VILNIAUS UNIVERSITETO
MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Vilijandas Bagdonavičius

Julius Jonas Kruopis

MATEMATINĖ STATISTIKA

Vadovėlis

II DALIS

TIESINIAI MODELIAI

Vilniaus universiteto leidykla
2015

UDK 519.2(075.8)

Apsvarstė ir rekomendavo spausdinti Vilniaus universiteto Matematikos ir informatikos fakulteto taryba (2015 m. vasario 17 d.; protokolas Nr 3); vadovėlio statusą suteikė Vilniaus universiteto Senatas (2015 m. balandžio 21 d. nutarimas Nr. S – 2015 – 4 –12).

Recenzavo:
prof. habil. dr. Algimantas Bikelis (Vytauto Didžiojo universitetas),
prof. habil. dr. Kęstutis Dučinskas (Klaipėdos universitetas)

ISBN 978–609–459–516–5

Turinys

Pratarmė 7
Trumpiniai ir žymenys 8

1 Tiesiniai modeliai **10**
1.1. Gauso ir Markovo tiesinis modelis 10
1.2. Mažiausių̨jų kvadratų įvertiniai ir jų savybės 12
1.2.1. Mažiausių̨jų kvadratų įvertiniai 12
1.2.2. Mažiausių̨jų kvadratų įvertinių savybės 13
1.3. Normaliojo skirstinio atvejis 16
1.3.1. Įvertinių savybės 16
1.3.2. Pasikliovimo intervalai ir hipotezių tikrinimas 21
1.4. Pratimai 27
1.5. Atsakymai ir nurodymai 30

2 Dispersinė analizė **31**
2.1. Vienfaktorė dispersinė analizė 31
2.1.1. Statistinis modelis 31
2.1.2. Mažiausių̨jų kvadratų įvertiniai 32
2.1.3. Vidurkių lygybės hipotezės tikrinimas 33
2.1.4. Kontrastų analizė 36
2.2. Dvifaktorė dispersinė analizė 41
2.2.1. Statistinis modelis 41
2.2.2. Mažiausių̨jų kvadratų įvertiniai 43
2.2.3. Faktorių įtakos hipotezių tikrinimas 44
2.2.4. Kontrastų analizė 47
2.2.5. Vieno stebėjimo langelyje atvejis 48
2.3. Dvifaktorė analizė, kai stebėjimų skaičius langeliuose skirtingas . 52
2.3.1. Statistinis modelis 52
2.3.2. Mažiausių̨jų kvadratų įvertiniai 53
2.3.3. Faktorių įtakos apibūdinimas 54
2.3.4. Faktorių sąveikos nebuvimo hipotezės tikrinimas 57
2.3.5. Faktorių įtakos nebuvimo hipotezės adityviajame modelyje tikrinimas 61
2.3.6. Faktorių įtakos nebuvimo hipotezės neadityviajame modelyje tikrinimas 64
2.4. Vienfaktorė dispersinė analizė, kai faktorius atsitiktinis 69

2.4.1. Statistinis modelis 69
2.4.2. Parametrų įvertiniai ir hipotezių tikrinimas 70
2.5. Dvifaktorė dispersinė analizė, kai faktoriai atsitiktiniai 73
2.5.1. Statistinis modelis 73
2.5.2. Kvadratų sumų skirstiniai 75
2.5.3. Parametrų įvertiniai ir hipotezių tikrinimas 76
2.6. Mišrusis dvifaktorės dispersinės analizės modelis 79
2.6.1. Statistinis modelis 79
2.6.2. Kvadratų sumų skirstiniai 81
2.6.3. Faktorių įtakos ir sąveikos nebuvimo hipotezių tikrinimas 84
2.6.4. Parametrų vertinimas 85
2.7. Daugiafaktorė dispersinė analizė 85
2.7.1. Statistinis modelis 86
2.7.2. Kvadratų sumų sudarymas 87
2.7.3. Laisvės laipsnių skaičių radimas 88
2.7.4. Vidurkių $\mathbf{E}(MS)$ radimas 88
2.7.5. Parametrų įvertiniai ir hipotezių tikrinimas 89
2.8. Dispersinės analizės eksperimentų planai 90
2.9. Dvifaktorė dispersinė analizė naudojant hierarchinę klasifikaciją . 93
2.9.1. Modelis, kai faktoriai pastovūs 93
2.9.2. Parametrų įvertiniai ir hipotezių tikrinimas 94
2.9.3. Modelis, kai faktoriai atsitiktiniai 97
2.10. Blokuotųjų duomenų dispersinė analizė 98
2.10.1. Vienfaktorė blokuotųjų duomenų dispersinė analizė 98
2.10.2. Dvifaktorė blokuotųjų duomenų dispersinė analizė 102
2.10.3. Nepilni subalansuoti blokai 105
2.11. Lotyniškieji kvadratai 111
2.11.1. Statistinis modelis 111
2.11.2. Parametrų įvertiniai ir hipotezių tikrinimas 112
2.12. Pratimai 116
2.13. Atsakymai ir nurodymai 125

3 Regresinė analizė **128**
3.1. Teoriniai regresijos pagrindai 128
3.1.1. Optimalioji prognozė 128
3.1.2. Tiesinė prognozė 130
3.1.3. Papildomos priklausomybės apibūdinimas 132
3.2. Tiesinė vieno kintamojo regresija 134
3.2.1. Statistinis modelis 134
3.2.2. Parametrų įvertiniai 135
3.2.3. Parametro β lygybės nuliui hipotezės tikrinimas 137
3.2.4. Tolesnio matavimo prognozė 137
3.2.5. Regresijos tiesiškumo hipotezės tikrinimas 139
3.2.6. Atsitiktinių kovariančių atvejis 140
3.2.7. Regresijos ir koreliacijos koeficientų sąryšis 141
3.2.8. Netiesinė regresija 143

3.3. Tiesinė keleto kintamųjų regresija . . . 145
3.3.1. Statistinis modelis . . . 145
3.3.2. Koeficientų $\boldsymbol{\beta}$ interpretacija . . . 146
3.3.3. Modelis, kai yra kovariančių sąveika . . . 147
3.3.4. Parametrų įvertiniai . . . 147
3.3.5. Koeficientų $\boldsymbol{\beta}$ ir jų tiesinių darinių pasikliovimo intervalai 149
3.3.6. Naujo stebėjimo reikšmės prognozė . . . 150
3.3.7. Hipotezių apie regresijos parametrų reikšmes tikrinimas . 150
3.3.8. Determinacijos koeficientas . . . 152
3.3.9. Empiriniai dalinės koreliacijos koeficientai . . . 153
3.3.10. Regresijos tiesinio pavidalo hipotezės tikrinimas . . . 154
3.3.11. Pažingsninė regresija . . . 156
3.4. Pratimai . . . 157
3.5. Atsakymai ir nurodymai . . . 162

4 Kovariacinė analizė 164
4.1. Kovariacinės analizės modeliai . . . 164
4.2. Vienfaktorė vieno kintamojo kovariacinė analizė . . . 166
4.2.1. Mažiausiųjų kvadratų įvertiniai . . . 166
4.2.2. Faktoriaus įtakos nebuvimo hipotezės tikrinimas . . . 167
4.2.3. Trukdančių parametrų įtakos nebuvimo hipotezės tikrinimas 168
4.2.4. Regresijos tiesių lygiagretumo hipotezės tikrinimas . . . 168
4.3. Bendrasis kovariacinės analizės atvejis . . . 169
4.3.1. Parametrų įvertiniai ir liekamoji kvadratinė forma . . . 169
4.3.2. Hipotezių tikrinimas . . . 171
4.4. Dispersinė ir kovariacinė analizė – atskiri regresinės analizės atvejai 173
4.4.1. Vienfaktorė dispersinė analizė . . . 174
4.4.2. Vienfaktorė kovariacinė analizė . . . 175
4.4.3. Dvifaktorė dispersinė analizė . . . 176
4.4.4. Dvifaktorė kovariacinė analizė . . . 178
4.5. Faktoriniai eksperimentai 2^m . . . 179
4.5.1. Faktoriniai eksperimentai 2^2 . . . 180
4.5.2. Faktoriniai eksperimentai 2^3 . . . 182
4.5.3. Faktoriniai eksperimentai 2^m . . . 183
4.5.4. Faktorinių eksperimentų 2^m replikos . . . 184
4.6. Pratimai . . . 187
4.7. Atsakymai ir nurodymai . . . 193

5 Apibendrintieji tiesiniai modeliai 195
5.1. Vienparametrių eksponentinio tipo skirstinių tiesiniai modeliai . 195
5.2. Apibendrintųjų tiesinių modelių pavyzdžiai . . . 199
5.2.1. Puasoninė regresija . . . 199
5.2.2. Gama regresija . . . 201
5.2.3. Neigiamoji binominė regresija . . . 202
5.2.4. Binominė regresija . . . 204
5.3. Logistinė regresija . . . 204

5.3.1. Logistinės regresijos modelis . . . 204
5.3.2. Regresinių parametrų interpretavimas . . . 205
5.3.3. Regresinių parametrų vertinimas . . . 208
5.3.4. Regresinių parametrų įvertinių savybės . . . 210
5.3.5. Tikėtinumų santykiai ir determinacijos koeficientas . . . 212
5.3.6. Regresijos parametrų lygybės nuliui hipotezių tikrinimas . . . 214
5.3.7. Įvykio A tikimybės ir regresinių parametrų pasikliovimo intervalai . . . 216
5.3.8. Klasifikavimo uždaviniai . . . 217
5.4. Pratimai . . . 221
5.5. Atsakymai ir nurodymai . . . 224

6 1 Priedas. Tiesinės algebros elementai 226
6.1. Vektoriai . . . 226
6.2. Matricos ir determinantai . . . 228

7 2 priedas. Atsitiktiniai vektoriai 231
7.1. Atsitiktinio vektoriaus skirstinys . . . 231
7.2. Marginalieji ir sąlyginiai skirstiniai . . . 232
7.3. Atsitiktinio vektoriaus momentai . . . 233
7.4. Daugiamatis normalusis skirstinys . . . 234
Literatūra . . . 235
Dalykinė rodyklė . . . 236

Pratarmė

Pirmojoje vadovėlio dalyje daugiausia buvo nagrinėjamos paprastosios imtys, kai jų elementai yra vienodai pasiskirstę nepriklausomi atsitiktiniai dydžiai. Šioje dalyje nagrinėjamos imtys nėra paprastosios, jų elementų skirstiniai gali priklausyti nuo vieno ar kelių kiekybinių ar kokybinių faktorių. Analizės tikslas būtent ir yra nustatyti, ar imties skirstinys priklauso ir kaip priklauso nuo dominančių faktorių.

Naudojami statistiniai modeliai vadinami tiesiniais, nes imties elementų vidurkiai aprašomi tiesinėmis nežinomų parametrų funkcijomis. Imties skirstinio priklausomybę nuo tam tikrų faktorių ar jų darinių ir apibūdina minėtieji parametrai.

Tiesiniai modeliai – viena iš svarbiausių matematinės statistikos dalių. Jie labai plačiai naudojami įvairiose mokslo ir praktikos srityse kylantiems uždaviniams spręsti. Yra daug matematinės statistikos knygų ir monografijų, skirtų tiesiniams modeliams nagrinėti. Jos skiriasi pateikiamos medžiagos matematiniu lygiu ir apimtimi, taip pat taikymų srities specifika. Skaitytojui rekomenduojame monografijas [2], [10], [12], [14].

Knygos paskirtis lėmė medžiagos iš šios plačios matematinės statistikos srities parinkimą. Daugiausia apsiribojama tiesiniais modeliais, kai jų paklaidos yra nepriklausomi ir normalieji atsitiktiniai dydžiai. Tokiu atveju teorija yra nusistovėjusi ir įgijusi užbaigtą pavidalą.

Pirmajame skyriuje pateikiami bendri tiesinių modelių analizės rezultatai, kurie tolesniuose skyriuose taikomi nagrinėjant specialius tiesinius modelius.

Antrajame skyriuje nagrinėjami dispersinės analizės, trečiajame – regresinės analizės, o ketvirtajame – kovariacinės analizės modeliai. Nors faktiškai visus šiuos modelius galima traktuoti kaip tam tikrus regresinės analizės atvejus, tačiau dėl jų taikymo specifikos matematinės statistikos literatūroje juos įprasta nagrinėti atskirai.

Penktajame skyriuje trumpai aptariami apibendrintieji tiesiniai modeliai ir detaliau logistinė regresija.

Prieduose pateikiami naudojami tiesinės algebros faktai ir kai kurios atsitiktinių vektorių savybės.

Vadovėlis parengtas remiantis paskaitomis, kurias autoriai skaitė Vilniaus universiteto Matematikos ir informatikos fakulteto statistikos studijų programos studentams.

Autoriai

Trumpiniai ir žymenys

A. d. – atsitiktinis dydis;
n. a. d. – nepriklausomi atsitiktiniai dydžiai;
a. v. – atsitiktinis vektorius;
n. a. v. – nepriklausomi atsitiktiniai vektoriai;
TG – tolygiai galingiausias (kriterijus);
TGN – tolygiai galingiausias nepaslinktasis (kriterijus);
DT – didžiausiojo tikėtinumo (funkcija, metodas, įvertinys);
MK – mažiausiųjų kvadratų (metodas, įvertinys);
TPP – taikomieji programų paketai;
$X,\ Y,\ Z, \ldots$ – atsitiktiniai dydžiai;
$\mathbf{X},\ \boldsymbol{Y},\ \boldsymbol{Z}, \ldots$ – atsitiktiniai vektoriai;
$\mathbf{X}^T$ – transponuotas vektorius, t. y. vektorius-eilutė;
$x(P)$ – P-asis kvantilis;
x_P – P-oji kritinė reikšmė;
$\mathbf{\Sigma} = [\sigma_{ij}]_{k\times k}$ – kovariacijų matrica;
$\rho = [\rho_{ij}]_{k\times k}$ – koreliacijos koeficientų matrica;
$\mathbf{P}\{A\}$ – įvykio A tikimybė;
$\mathbf{P}\{A|B\}$ – įvykio A sąlyginė tikimybė;
$\mathbf{P}_\theta\{A\},\ \mathbf{P}\{A|\theta\}$ – tikimybė, priklausanti nuo parametro θ;
$F_\theta(x),\ F(x;\ \theta),\ F(x|\theta)$ – pasiskirstymo funkcija, priklausanti nuo parametro θ (analogiškai tankio funkcijai);
$\mathbf{E}X$ – a. d. X vidurkis;
$\boldsymbol{V}X$ – a. d. X dispersija;
$\mathbf{E}_\theta(X),\ \mathbf{E}(X|\theta),\ \boldsymbol{V}_\theta(X),\ \boldsymbol{V}(X|\theta)$ – a. d. X vidurkis ar dispersija, priklausantys nuo parametro θ;
$\mathbf{E}(\mathbf{X})$ – a. v. $\mathbf{X}$ vidurkių vektorius;
$\boldsymbol{V}(\mathbf{X})$ – a. v. $\mathbf{X}$ kovariacijų matrica;
$\mathbf{Cov}\,(X,\ Y)$ – a. d. X ir Y kovariacija;
$\mathbf{Cov}\,(\mathbf{X},\ \boldsymbol{Y})$ – a. v. $\mathbf{X}$ ir $\boldsymbol{Y}$ kovariacijų matrica;
$N(0,\ 1)$ – standartinis normalusis skirstinys;
$N(\mu,\ \sigma^2)$ – normalusis skirstinys su parametrais μ ir σ^2;
$\chi^2(n)$ – chi kvadrato skirstinys su n laisvės laipsnių;
$\chi^2(n;\ \delta)$ – necentrinis chi kvadrato skirstinys su n laisvės laipsnių ir necentriškumo parametru δ;
$S(n)$ – Stjudento skirstinys su n laisvės laipsnių;
$S(n;\ \delta)$ – necentrinis Stjudento skirstinys su n laisvės laipsnių ir necentriškumo parametru δ;
$F(m,\ n)$ – Fišerio skirstinys su m ir n laisvės laipsnių;
$F(m,\ n;\ \delta)$ — necentrinis Fišerio skirstinys su m ir n laisvės laipsnių ir necentriškumo parametru δ;

z_α – standartinio normaliojo skirstinio α kritinė reikšmė;
$t_\alpha(n)$ – Stjudento skirstinio su n laisvės laipsnių α kritinė reikšmė;
$\chi^2_\alpha(n)$ – chi kvadrato skirstinio su n laisvės laipsnių α kritinė reikšmė;
F_α(m, n) – Fišerio skirstinio su m ir n laisvės laipsnių α kritinė reikšmė;
$N_k(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ – k-matis normalusis skirstinys su vidurkių vektoriumi $\boldsymbol{\mu}$ ir kovariacijų matrica $\boldsymbol{\Sigma}$;
$X \sim N(\mu,\ \sigma^2)$ – a. d. X, pasiskirstęs pagal normalųjį dėsnį su parametrais μ ir σ^2 (analogiškai kitų skirstinių atveju);
$X_n \xrightarrow{P} X$ – konvergavimas pagal tikimybę ($n \to \infty$);
$X_n \xrightarrow{b.t.} X$ – konvergavimas su tikimybe 1 arba beveik tikrai ($n \to \infty$);
$X_n \xrightarrow{kv.v.} X$ – konvergavimas pagal kvadratinį vidurkį ($n \to \infty$);
$X_n \xrightarrow{d} X$, $F_n(x) \xrightarrow{d} F(x)$ – konvergavimas pagal pasiskirstymą (silpnasis; $n \to \infty$);
$X_n \xrightarrow{d} X \sim N(\mu,\ \sigma^2)$ – a. d. X_n asimptotiškai ($n \to \infty$) turi normalųjį skirstinį su parametrais μ ir σ^2;
$X_n \sim Y_n$ – a. d. X_n ir Y_n asimptotiškai ($n \to \infty$) ekvivalentūs ($X_n - Y_n \xrightarrow{P} 0$);
$||\boldsymbol{x}||$ – kai $\boldsymbol{x} = (x_1, ..., x_k)^T$ yra vektorius, reiškia atstumą $(\boldsymbol{x}^T\boldsymbol{x})^{1/2} = (\sum_i x_i^2)^{1/2}$;
$||\boldsymbol{A}||$ – kai $\boldsymbol{A} = [a_{ij}]$ yra matrica, reiškia $(\sum_i \sum_j a_{ij}^2)^{1/2}$;
$\boldsymbol{A} > \boldsymbol{B}$ $(\boldsymbol{A} \geq \boldsymbol{B})$ — kai $\boldsymbol{A}$ ir $\boldsymbol{B}$ yra vienodos dimensijos kvadratinės matricos, reiškia, kad matrica $\boldsymbol{A} - \boldsymbol{B}$ yra teigiamai (neneigiamai) apibrėžta.

1 skyrius

Tiesiniai modeliai

1.1. Gauso ir Markovo tiesinis modelis

Tiesiniai modeliai – plati modelių klasė ir yra, ko gero, labiausiai naudojami taikomojoje statistikoje. Juos naudojant tiriama stebimų diskrečiųjų ar tolydžiųjų kintamųjų (kovariančių, faktorių) įtaka nagrinėjamų požymių skirstiniams. Modeliai vadinami tiesiniais, nes požymio vidurkiai nusakomi tiesinėmis nežinomų parametrų funkcijomis.

Tarkime, kad imties $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ nariai Y_i yra nepriklausomi tokio pavidalo atsitiktiniai dydžiai:

$$Y_i = \boldsymbol{a}_i{}^T\boldsymbol{\beta} + e_i = a_{i1}\beta_1 + \cdots + a_{im}\beta_m + e_i, \;\; i = 1, 2, ..., n; \tag{1.1.1}$$

čia $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^T \in \mathbf{R}^m$ yra nežinomų parametrų vektorius, $\boldsymbol{a}_i = (a_{i1}, \cdots, a_{im})^T$ – žinomas vektorius, $\boldsymbol{e} = (e_1, ..., e_n)^T$ – atsitiktinis vektorius, kurio koordinatės yra vienodai pasiskirstę n. a. d. su nuliniais vidurkiais ir nežinomomis vienodomis dispersijomis σ^2. Toks modelis vadinamas *Gauso ir Markovo tiesiniu modeliu.*

Imtis nėra paprastoji, nes a. d. Y_i nėra vienodai pasiskirstę.

Šiame skyriuje $\boldsymbol{a}_i = (a_{i0}, \cdots, a_{im})^T$ neatsitiktiniai arba yra atsitiktinių dydžių realizacijos. Šiuo atveju analizė yra sąlyginė, žinant šias realizacijas.

Pažymėjus $\boldsymbol{A} = [a_{ij}]_{n\times m}$ matricą, sudarytą iš koeficientų a_{ij}, modelį (1.1.1) galima trumpiau užrašyti matricine forma:

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}, \;\; \mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\beta}, \;\; \boldsymbol{V}(\boldsymbol{Y}) = \boldsymbol{V}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}. \tag{1.1.2}$$

Matrica $\boldsymbol{A}$ vadinama *eksperimento plano matrica.*

Remdamiesi pateikiamais bendrais rezultatais, toliau nagrinėsime konkrečius (1.1.2) modelio atvejus. Matematinės statistikos uždavinių, susijusių su tais konkrečiais modeliais, sprendimas vadinamas *dispersine ir regresine analize.*

Minėta, kad tiesinių modelių analizė susijusi su a. d. $Y_1, ..., Y_n$ skirstinių priklausomybės nuo tam tikrų faktorių (jų įtaką modelyje (1.1.2) nusako parametrai

$\beta_1, ...\beta_m$) tyrimu. Tie faktoriai gali būti kokybiniai arba kiekybiniai. Pavyzdžiui, gali būti tiriama kviečių derlingumo priklausomybė nuo jų veislės ir auginimo metodikos; išdirbio priklausomybė nuo darbininko kvalifikacijos ir staklių markės ir pan. Kviečių veislė ir auginimo metodika, darbininko kvalifikacija ir staklių markė – kokybiniai faktoriai. *Dispersinės analizės* pavadinimas yra paliktas būtent tokiems modeliams, kai tiriama a. v. $\boldsymbol{Y}$ skirstinio priklausomybė nuo kokybinių faktorių.

1.1.1 pavyzdys. Norint ištirti kviečių derlingumo priklausomybę nuo jų veislės, atliekamas toks eksperimentas. Atsitiktinai parinktų n_1 sklypelių apsėjama pirmąja kviečių veisle; n_2 – antrąja ir t. t., ir pagaliau n_m sklypelių – m-ąja veisle. Pažymėkime Y_{ij} i-osios kviečių veislės derlingumą j-ajame sklypelyje. Tarkime, kad derlingumo pokyčiai pereinant nuo vienos kviečių veislės prie kitos nekeičia dispersijos, o gali keisti tik vidurkį. Gauname modelį

$$Y_{ij} = \beta_i + e_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., m.$$

Tegu e_{ij} nekoreliuoti vienodų dispersijų a. d., o $\beta_i = \mathbf{E}Y_{ij}$ žymi i-osios kviečių veislės vidutinį derlingumą.

Sujungę stebėjimus Y_{ij} į vieną bendrą vektorių

$$\boldsymbol{Y} = (Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2}, ..., Y_{m1}, ..., Y_{mn_m})^T$$

ir pažymėję nežinomų parametrų vektorių $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^T$, gausime modelio aprašymą matricine forma (1.1.2). Matrica $\boldsymbol{A}$ turi $n = n_1 + \cdots + n_m$ eilučių ir m stulpelių. Pirmosios n_1 eilutės turi pavidalą $(1, 0, \ldots, 0)$, paskui n_2 eilučių turi pavidalą $(0, 1, 0, \ldots, 0)$, pagaliau paskutinės n_m eilučių turi pavidalą $(0, 0, \ldots, 0, 1)$.

Atsitiktinio vektoriaus $\boldsymbol{Y}$ skirstinio priklausomybės nuo kiekybinių faktorių (pvz., trašų kiekio, temperatūros, svorio ir kt.) analizė vadinama *regresine analize*. Regresinėje analizėje dydžiai $a_{i1}, \cdots, a_{im}$ yra interpretuojami kaip m kovariančių i-osios reikšmės, tada koeficientai $\beta_1, \ldots, \beta_m$ parodo atitinkamų kovariančių įtaką tiriamo požymio vidurkiui.

1.1.2 pavyzdys. Norint ištirti vyrų sistolinio kraujo spaudimo Y priklausomybę nuo jų svorio X_1 ir amžiaus X_2, atsitiktinai atrenkama n vyrų ir jiems pamatuojamos a. v. $(Y, X_1, X_2)^T$ reikšmės $(Y_i, x_{1i}, x_{2i})^T$, $i = 1, ..., n$.

Tarkime, kad stebėjimai, kai skirtingi $i = 1, ..., n$ yra nepriklausomi a. d.; Y sąlyginio skirstinio, kai $\mathbf{X} = (X_1, X_2)^T = (x_1, x_2)^T$ yra fiksuotas, dispersija lygi σ^2 ir nepriklauso nuo $\boldsymbol{x} = (x_1, x_2)^T$, o Y vidurkis yra tiesinė funkcija: $\mathbf{E}(Y|\mathbf{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Tardami, kad a. v. $(X_1, X_2)^T$ realizacijos $(x_{1i}, x_{2i}), i = 1, ..., n$ yra fiksuotos, gauname modelį

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad i = 1, ..., n.$$

Pažymėjus $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ ir $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ nežinomų parametrų vektorių, modelį galima užrašyti matriciniu pavidalu (1.1.2) su matrica $\boldsymbol{A}$, kuri turi n eilučių ir 3 stulpelius: i-oji eilutė turi pavidalą $(1, x_{1i}, x_{2i})$.

Kaip matome, formalus dispersinės ir regresinės analizės schemų skirtumas yra toks. Pirmuoju atveju matricos $\boldsymbol{A}$ elementai yra arba 0, arba 1, t. y. jie parodo, ar stebėjimas Y gautas veikiant tam tikram faktoriaus lygmeniui. Regresinėje analizėje matricos $\boldsymbol{A}$ elementai gali būti bet kokie realūs skaičiai. Abiem modeliams yra pritaikomi toliau pateikiami bendrieji rezultatai. Tačiau dispersinės ir regresinės analizės uždaviniai skiriasi savo taikymo specifika, todėl matematinės statistikos literatūroje šios schemos nagrinėjamos atskirai.

1.2. Mažiausiųjų kvadratų įvertiniai ir jų savybės

1.2.1. Mažiausiųjų kvadratų įvertiniai

Modelio (1.1.2) nežinomo parametro $\boldsymbol{\beta}$ įvertinio $\hat{\boldsymbol{\beta}}$ ieškoma minimizuojant kvadratinę formą

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - a_{i1}\beta_1 - \cdots - a_{im}\beta_m)^2, \quad (1.2.1)$$

kuri lygi atstumų tarp stebėjimų Y_i ir jų vidurkių $\mathbf{E}Y_i = a_{i1}\beta_1 + ... + a_{im}\beta_m$ kvadratų sumai.

Diferencijuodami (1.2.1) pagal parametrą $\boldsymbol{\beta}$ ir prilyginę išvestinę $\mathbf{0}$, gauname lygčių sistemą

$$-2\boldsymbol{A}^T(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta}) = \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{A}^T\boldsymbol{Y}. \quad (1.2.2)$$

Tarsime, kad $m \times m$ matrica $\boldsymbol{A}^T\boldsymbol{A}$ neišsigimusi (o tai ekvivalentu, kad matricų $\boldsymbol{A}$ ir $\boldsymbol{A}^T\boldsymbol{A}$ rangai lygūs m). Tada sistemos (1.2.2) sprendinys yra vienintelis (žr. 1 priedą (6.2.22)):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}. \quad (1.2.3)$$

1.2.1 pastaba. Jeigu $Rang(\boldsymbol{A}) < m$, tai modelį (1.2.1) galima modifikuoti sumažinant parametrų skaičių taip, kad naujame modelyje matricos $\boldsymbol{A}$ rangas sutaptų su nežinomų parametrų skaičiumi. Todėl sąlyga $Rang(\boldsymbol{A}) = m$ iš esmės nemažina bendrumo.

1.2.1 apibrėžimas. Įvertinys (1.2.3) vadinamas parametro $\boldsymbol{\beta}$ *mažiausiųjų kvadratų (MK) įvertiniu.*

1.2.1 pavyzdys. Tarkime, imties $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^T$ elementai turi tokią struktūrą

$$\begin{aligned} Y_1 &= \alpha + \beta_2 + e_1, \\ Y_2 &= \alpha + 2\beta_1 - \beta_2 + e_2, \\ Y_3 &= \alpha + \beta_2 + e_3, \\ Y_4 &= \alpha - 2\beta_1 + e_4, \\ Y_5 &= \alpha - \beta_2 + e_5; \end{aligned}$$

čia $e_1, ..., e_5$ yra n. a. d., turintys nulinius vidurkius ir vienodas dispersijas σ^2. Rasime parametro $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2)^T$ MK įvertinį ir jo realizaciją (įvertį), kai a. v. $\boldsymbol{Y}$ realizacija yra $(5,2; 0,8; 4,9; 0,4; -0,7)^T$.

Turime tiesinį Gauso ir Markovo modelį $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}$, kuriame

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 0 & 1 \\ 1 & -2 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}.$$

Gauname

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 8 & -2 \\ 0 & -2 & 4 \end{pmatrix}, \ (\boldsymbol{A}^T\boldsymbol{A})^{-1} = \begin{pmatrix} 1/5 & 0 & 0 \\ 0 & 1/7 & 1/14 \\ 0 & 1/14 & 2/7 \end{pmatrix},$$

$$\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} Y_1+Y_2+Y_3+Y_4+Y_5 \\ 2Y_2-2Y_4 \\ Y_1-Y_2+Y_3-Y_5 \end{pmatrix},$$

ir

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} (Y_1+Y_2+Y_3+Y_4+Y_5)/5 \\ (Y_1+3Y_2+Y_3-4Y_4-Y_5)/14 \\ (2Y_1-Y_2+2Y_3-Y_4-2Y_5)/7 \end{pmatrix}.$$

Pagal turimą imties realizaciją gauname įvertį

$$\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} 10,6 \\ 0,8 \\ 10 \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 2,12 \\ 5,8/7 \\ 20,4/7 \end{pmatrix} \approx \begin{pmatrix} 2,1200 \\ 0,8286 \\ 2,9143 \end{pmatrix}.$$

1.2.2. Mažiausiųjų kvadratų įvertinių savybės

1.2.1 teorema. *MK įvertinys $\hat{\boldsymbol{\beta}}$ minimizuoja (1.2.1) kvadratinę formą.*

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta}) \geq (\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}}) =: SS_E. \qquad (1.2.4)$$

Įrodymas. Turime:

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}}+\boldsymbol{A}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}}+\boldsymbol{A}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))$$

$$= (\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{A}^T\boldsymbol{A}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \geq (\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}}) = SS_E.$$

▲

1.2.2 pastaba. Liekamoji kvadratinė forma SS_E gali būti užrašyta tokiu pavidalu:

$$SS_E = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{Y} = \sum_{i=1}^{n} Y_i^2 - \sum_{j=1}^{m} \hat{\beta}_j h_j, \qquad (1.2.5)$$

čia h_j yra j-oji vektoriaus $\boldsymbol{A}^T\boldsymbol{Y}$ koordinatė.

Įrodymas. Turime

$$SS_E = (\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{Y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}} =$$

$$= \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T(\boldsymbol{A}^T\boldsymbol{Y} - \boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{Y},$$

nes $\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}}$. ▲

1.2.2 pavyzdys (**1.2.1** pavyzdžio tęsinys). Rasime liekamosios kvadratų sumos SS_E realizaciją pagal **1.2.1** pavyzdžio duomenis.

Gauname

$$SS_E = (Y_1-\hat{\alpha}-\hat{\beta}_2)^2+(Y_2-\hat{\alpha}-2\hat{\beta}_1+\hat{\beta}_2)^2+(Y_3-\hat{\alpha}-\hat{\beta}_2)^2+(Y_4-\hat{\alpha}+2\hat{\beta}_1)^2+(Y_5-\hat{\alpha}+\hat{\beta}_2)^2 = 0,0623.$$

Liekamąją kvadratų sumą galima rasti ir pagal (1.2.5) formulę:

$$SS_E = Y_1^2+Y_2^2+Y_3^2+Y_4^2+Y_5^2 - 10,6\hat{\alpha} - 0,8\hat{\beta}_1 - 10\hat{\beta}_2 = 0,0623.$$

1.2.2 teorema. *Jei* $Rang(\boldsymbol{A}^T\boldsymbol{A}) = m$, *tai*

a) MK įvertinys $\hat{\boldsymbol{\beta}}$ *yra nepaslinktasis ir jo pirmieji du momentai yra*

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}.$$

b) Dispersijos σ^2 *nepaslinktasis įvertinys* s^2 *yra*

$$s^2 = \frac{SS_E}{n-m} = \frac{(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})}{n-m}. \tag{1.2.6}$$

Įrodymas. a) Gauname (žr. 2 priedą (7.3.4))

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\mathbf{E}(\boldsymbol{Y}) = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta},$$

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\sigma^2\boldsymbol{I}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}.$$

b) Liekamąją kvadratinę formą SS_E galima užrašyti taip (žr. (1.2.5))

$$SS_E = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{Y}^T[\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T]\boldsymbol{Y}.$$

Laužtiniuose skliaustuose parašytoji matrica (pažymėkime ją $\boldsymbol{B}$) yra simetriška. Kadangi

$$\mathbf{E}(\boldsymbol{Y})^T\boldsymbol{B}\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{\beta}^T\boldsymbol{A}^T(\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{\beta} = 0, \quad \boldsymbol{B}\boldsymbol{B} = \boldsymbol{B},$$

tai, remdamiesi matricos pėdsako savybe (žr. 1 priedą (6.2.4)), gauname

$$\mathbf{E}(SS_E) = \mathbf{E}(\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}) = \mathbf{E}[(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))^T\boldsymbol{B}(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))]$$

$$= \mathbf{E}\mathbf{Tr}[(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))^T\boldsymbol{B}(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))] = \mathbf{E}\mathbf{Tr}[\boldsymbol{B}(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))^T]$$

$$= \mathbf{Tr}(\boldsymbol{B}\mathbf{E}[(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))(\boldsymbol{Y} - \mathbf{E}(\boldsymbol{Y}))^T]) = \sigma^2\mathbf{Tr}(\boldsymbol{B}\boldsymbol{I}) = \sigma^2\mathbf{Tr}(\boldsymbol{B})$$

$$= \sigma^2[\mathbf{Tr}(\boldsymbol{I}) - \mathbf{Tr}(\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1})] = \sigma^2(n-m).$$

▲

1.2.3 pavyzdys (**1.2.1** pavyzdžio tęsinys). Rasime įvertinio $\hat{\boldsymbol{\beta}}$ kovariacijų matricą ir dispersijos σ^2 įvertį.

Remiantis **1.2.2** teorema

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \sigma^2 \begin{pmatrix} 1/5 & 0 & 0 \\ 0 & 1/7 & 1/14 \\ 0 & 1/14 & 2/7 \end{pmatrix}.$$

Iš čia $\boldsymbol{V}\hat{\alpha} = \sigma^2/5$, $\boldsymbol{V}\hat{\beta}_1 = \sigma^2/7$, $\boldsymbol{V}\hat{\beta}_2 = 2\sigma^2/7$, $\mathbf{Cov}\,(\hat{\beta}_1, \hat{\beta}_2) = \sigma^2/14$, $\mathbf{Cov}\,(\hat{\alpha}, \hat{\beta}_1)$=0, $\mathbf{Cov}\,(\hat{\alpha}_1, \hat{\beta}_2) = 0$. Dispersijos σ^2 nepaslinktasis įvertis

$$s^2 = \frac{SS_E}{n-m} = \frac{SS_E}{2} = 0,03115.$$

Dažnai tenka vertinti ne tiktai parametrus $\beta_1, \ldots, \beta_m$, bet ir jų tiesines funkcijas. Pažymėkime $\mathcal{G}_{\boldsymbol{L}}$ parametro $\theta = \boldsymbol{L}^T\boldsymbol{\beta} = L_1\beta_1 + \cdots + L_m\beta_m$ *tiesinių nepaslinktųjų įvertinių klasę.*

1.2.3 teorema. (Gauso ir Markovo). *Jei* $Rang(\boldsymbol{A}^T\boldsymbol{A}) = m$, *tai* $\boldsymbol{L}^T\hat{\boldsymbol{\beta}}$ *yra vienintelis minimalios dispersijos įvertinys klasėje* $\mathcal{G}_{\boldsymbol{L}}$.

Šio įvertinio pirmieji du momentai ir kovariacija su kitos tiesinės funkcijos $\boldsymbol{K}^T\boldsymbol{\beta}$ *įvertiniu* $\boldsymbol{K}^T\hat{\boldsymbol{\beta}}$ *yra*

$$\mathbf{E}(\hat{\theta}) = \theta, \quad \boldsymbol{V}(\hat{\theta}) = \sigma^2\boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{L}, \quad \mathbf{Cov}\,(\boldsymbol{L}^T\hat{\boldsymbol{\beta}},\ \boldsymbol{K}^T\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{K}. \tag{1.2.7}$$

Jei $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)^T = (\boldsymbol{H}_1^T\boldsymbol{\beta}, ..., \boldsymbol{H}_k^T\boldsymbol{\beta})^T = \boldsymbol{H}\boldsymbol{\beta},\ \boldsymbol{H} = [h_{ij}]_{k\times m}$, *yra* k*-matis parametras, tai* $\hat{\boldsymbol{\theta}} = \boldsymbol{H}\hat{\boldsymbol{\beta}}$ *yra nepaslinktasis parametro* $\boldsymbol{\theta}$ *įvertinys ir*

$$\mathbf{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}, \quad \boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \sigma^2\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T. \tag{1.2.8}$$

Įrodymas. Jei $\boldsymbol{M}^T\boldsymbol{Y} \in \mathcal{G}_{\boldsymbol{L}}$, tai

$$\boldsymbol{L}^T\boldsymbol{\beta} = \mathbf{E}(\boldsymbol{M}^T\boldsymbol{Y}) = \mathbf{E}(\boldsymbol{M}^T\boldsymbol{Y} - \boldsymbol{L}^T\hat{\boldsymbol{\beta}} + \boldsymbol{L}^T\hat{\boldsymbol{\beta}}) = (\boldsymbol{M}^T\boldsymbol{A} - \boldsymbol{L}^T)\boldsymbol{\beta} + \boldsymbol{L}^T\boldsymbol{\beta}.$$

Taigi

$$(\boldsymbol{M}^T\boldsymbol{A} - \boldsymbol{L}^T)\boldsymbol{\beta} = 0 \quad \text{su visais} \quad \boldsymbol{\beta} \in \boldsymbol{R}^m$$

ir

$$\boldsymbol{M}^T\boldsymbol{A} - \boldsymbol{L}^T = \mathbf{0}. \tag{1.2.9}$$

Turime

$$\boldsymbol{V}(\boldsymbol{M}^T\boldsymbol{Y}) = \boldsymbol{V}(\boldsymbol{M}^T\boldsymbol{Y} - \boldsymbol{L}^T\hat{\boldsymbol{\beta}} + \boldsymbol{L}^T\hat{\boldsymbol{\beta}})$$

$$= \boldsymbol{V}(\boldsymbol{M}^T\boldsymbol{Y} - \boldsymbol{L}^T\hat{\boldsymbol{\beta}}) + \boldsymbol{V}(\boldsymbol{L}^T\hat{\boldsymbol{\beta}}) + 2\mathbf{Cov}\,(\boldsymbol{M}^T\boldsymbol{Y} - \boldsymbol{L}^T\hat{\boldsymbol{\beta}}, \boldsymbol{L}^T\hat{\boldsymbol{\beta}}).$$

Žymėkime $\boldsymbol{B} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. Iš lygybės (1.2.9) gaunama (žr. 2 priedą (7.3.4))

$$\mathbf{Cov}\,(\boldsymbol{M}^T\boldsymbol{Y} - \boldsymbol{L}^T\hat{\boldsymbol{\beta}}, \boldsymbol{L}^T\hat{\boldsymbol{\beta}}) = \mathbf{Cov}\,((\boldsymbol{M}^T - \boldsymbol{L}^T\boldsymbol{B})\boldsymbol{Y}, \boldsymbol{L}^T\boldsymbol{B}\boldsymbol{Y})$$

$$= (\boldsymbol{M}^T - \boldsymbol{L}^T\boldsymbol{B})\sigma^2\mathbf{I}\boldsymbol{B}^T\boldsymbol{L} = \sigma^2(\boldsymbol{M}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}$$

$$-\boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1})\boldsymbol{L} = \sigma^2(\boldsymbol{M}^T\boldsymbol{A} - \boldsymbol{L}^T)(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{L} = 0,$$

Taigi

$$\boldsymbol{V}(\boldsymbol{M}^T\boldsymbol{Y}) = \boldsymbol{V}(\boldsymbol{L}^T\hat{\boldsymbol{\beta}}) + \boldsymbol{V}((\boldsymbol{M}^T - \boldsymbol{L}^T\boldsymbol{B})\boldsymbol{Y}) \geq \boldsymbol{V}(\boldsymbol{L}^T\hat{\boldsymbol{\beta}}).$$

Lygybė teisinga tada ir tik tada, kai $\boldsymbol{M}^T = \boldsymbol{L}^T\boldsymbol{B}$.

Randame

$$\mathbf{Cov}\,(\boldsymbol{L}^T\hat{\boldsymbol{\beta}},\ \boldsymbol{K}^T\hat{\boldsymbol{\beta}}) = \mathbf{Cov}\,(\boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y},\ \boldsymbol{K}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y})$$

$$= \sigma^2\boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{K}.$$

▲

1.2.4 pavyzdys (**1.2.1** pavyzdžio tęsinys). Tarkime, kad **1.2.1** pavyzdžio sąlygomis mus domina parametras $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, $\theta_1 = \alpha - 2\beta_1$, $\theta_2 = \alpha + \beta_1 - \beta_2$. Rasime parametro $\boldsymbol{\theta}$ įvertinį ir jo kovariacijų matricą.

Parametras $\boldsymbol{\theta}$ yra tiesinė $\boldsymbol{\beta}$ funkcija

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^T\boldsymbol{\beta} \\ \boldsymbol{K}^T\boldsymbol{\beta} \end{pmatrix}; \quad \boldsymbol{L} = (1;\ -2,\ 0)^T, \quad \boldsymbol{K} = (1,\ 1,\ -1)^T$$

Remiantis **1.2.3** teorema, vienintelis mažiausios dispersijos įvertinys nepaslinktų tiesinių įvertinių aibėje yra

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^T\hat{\boldsymbol{\beta}} \\ \boldsymbol{K}^T\hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \hat{\alpha} - 2\hat{\beta}_1 \\ \hat{\alpha} + \hat{\beta}_1 - \hat{\beta}_2 \end{pmatrix},$$

o jo kovariacijų matrica

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \sigma^2 \begin{pmatrix} \boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{L} & \boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{K} \\ \boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{K} & \boldsymbol{K}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{K} \end{pmatrix} = \frac{\sigma^2}{35}\begin{pmatrix} 27 & 2 \\ 2 & 17 \end{pmatrix}.$$

Įvertinio $\hat{\boldsymbol{\theta}}$ realizacija yra $(0,4628;\ 0,0343)^T$.

1.3. Normaliojo skirstinio atvejis

Sakykime, kad paklaidos e_i turi normalųjį skirstinį $N(0,\sigma^2)$. Tada a. v. $\boldsymbol{Y}$ skirstinys taip pat normalusis (žr 2 priedą 7.4 skyrelį):

$$\boldsymbol{Y} \sim N_n(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}). \tag{1.3.1}$$

Modelis turi $m+1$ nežinomą parametrą $\beta_1, ..., \beta_m, \sigma^2$.

Šiuo atveju galima ne tik rasti įvertinių $\hat{\boldsymbol{\beta}}$, s^2 momentus, bet ir jų tikimybinius skirstinius. Jais naudojamasi sudarant parametrų pasikliovimo intervalus ir kuriant kriterijus hipotezėms apie parametrų reikšmes tikrinti.

1.3.1. Įvertinių savybės

Atsitiktinio vektoriaus $\boldsymbol{Y}$ didžiausiojo tikėtinumo funkcija turi tokį pavidalą

$$L = L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta})\} =$$

$$= \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\{-\frac{1}{2\sigma^2}SS(\boldsymbol{\beta})\}.$$

Gauname

$$\ln L = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}SS(\boldsymbol{\beta}),$$

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{A}^T(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta}), \quad \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{SS(\boldsymbol{\beta})}{2\sigma^4}.$$

Taigi parametro $\boldsymbol{\beta}$ DT įvertinys sutampa su MK įvertiniu. Parametro σ^2 DT įvertinys

$$\hat{\sigma}^2 = \frac{SS_E}{n}.$$

Jis paslinktasis, tačiau, remiantis 1.2.2 teorema, poslinkį galima atitaisyti imant įvertinį $s^2 = n\hat{\sigma}^2/(n-m) = SS_E/(n-m)$.

Remiantis 1.2.1 teoremos įrodymu, funkciją L galima perrašyti šitaip:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\{-\frac{1}{2\sigma^2}(SS_E + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{A}^T\boldsymbol{A}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\}. \tag{1.3.2}$$

Tada pagal faktorizacijos kriterijų įsitikiname, kad $\boldsymbol{T} = (SS_E, \hat{\beta}_1, ..., \hat{\beta}_m)^T$ yra pakankamoji statistika. Pertvarkę eksponentės argumentą įsitikiname, kad imties skirstinys priklauso $(m+1)$-mačių eksponentinių skirstinių šeimai ir statistika $\boldsymbol{T}$ ne tik pakankamoji, bet ir pilnoji (žr. I dalies, 3.3 skyrelį). Todėl visos $\boldsymbol{T}$ funkcijos yra savo vidurkių NMD įvertiniai. Pavyzdžiui, $\hat{\beta}_i, \boldsymbol{L}^T\hat{\boldsymbol{\beta}}, s^2$ yra parametrų $\beta_i, \boldsymbol{L}^T\boldsymbol{\beta}, \sigma^2$ įvertiniai, turintys minimalią dispersiją nepaslinktųjų įvertinių klasėje.

Rasime nežinomų parametrų įvertinių skirstinius.

1.3.1 teorema. *Jei* $Rang(\boldsymbol{A}^T\boldsymbol{A}) = m$, *tai įvertiniai* $\hat{\boldsymbol{\beta}}$ *ir* s^2 *yra nepriklausomi ir*

$$\hat{\boldsymbol{\beta}} \sim N_m(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}), \quad \frac{(n-m)s^2}{\sigma^2} = \frac{SS_E}{\sigma^2} \sim \chi^2(n-m). \tag{1.3.3}$$

Be to, $(SS(\boldsymbol{\beta}) - SS_E)/\sigma^2 \sim \chi^2(m)$ *ir a. d.* SS_E *bei* $SS(\boldsymbol{\beta}) - SS_E$ *yra nepriklausomi.*

Įrodymas. Įvertinys $\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$ yra tiesinė normaliojo a. v. $\boldsymbol{Y} \sim N_n(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ funkcija, todėl jis turi normalųjį skirstinį, kurio parametrai nurodyti 1.2.2 teoremoje (žr. 2 priedą 7.4 skyrelį):

$$\hat{\boldsymbol{\beta}} \sim N_m(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}).$$

Kadangi $n \times m$ matricos $\boldsymbol{A}$ rangas yra m, tai egzistuoja ortonormuotų n-mačių vektorių $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_m$ sistema (bazė), kad kiekvienas matricos $\boldsymbol{A}$ stulpelis yra šios sistemos vektorių tiesinė forma (žr. 1 priedą). Tegu $\boldsymbol{F}$ yra $n \times m$ matrica, kurios stulpeliai yra vektoriai $\boldsymbol{F}_i$. Papildykime matricą $\boldsymbol{F}$ eilės $n \times (n-m)$ matrica $\boldsymbol{G}$, kad matrica $\boldsymbol{D} = (\boldsymbol{F}\vdots\boldsymbol{G})$ būtų ortogonali: $\boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{D}^T\boldsymbol{D} = \boldsymbol{I}$. Pagal konstrukciją

$$\boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{I}, \quad \boldsymbol{G}^T\boldsymbol{G} = \boldsymbol{I}, \quad \boldsymbol{F}^T\boldsymbol{G} = \boldsymbol{0}, \quad \boldsymbol{A}^T\boldsymbol{G} = \boldsymbol{0}, \quad \boldsymbol{G}^T\boldsymbol{A} = \boldsymbol{0}. \tag{1.3.4}$$

Nagrinėkime n-matį vektorių :

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{F}^T\boldsymbol{Y} \\ \boldsymbol{G}^T\boldsymbol{Y} \end{pmatrix} = \boldsymbol{D}^T\boldsymbol{Y}. \tag{1.3.5}$$

Vektorius $\boldsymbol{Z}$ turi normalųjį skirstinį ir jo momentai yra

$$\mathbf{E}(\boldsymbol{Z}_1) = \boldsymbol{F}^T\boldsymbol{A}\boldsymbol{\beta}, \quad \mathbf{E}(\boldsymbol{Z}_2) = \boldsymbol{G}^T\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}, \quad \boldsymbol{V}(\boldsymbol{Z}_1) = \sigma^2\boldsymbol{I},$$

$$\boldsymbol{V}(\boldsymbol{Z}_2) = \sigma^2\boldsymbol{I}, \quad \mathbf{Cov}\,(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \boldsymbol{0}. \tag{1.3.6}$$

A. v. $\boldsymbol{Z}$ koordinatės yra normalieji n. a. d., turintys tą pačią dispersiją σ^2 kaip ir pradinio a. v. $\boldsymbol{Y}$ koordinatės. Be to, a. v. $\boldsymbol{Z}_1$ ir $\boldsymbol{Z}_2$ yra nepriklausomi.

Kadangi, atliekant ortonormuotą transformaciją, atstumai tarp taškų lieka nepakitę

$$(\boldsymbol{Z} - \mathbf{E}\boldsymbol{Z})^T(\boldsymbol{Z} - \mathbf{E}\boldsymbol{Z}) = (\boldsymbol{Y} - \mathbf{E}\boldsymbol{Y})^T(\boldsymbol{Y} - \mathbf{E}\boldsymbol{Y}),$$

tai

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{A\beta})^T(\boldsymbol{Y}-\boldsymbol{A\beta}) = (\boldsymbol{Z}_1-\boldsymbol{F}^T\boldsymbol{A\beta})^T(\boldsymbol{Z}_1-\boldsymbol{F}^T\boldsymbol{A\beta})+\boldsymbol{Z}_2^T\boldsymbol{Z}_2. \quad (1.3.7)$$

Minimizuodami kvadratinę formą $SS(\boldsymbol{\beta})$ gauname

$$SS_E = \min_{\boldsymbol{\beta}} SS(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} (\boldsymbol{Z}_1 - \boldsymbol{F}^T\boldsymbol{A\beta})^T(\boldsymbol{Z}_1 - \boldsymbol{F}^T\boldsymbol{A\beta}) + \boldsymbol{Z}_2^T\boldsymbol{Z}_2 = \boldsymbol{Z}_2^T\boldsymbol{Z}_2, \quad (1.3.8)$$

jeigu yra toks $\boldsymbol{\beta}$, kad

$$\boldsymbol{F}^T\boldsymbol{A\beta} = \boldsymbol{Z}_1. \quad (1.3.9)$$

Šioje lygčių sistemoje yra m nežinomųjų, o $m \times m$ matricos $\boldsymbol{F}^T\boldsymbol{A}$ rangas lygus m:

$$m = Rang(\boldsymbol{A}^T) = Rang(\boldsymbol{A}^T(\boldsymbol{F}\vdots\boldsymbol{G})) = Rang(\boldsymbol{A}^T\boldsymbol{F}\vdots\boldsymbol{A}^T\boldsymbol{G}) =$$

$$= Rang(\boldsymbol{A}^T\boldsymbol{F}\vdots\boldsymbol{0}) = Rang(\boldsymbol{A}^T\boldsymbol{F}).$$

Taigi matrica $\boldsymbol{F}^T\boldsymbol{A}$ turi atvirkštinę ir lygčių sistemos (1.3.9) sprendinys vienintelis $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{A})^{-1}\boldsymbol{Z}_1$ (žr. 1 priedą (6.2.22)). Kadangi

$$\frac{SS_E}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{Z}_2^T\boldsymbol{Z}_2,$$

yra $n-m$ vienodai pasiskirsčiusių pagal $N(0,\ 1)$ a. d. kvadratų suma, tai jos skirstinys yra $\chi^2(n-m)$.

Kadangi SS_E yra $\boldsymbol{Z}_2$ funkcija, $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{A})^{-1}\boldsymbol{Z}_1$ yra $\boldsymbol{Z}_1$ funkcija, o $\boldsymbol{Z}_1$ ir $\boldsymbol{Z}_2$ yra nepriklausomi, tai $\hat{\boldsymbol{\beta}}$ ir SS_E taip pat nepriklausomi.

Pagal (1.3.7)

$$SS(\boldsymbol{\beta}) - SS_E = (\boldsymbol{Z}_1 - \boldsymbol{F}^T\boldsymbol{A\beta})^T(\boldsymbol{Z}_1 - \boldsymbol{F}^T\boldsymbol{A\beta}),$$

o pagal (1.3.6) $\boldsymbol{Z}_1 - \boldsymbol{F}^T\boldsymbol{A\beta} \sim N_m(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Todėl a. d. SS_E ir $SS(\boldsymbol{\beta}) - SS_E$ nepriklausomi ir $(SS(\boldsymbol{\beta}) - SS_E)/\sigma^2 \sim \chi^2(m)$. ▲

1.3.1 pastaba. Tiesiniuose modeliuose dažnai tenka tikrinti pavidalo $\beta_{j_1} = \cdots = \beta_{j_k} = 0$ arba $\beta_{j_1} = \cdots = \beta_{j_k}$, $0 \le j_1 < \cdots < j_k \le m$, hipotezes arba dar bendresnes hipotezes.

Jei pažymėsime $H_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, čia 1 yra j_i pozicijoje, tai pirmąją hipotezę galima užrašyti pavidalu $\boldsymbol{H\beta} = \boldsymbol{0}$, kur $\boldsymbol{H}$ yra $k \times m$ matrica, kurios eilutės yra H_i^T, $i = 1, \ldots, k$.

Analogiškai, jei žymėsime $H_i = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T$, čia 1 yra j_1, -1 yra j_i pozicijose $(i = 2, \ldots, k)$, tai antrąją hipotezę taip pat galima užrašyti šitaip $\boldsymbol{H\beta} = \boldsymbol{0}$, čia $\boldsymbol{H}$ – $(k-1) \times m$ matrica, kurios eilutės yra H_i^T $(i = 2, \ldots, k)$.

Apibendrinant, dažnai tenka tikrinti hipotezes, tvirtinančias, kad $\boldsymbol{H\beta} = \boldsymbol{\theta}_0$, čia $\boldsymbol{H}$ yra dimensijos $k \times m$ žinoma matrica ir $\boldsymbol{\theta}_0$ yra žinomas k-matis vektorius.

Tikėtinumų santykis šiai hipotezei tikrinti yra

$$\Lambda = \frac{L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)}{L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}}{\tilde{\sigma}}\right)^n \exp\{-\frac{1}{2}(SS(\tilde{\boldsymbol{\beta}})/\tilde{\sigma}^2 - SS(\hat{\boldsymbol{\beta}})/\hat{\sigma}^2)\},$$

čia $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ yra parametrų $\boldsymbol{\beta}, \sigma^2$ DT įvertiniai, o $\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2$ maksimizuoja L su sąlyga, kad $\boldsymbol{H\beta} = \boldsymbol{\theta}_0$.

Turėjome, kad DT įvertinys $\hat{\boldsymbol{\beta}}$ tenkina sąlygą $SS_E = SS(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} SS(\boldsymbol{\beta})$, $\hat{\sigma}^2 = SS_E/n$. Visiškai analogiškai įvertinys $\tilde{\boldsymbol{\beta}}$ tenkina sąlygą

$$SS(\tilde{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}:\boldsymbol{H\beta}=\boldsymbol{\theta}_0} SS(\boldsymbol{\beta}), \quad \tilde{\sigma}^2 = SS_{EH}/n,$$

čia $SS_{EH} = \min_{\boldsymbol{\beta}:\boldsymbol{H\beta}=\boldsymbol{\theta}_0} SS(\boldsymbol{\beta})$. Taigi

$$\Lambda = \left(\frac{SS_E}{SS_{EH}}\right)^{n/2}.$$

Kritinės srities pavidalas $SS_{EH}/SS_E > c$.

Matome, kad minėtoms hipotezėms tikrinti svarbu rasti ne tik statistikos SS_E, bet ir statistikos SS_{EH} skirstinį.

1.3.2 teorema. *Sakykime, kad* $Rang(\boldsymbol{A}^T\boldsymbol{A}) = m$, *o* $\boldsymbol{H}$ *yra* $k \times m$ *matrica, kurios rangas* $Rang(\boldsymbol{H}) = k \leq m$. *Pažymėkime*

$$SS_{EH} = \min_{\boldsymbol{\beta}:\boldsymbol{H\beta}=\boldsymbol{\theta}_0} (\boldsymbol{Y} - \boldsymbol{A\beta})^T(\boldsymbol{Y} - \boldsymbol{A\beta}) \tag{1.3.10}$$

kvadratinės formos $SS(\boldsymbol{\beta})$ *sąlyginį minimumą, kai* $\boldsymbol{H\beta}$ *lygus žinomam vektoriui* $\boldsymbol{\theta}_0$. *Tada:*

1) SS_E *ir* $SS_{EH} - SS_E$ *yra nepriklausomi;*

2) $SS_E/\sigma^2 \sim \chi^2(n-m)$, $(SS_{EH} - SS_E)/\sigma^2 \sim \chi^2(k;\lambda)$; *necentriškumo parametras* λ *apibrėžiamas lygybe*

$$\lambda = \frac{1}{\sigma^2}(\boldsymbol{H\beta} - \boldsymbol{\theta}_0)^T(\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A}))^{-1}\boldsymbol{H}^T)^{-1}(\boldsymbol{H\beta} - \boldsymbol{\theta}_0); \tag{1.3.11}$$

3) Jeigu hipotezė $\boldsymbol{H\beta} = \boldsymbol{\theta}_0$ *yra teisinga, tai* $\lambda = 0$ *ir* $(SS_{EH} - SS_E)/\sigma^2 \sim \chi^2(k)$, *t. y. santykis*

$$F = \frac{(SS_{EH} - SS_E)(n-m)}{kSS_E} \sim F(k, n-m), \tag{1.3.12}$$

pasiskirstęs pagal Fišerio skirstinį su k *ir* $n-m$ *laisvės laipsnių.*

Įrodymas. Imkime $k \times n$ matricą $\boldsymbol{D}_1 = \boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. Jos eilutės yra tiesinės matricos $\boldsymbol{A}^T$ eilučių formos. Kadangi $k \times m$ matricos $\boldsymbol{H}$ rangas yra k, o $m \times m$ ir $m \times n$ matricų $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ ir $\boldsymbol{A}^T$ rangai yra m, tai jų sandaugos $\boldsymbol{D}_1$ rangas yra k. Taigi matricos $\boldsymbol{D}_1$ eilutės yra k tiesiškai nepriklausomų vektorių, priklausančių matricos $\boldsymbol{A}^T$ eilučių tiesinių darinių erdvei. Šią matricą

papildykime tokia eilės $(m-k)\times n$ matrica $\boldsymbol{D}_2$, kad jos eilutės būtų vienetinio ilgio, ortogonalios tarpusavyje ir ortogonalios matricos $\boldsymbol{D}_1$ eilutėms, o abiejų matricų $\boldsymbol{D}_1$ ir $\boldsymbol{D}_2$ eilutės sudarytų matricos $\boldsymbol{A}^T$ eilučių tiesinių darinių erdvės bazę (žr. 1 priedą). Pagal konstrukciją

$$\boldsymbol{D}_1\boldsymbol{D}_2^T=\boldsymbol{0},\quad \boldsymbol{D}_2\boldsymbol{D}_1^T=\boldsymbol{0},\quad \boldsymbol{D}_2\boldsymbol{D}_2^T=\boldsymbol{I}.$$

Pagaliau parinkime $(n-m)\times n$ matricą $\boldsymbol{D}_3$, kurių eilutės yra vienetinio ilgio, ortogonalios tarpusavyje ir ortogonalios matricų $\boldsymbol{D}_1$ ir $\boldsymbol{D}_2$ eilutėms. Pagal konstrukciją

$$\boldsymbol{D}_3\boldsymbol{D}_1^T=\boldsymbol{0},\quad \boldsymbol{D}_3\boldsymbol{D}_2^T=\boldsymbol{0},\quad \boldsymbol{D}_3\boldsymbol{D}_3^T=\boldsymbol{I},\quad \boldsymbol{D}_3\boldsymbol{A}=\boldsymbol{0}.$$

Sujungę matricų $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ ir $\boldsymbol{D}_3$ eilutes, gausime rango n kvadratinę $n\times n$ matricą $\boldsymbol{D}$, kuri yra erdvės $\mathbf{R}^n$ bazė.

Nagrinėkime a. v.

$$\boldsymbol{Z}=\begin{pmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\\ \boldsymbol{Z}_3\end{pmatrix}=\begin{pmatrix}\boldsymbol{D}_1\boldsymbol{Y}\\ \boldsymbol{D}_2\boldsymbol{Y}\\ \boldsymbol{D}_3\boldsymbol{Y}\end{pmatrix}=\boldsymbol{D}\boldsymbol{Y}. \tag{1.3.13}$$

Atsitiktinio vektoriaus $\boldsymbol{Z}$ komponenčių vidurkiai yra

$$\mathbf{E}(\boldsymbol{Z}_1)=\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{H}\boldsymbol{\beta},\quad \mathbf{E}(\boldsymbol{Z}_2)=\boldsymbol{D}_2\boldsymbol{A}\boldsymbol{\beta},\quad \mathbf{E}(\boldsymbol{Z}_3)=\boldsymbol{D}_3\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{0}. \tag{1.3.14}$$

Pagal parinkimą a. v. $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ ir $\boldsymbol{Z}_3$ yra nekoreliuoti (taigi ir nepriklausomi), o jų kovariacijų matricos

$$\boldsymbol{V}(\boldsymbol{Z}_1)=\sigma^2\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T,\quad \boldsymbol{V}(\boldsymbol{Z}_2)=\sigma^2\boldsymbol{I},\quad \boldsymbol{V}(\boldsymbol{Z}_3)=\sigma^2\boldsymbol{I}. \tag{1.3.15}$$

Remdamiesi atvirkštinės matricos, transponavimo operacijos ir veiksmų su blokinėmis matricomis savybėmis (žr. 1priedą), gauname

$$SS(\boldsymbol{\beta})=(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{\beta})=(\boldsymbol{Z}-\mathbf{E}\boldsymbol{Z})^T(\boldsymbol{D}\boldsymbol{D}^T)^{-1}(\boldsymbol{Z}-\mathbf{E}\boldsymbol{Z})=$$

$$(\boldsymbol{Z}_1-\boldsymbol{H}\boldsymbol{\beta})^T(\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T)^{-1}(\boldsymbol{Z}_1-\boldsymbol{H}\boldsymbol{\beta})+$$

$$+(\boldsymbol{Z}_2-\boldsymbol{D}_2\boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Z}_2-\boldsymbol{D}_2\boldsymbol{A}\boldsymbol{\beta})+\boldsymbol{Z}_3^T\boldsymbol{Z}_3. \tag{1.3.16}$$

Funkcijos $SS(\boldsymbol{\beta})$ minimumas

$$SS_E=\boldsymbol{Z}_3^T\boldsymbol{Z}_3\sim\sigma^2\chi^2_{n-m},$$

jei tik galime parinkti $\boldsymbol{\beta}$ taip, kad pirmieji du dėmenys būtų lygūs 0, t. y. parinkti $\boldsymbol{\beta}$ iš lygčių sistemos

$$\begin{cases}\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{Z}_1,\\ \boldsymbol{D}_2\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{Z}_2.\end{cases} \tag{1.3.17}$$

Šioje sistemoje yra m lygčių ir m nežinomųjų, o matricos prie $\boldsymbol{\beta}$ rangas lygus m:

$$m = Rang(\boldsymbol{A}) = Rang(\boldsymbol{DA}) = Rang \begin{pmatrix} \boldsymbol{D}_1\boldsymbol{A} \\ \boldsymbol{D}_2\boldsymbol{A} \\ \boldsymbol{D}_3\boldsymbol{A} \end{pmatrix} = Rang \begin{pmatrix} \boldsymbol{H} \\ \boldsymbol{D}_2\boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix}, \tag{1.3.18}$$

todėl egzistuoja vienintelis sprendinys $\hat{\boldsymbol{\beta}}$, tenkinantis (žr 1 priedą (6.2.22)) (1.3.17).

Ieškodami sąlyginio $SS(\boldsymbol{\beta})$ minimumo SS_{EH} pažymėsime, kad tereikia minimizuoti antrąjį dėmenį lygybės (1.3.16) dešinėje pusėje, nes kiti dėmenys išlieka pastovūs, imant bet kokias $\boldsymbol{\beta}$ reikšmes, tenkinančias sąlygą $\boldsymbol{H\beta} = \boldsymbol{\theta}_0$. Parodysime, kad antrojo dėmens minimumas lygus 0. Jis virsta nuliu taške $\boldsymbol{\beta}$, tenkinančiame sąlygas

$$\boldsymbol{H\beta} = \boldsymbol{\theta}_0, \quad \boldsymbol{D}_2\boldsymbol{A\beta} = \boldsymbol{Z}_2.$$

Ką tik parodėme, kad tokia sistema turi vienintelį sprendinį (imame $\boldsymbol{\theta}_0$ vietoje $\boldsymbol{Z}_1$ sistemoje (1.3.17)). Taigi

$$SS_{EH} = \min_{\boldsymbol{\beta}:\boldsymbol{H\beta}=\boldsymbol{\theta}_0} SS(\boldsymbol{\beta}) = (\boldsymbol{Z}_1 - \boldsymbol{\theta}_0)^T(\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T)^{-1}(\boldsymbol{Z}_1 - \boldsymbol{\theta}_0) + SS_E;$$

$$SS_{EH} - SS_E = (\boldsymbol{Z}_1 - \boldsymbol{\theta}_0)^T(\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T)^{-1}(\boldsymbol{Z}_1 - \boldsymbol{\theta}_0). \tag{1.3.19}$$

Kadangi $SS_{EH} - SS_E$ išreiškiamas vektoriumi $\boldsymbol{Z}_1$, SS_E – vektoriumi $\boldsymbol{Z}_3$, o $\boldsymbol{Z}_1$ ir $\boldsymbol{Z}_3$ yra n. a. v., tai darome išvadą, kad $SS_{EH} - SS_E$ ir SS_E yra nepriklausomi.

A. v. $\boldsymbol{Z}_1$ turi k-matį normalųjį skirstinį $N_k(\boldsymbol{H\beta},\ \sigma^2\boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A}))^{-1}\boldsymbol{H}^T)$. Pagal daugiamačio normaliojo skirstinio savybes kvadratinė forma $(SS_{EH}-SS_E)/\sigma^2$ turi necentrinį chi kvadrato skirstinį su k laisvės laipsnių ir necentriškumo parametru λ, kurio išraiška yra (1.3.11) formulėje (žr. 2 priedą).

Jeigu $\boldsymbol{H\beta} = \boldsymbol{\theta}_0$, tai necentriškumo parametras $\lambda = 0$ ir $(SS_{EH} - SS_E)/\sigma^2$ turi centrinį chi kvadrato skirstinį $\chi^2(k)$. ▲

1.3.2 pastaba. Kadangi $\boldsymbol{Z}_1 = \boldsymbol{D}_1\boldsymbol{Y} = \boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{H}\hat{\boldsymbol{\beta}}$, tai iš (1.2.3) formulės gauname, kad

$$SS_{EH} - SS_E = (\boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0); \tag{1.3.20}$$

čia $\boldsymbol{\Sigma} = \boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T$.

1.3.2. Pasikliovimo intervalai ir hipotezių tikrinimas

Remdamiesi įvertinių savybėmis bei 1.3.1 ir 1.3.2 teoremomis galime sudaryti parametrų pasikliovimo intervalus ir sukurti kriterijus hipotezėms apie parametrų reikšmes tikrinti.

1) **Dispersija σ^2.** Dispersijos σ^2 NMD įvertinys yra $s^2 = SS_E/(n-m)$. Pagal 1.3.1 teoremą

$$\frac{(n-m)s^2}{\sigma^2} \sim \chi^2(n-m). \tag{1.3.21}$$

Taigi dispersijos σ^2 lygmens $Q = 1-\alpha$ pasikliovimo intervalo $(\underline{\sigma}^2, \overline{\sigma}^2)$ rėžiai yra

$$\underline{\sigma}^2 = SS_E/\chi^2_{\alpha/2}(n-m), \quad \overline{\sigma}^2 = SS_E/\chi^2_{1-\alpha/2}(n-m). \tag{1.3.22}$$

Hipotezė $H_0 : \sigma = \sigma_0$, kai alternatyvos yra $H_1 : \sigma > \sigma_0$ arba $H_2 : \sigma < \sigma_0$, yra atmetamos reikšmingumo lygmens α kriterijumi, kai atitinkamai teisingos nelygybės

$$SS_E > \sigma_0^2\chi^2_\alpha(n-m), \quad SS_E < \sigma_0^2\chi^2_{1-\alpha}(n-m),$$

arba P reikšmių terminais, kai atitinkamai teisingos nelygybės

$$pv = \mathbf{P}\{\chi^2_{n-m} > y\} \le \alpha, \quad pv = \mathbf{P}\{\chi^2_{n-m} < y\} \le \alpha;$$

čia y yra statistikos SS_E/σ_0^2 realizacija.

Kai alternatyva $H_3 : \sigma \ne \sigma_0$ dvipusė, hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, kai

$$SS_E > \sigma_0^2\chi^2_{\alpha/2}(n-m), \quad \text{arba} \quad SS_E < \sigma_0^2\chi^2_{1-\alpha/2}(n-m),$$

arba P reikšmių terminais, kai teisinga nelygybė

$$pv = 2\min(\mathbf{P}\{\chi^2_{n-m} > y\}, \mathbf{P}\{\chi^2_{n-m} < y\}) \le \alpha.$$

2) **Viena tiesinė parametro $\boldsymbol{\beta}$ funkcija.** Sudarykime tiesinės funkcijos $\theta = \boldsymbol{L}^T\boldsymbol{\beta} = L_1\beta_1 + \cdots + L_m\beta_m$ pasikliovimo intervalą. Atskiru atveju gausime bet kurio iš parametrų β_j pasikliovimo intervalą. Įvertinio $\hat\theta = \boldsymbol{L}^T\hat{\boldsymbol{\beta}}$ skirstinys yra normalusis $\hat\theta \sim N(\theta, b^2\sigma^2)$; čia pažymėta $b^2 = \boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{L}$.

Pagal 1.3.1 teoremą a. d. $\hat\theta$ ir s^2 nepriklausomi, todėl, remdamiesi (1.3.3), gauname, kad a. d.

$$\frac{\hat\theta - \theta}{s\,b} \sim S(n-m) \tag{1.3.23}$$

turi Stjudento skirstinį su $n-m$ laisvės laipsnių.

Iš čia gauname parametro θ pasikliovimo intervalą

$$(\underline{\theta}, \bar\theta) = (\hat\theta - b\,s\,t_{\alpha/2}, \hat\theta + b\,s\,t_{\alpha/2}), \tag{1.3.24}$$

kai pasikliovimo lygmuo $Q = 1-\alpha$; čia $t_\alpha = t_\alpha(n-m)$ yra Stjudento skirstinio su $n-m$ laisvės laipsnių α kritinė reikšmė.

Panašiai hipotezė $H_0 : \theta = \theta_0$, kai alternatyvos yra $H_1 : \theta > \theta_0, H_2 : \theta < \theta_0$, $H_3 : \theta \ne \theta_0$, atmetama reikšmingumo lygmens α kriterijumi, kai atitinkamai

$$T > t_\alpha, \quad T < t_\alpha, \quad |T| > t_{\alpha/2}; \tag{1.3.25}$$

čia $T = (\hat\theta - \theta_0)/(s\,b)$. Tegu t yra statistikos T realizacija, o $F(x|n-m)$ – Stjudento skirstinio su $n-m$ laisvės laipsnių pasiskirstymo funkcija. Tada P reikšmių terminais hipotezė H_0 atmetama, kai atitinkamai teisingos nelygybės

$$pv = 1-F(t|n-m) \le \alpha, \quad pv = F(t|n-m) \le \alpha, \quad pv = 2(1-F(|t||n-m)) \le \alpha.$$

1.3.1 pavyzdys (**1.2.1** pavyzdžio tęsinys). Tarkime, kad **1.2.1** pavyzdyje paklaidos $e_1, ..., e_5$ turi normaliuosius skirstinius $N(0,\ \sigma^2)$ su vienodomis dispersijomis σ^2. Rasime dispersijos σ^2 ir parametrų α, β_1, β_2 pasikliovimo intervalus, kai pasikliovimo lygmuo $Q = 0,95$.

Remiantis **1.3.2** teorema

$$\frac{2s^2}{\sigma^2} = \frac{SS_E}{\sigma^2} \sim \chi^2(2)$$

turi χ^2 skirstinį su dviem laisvės laipsniais. Pagal (1.3.22) randame pasikliovimo intervalą

$$(\underline{\sigma^2};\ \overline{\sigma^2}) = (SS_E/\chi^2_{0,025}(2);\ SS_E/\chi^2_{0,975}(2)) = (0,0084;\ 1,2304).$$

Imdami paeiliui $\boldsymbol{L} = (1;\ 0;\ 0)^T, \boldsymbol{L} = (0;\ 1;\ 0)^T, \boldsymbol{L} = (0;\ 0;\ 1)^T$ ir pasinaudoję **1.2.3** pavyzdyje surastomis dispersijų išraiškomis pagal (1.3.24) gauname pasikliovimo intervalus

$$(\underline{\alpha};\ \overline{\alpha}) = (\hat{\alpha} - st_{0,025}/\sqrt{5};\ \hat{\alpha} + st_{0,025}/\sqrt{5}) = (1,7804;\ 2,4596);$$

$$(\underline{\beta_1};\ \overline{\beta}_1) = (\hat{\beta}_1 - st_{0,025}/\sqrt{7};\ \hat{\beta}_1 + st_{0,025}/\sqrt{7}) = (0,5416;\ 1,1156);$$

$$(\underline{\beta_2};\ \overline{\beta}_2) = (\hat{\beta}_2 - st_{0,025}\sqrt{2/7};\ \hat{\beta}_2 + st_{0,025}\sqrt{2/7}) = (2,5804;\ 3,3202).$$

Kelios tiesinės parametro $\boldsymbol{\beta}$ funkcijos. Tarkime, $\boldsymbol{\theta} = \boldsymbol{H\beta}$ yra k-matis vektorius; čia $\boldsymbol{H}$ dimensijos $k \times m$ matrica $(k \le m)$, $Rang(\boldsymbol{H}) = k$.

Vektoriaus $\boldsymbol{\theta}$ įvertinys $\hat{\boldsymbol{\theta}} = \boldsymbol{H}\hat{\boldsymbol{\beta}}$ yra k-matis normalusis vektorius:

$$\hat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T.$$

Vadinasi,

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})/\sigma^2 \sim \chi^2(k), \tag{1.3.26}$$

ir pagal 1.3.1 teoremą ši kvadratinė forma nepriklauso nuo $SS_E = s^2(n-m)$.

Pažymėkime $F_\alpha(k, n-m)$ Fišerio skirstinio su k ir $n-m$ laisvės laipsnių α kritinę reikšmę ir apibrėžkime k-matės erdvės poaibį

$$\boldsymbol{C}(\hat{\boldsymbol{\theta}}, s) = \{\boldsymbol{\theta} : \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{ks^2} < F_\alpha(k, n-m)\}. \tag{1.3.27}$$

Tada $\boldsymbol{C}(\hat{\boldsymbol{\theta}}, s)$ yra vektorinės funkcijos $\boldsymbol{\theta}$ pasikliovimo sritis, kai pasikliovimo lygmuo $Q = 1 - \alpha$:

$$\mathbf{P}_{\boldsymbol{\theta}}\{\boldsymbol{\theta} \in \boldsymbol{C}(\hat{\boldsymbol{\theta}}, s)\} = Q. \tag{1.3.28}$$

Sprendžiant praktinius uždavinius dažnai reikia kurti kriterijus hipotezėms apie kelių tiesinių parametro $\boldsymbol{\beta}$ funkcijų reikšmes. Sakykime, reikia patikrinti hipotezę

$$H_0 : \boldsymbol{H\beta} = \boldsymbol{\theta}_0.$$

Matėme, kad jos atskiri atvejai yra kelių koeficientų lygybės nuliui, kelių koeficientų lygybės ir kitos hipotezės.

Jei teisinga hipotezė H_0, tai remiantis (1.3.20) ir 1.3.2 teorema

$$SS_{EH} - SS_E = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \sigma^2\chi^2_k.$$

Tikėtinumo santykis hipotezei tikrinti yra

$$\Lambda = \left(\frac{SS_E}{SS_{EH}}\right)^{n/2}.$$

Taigi kritinė sritis turėtų būti $SS_E/SS_{EH} < c$ pavidalo, o tai ekvivalentu nelygybei $(SS_{EH} - SS_E)/SS_E > c_1$. Kadangi statistika

$$F = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}{ks^2} = \frac{(SS_{EH} - SS_E)/k}{SS_E/(n-m)} \tag{1.3.29}$$

pasiskirsčiusi pagal Fišerio dėsnį su k ir $n-m$ laisvės laipsniais, jeigu hipotezė H_0 teisinga, tai hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, kai

$$F > F_\alpha(k, n-m), \tag{1.3.30}$$

arba P reikšmių terminais, kai

$$pv = \mathbf{P}\{F_{k,n-m} > f\} \leq \alpha;$$

čia f – statistikos F realizacija. Jei hipotezė nėra teisinga, tai pagal 1.3.2 teoremos 2) teiginį statistika F pasiskirsčiusi pagal necentrinį Fišerio skirstinį su k ir $n-m$ laisvės laipsniais ir necentriškumo parametru

$$\lambda = \frac{1}{\sigma^2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Kriterijaus galia išreiškiama necentrinio Fišerio skirstinio pasiskirstymo funkcija.

1.3.2 pavyzdys (**1.2.4** pavyzdžio tęsinys). Priėmę normalumo prielaidą sudarysime parametro $\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (\alpha - 2\beta_1, \alpha + \beta_1 - \beta_2)^T$ pasikliovimo sritį, kai pasikliovimo lygmuo $Q = 0,95$, ir patikrinsime hipotezę $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0 = \mathbf{0}$.

Parametro $\boldsymbol{\theta}$ įvertis $\hat{\boldsymbol{\theta}}$ ir jo kovariacijų matrica surasti **1.2.4** pavyzdyje

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^T \hat{\boldsymbol{\beta}} \\ \boldsymbol{K}^T \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \hat{\alpha} - 2\hat{\beta}_1 \\ \hat{\alpha} + \hat{\beta}_1 - \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 0,463 \\ 0,034 \end{pmatrix},$$

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \frac{\sigma^2}{35} \begin{pmatrix} 27 & 2 \\ 2 & 17 \end{pmatrix} = \sigma^2 \boldsymbol{\Sigma}.$$

Remiantis (1.3.27) pasikliovimo sritis

$$C(\hat{\boldsymbol{\theta}}, s) = \{\boldsymbol{\theta} : \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{2s^2} < F_{0,05}(2,\ 2)\} =$$

$$= \{\boldsymbol{\theta} : 17(\hat{\theta_1} - \theta_1)^2 - 4(\hat{\theta_1} - \theta_1)(\hat{\theta_2} - \theta_2) + 27(\hat{\theta_2} - \theta_2)^2 < 15,3881\}.$$

Imdami $\boldsymbol{\theta}_0 = \mathbf{0}$ gauname statistikos F iš (1.3.29) realizaciją

$$F = \frac{\hat{\boldsymbol{\theta}}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\theta}}}{2s^2} = \frac{0,2779}{0,0623} = 4,4575.$$

Kadangi P reikšmė $pv = \mathbf{P}\{F_{2;2} > 4,4575\} = 0,1832$, atmesti hipotezę nėra pagrindo.

Hipotezę $H : \boldsymbol{\theta} = \mathbf{0}$ galima patikrinti neskaičiuojant matricos $\boldsymbol{\Sigma}$ ir jos atvirkštinės, o tiesiogiai remiantis 1.3.2 teorema. Kai hipotezė H teisinga, tai $\beta_1 = \alpha/2, \beta_2 = 3\alpha/2$. Taigi turime tiesinį modelį su vienu nežinomu parametru α:

$$Y_1 = 3\alpha/2 + e_1,\ Y_2 = \alpha/2 + e_2,\ Y_3 = 3\alpha/2 + e_3,\ Y_4 = e_4,\ Y_5 = -\alpha/2 + e_5.$$

Parametro α įvertis $\hat{\alpha} = 2$, o sąlyginė liekamoji kvadratinė forma

$$SS_{EH} = (Y_1 - 5)^2 + (Y_2 - 1)^2 + (Y_3 - 5)^2 + y_4^2 + (Y_5 + 1)^2 = 0,34.$$

Statistika F iš (1.3.29), suprantama, įgyja tą pačią reikšmę

$$F = \frac{SS_{EH} - SS_E}{SS_E} = 4,4575.$$

4) **Pasikliovimo intervalų rinkiniai.** Vietoje pasikliovimo srities (1.3.27) kartais pageidautina turėti pasikliovimo intervalų rinkinį, kuris uždengtų visus dominančius parametrus su tikimybe Q.

Naudojantis pasikliovimo sritimi (1.3.27) galima sudaryti iš karto visų tiesinių funkcijų $\boldsymbol{c}^T\boldsymbol{\beta}$, $\boldsymbol{c} \in \mathbf{R}^m$ pasikliovimo intervalus.

Pažymėkime $\mathcal{L}$ aibę, kuri gaunama imant tiesines vektoriaus $\boldsymbol{\beta}$ funkcijas: $\mathcal{L} = \{\boldsymbol{c}^T\boldsymbol{\beta} : \boldsymbol{c} \in \mathbf{R}^m\}$.

1.3.3 teorema. *Tarkime, kad* $Rang(\boldsymbol{A}) = m$.

Tada su tikimybe $Q = 1 - \alpha$ *iš karto visoms funkcijoms* $\boldsymbol{c}^T\boldsymbol{\beta} \in \mathcal{L}$ *galioja nelygybės*

$$\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \delta_\alpha\sqrt{\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}} \le \boldsymbol{c}^T\boldsymbol{\beta} \le \boldsymbol{c}^T\hat{\boldsymbol{\beta}} + \delta_\alpha\sqrt{\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}}, \tag{1.3.31}$$

čia $\boldsymbol{\Sigma} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}$, $\delta_\alpha = s\sqrt{mF_\alpha(m, n-m)}$, *o* $F_\alpha(m, n-m)$ *– Fišerio skirstinio su* m *ir* $n - m$ *laisvės laipsnių* α *kritinė reikšmė.*

Nelygybes (1.3.31) galima traktuoti kaip pasikliovimo intervalus, sudarytus iš karto visoms tiesinėms funkcijoms $\boldsymbol{c}^T\boldsymbol{\beta} \in \mathcal{L}$. *Jeigu imsime vieną funkciją* $\boldsymbol{c}^T\boldsymbol{\beta}$ *(arba keletą tokių funkcijų), tai intervalų pasikliovimo lygmuo ne mažesnis už* Q.

Įrodymas. Pagal Koši ir Švarco nelygybę (žr. 1 priedą) bet kuriems vienodos dimensijos vektoriams $\boldsymbol{U}$ ir $\boldsymbol{V}$ galioja nelygybė

$$(\boldsymbol{U}^T\boldsymbol{V})^2 \le (\boldsymbol{U}^T\boldsymbol{U})(\boldsymbol{V}^T\boldsymbol{V}).$$

Kadangi $\boldsymbol{\Sigma}$ teigiamai apibrėžta simetriška matrica, tai egzistuoja tokia kvadratinė teigiamai apibrėžta matrica $\boldsymbol{B}$, kad $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$. Pritaikę Koši ir Švarco nelygybę vektoriams $\boldsymbol{B}\boldsymbol{U}$ ir $(\boldsymbol{B}^{-1})^T\boldsymbol{V}$, gausime

$$(\boldsymbol{U}^T\boldsymbol{V})^2 = ((\boldsymbol{B}\boldsymbol{U})^T(\boldsymbol{B}^{-1})^T\boldsymbol{V})^2 \le (\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{U})(\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V}),$$

arba

$$\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \ge \frac{(\boldsymbol{U}^T\boldsymbol{V})^2}{\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{U}}, \quad \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} = \sup_{\boldsymbol{U}} \frac{(\boldsymbol{U}^T\boldsymbol{V})^2}{\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{U}}.$$

Supremumas pasiekiamas imant $\boldsymbol{U} = \boldsymbol{\Sigma}^{-1}\boldsymbol{V}$.

Pažymėkime $\boldsymbol{V} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, $\boldsymbol{U} = \boldsymbol{c}$. Taikydami (1.3.27), kai $\boldsymbol{\theta} = \boldsymbol{\beta}$, t. y. kai $\boldsymbol{H}$ yra vienetinė $m \times m$ matrica, gauname

$$1 - \alpha = \mathbf{P}_{\boldsymbol{\beta}}\left\{\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{m\,s^2} < F_\alpha(m, n-m)\right\} =$$

$$\mathbf{P}_{\boldsymbol{\beta}}\left\{\sup_{\boldsymbol{c}} \frac{|\boldsymbol{c}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|}{\sqrt{\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}}} < \delta_\alpha\right\} =$$

$$\mathbf{P}_{\boldsymbol{\beta}}\{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \delta_\alpha\sqrt{\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}} \le \boldsymbol{c}^T\boldsymbol{\beta} \le \boldsymbol{c}^T\hat{\boldsymbol{\beta}} + \delta_\alpha\sqrt{\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}},\ \forall\, \boldsymbol{c} \in \mathbf{R^m}\} = Q.$$

Aišku, kad vienai funkcijai $\boldsymbol{c}^T\boldsymbol{\beta}$ (arba keletui tokių funkcijų) pasikliovimo lygmuo yra ne mažesnis už Q.

Imdami paeiliui $\boldsymbol{c}_1 = (1,\ 0, ..., 0)^T$, $\boldsymbol{c}_2 = (0,\ 1, ..., 0)^T$, ..., $\boldsymbol{c}_m = (0,\ 0, ..., 1)^T$ pagal (1.3.31) gauname parametrų $\beta_1, ..., \beta_m$ pasikliovimo intervalų sistemą ($\underline{\beta}_i$, $\overline{\beta}_i$), $i = 1, ..., m$, kuriai

$$\mathbf{P}_{\boldsymbol{\beta}}\{\underline{\beta}_i < \beta_i < \overline{\beta}_i,\ \forall\ i = 1, ..., m\} \ge Q = 1 - \alpha, \tag{1.3.32}$$

$$\underline{\beta}_i = \hat{\beta}_i - \delta_\alpha\sqrt{b_{ii}}, \quad \overline{\beta}_i = \hat{\beta}_i + \delta_\alpha\sqrt{b_{ii}};$$

čia b_{ij} yra matricos $(\boldsymbol{A}^T\boldsymbol{A})^{-1} = [b_{ij}]_{m\times m}$ elementai.

Tikimybė, kad visi intervalai (1.3.32) uždengs tikrąsias parametrų β_i reikšmes, gali būti kur kas didesnė už $Q = 1 - \alpha$, nes vietoje visų vektorių $\boldsymbol{c} \in \mathbf{R}^m$ imame tik m vektorių $\boldsymbol{c}_1, ..., \boldsymbol{c}_m$.

Jeigu sudarytume pasikliovimo intervalus kiekvienam parametrui β_i pagal (1.3.24):

$$\hat{\beta}_i - s\sqrt{b_{ii}}t_{\alpha/2}(n-m), \quad \hat{\beta}_i + s\sqrt{b_{ii}}t_{\alpha/2}(n-m), \quad i = 1, ..., m, \tag{1.3.33}$$

tai toks intervalų rinkinys nebus ieškomasis, nes tikimybė, kad visi intervalai (1.3.33) uždengs tikrąsias visų parametrų reikšmes, gali būti gerokai mažesnė už $Q = 1 - \alpha$. Pavyzdžiui, jeigu įvertiniai $\hat{\beta}_i, i = 1, ..., m$ yra nepriklausomi, tai intervalai (1.3.33) uždengia visas parametrų reikšmes su tikimybe Q^m. Taigi intervalai (1.3.33) yra trumpesni negu reikėtų. Kita vertus, tikimybė, kad visi intervalai (1.3.32) uždengs tikrąsias parametrų reikšmes, gali būti gerokai didesnė už $Q = 1 - \alpha$, t. y. intervalai (1.3.32) yra ilgesni negu reikėtų.

Kitokį negu (1.3.32) intervalų rinkinio variantą galima gauti naudojant Bonferonio nelygybę. Tegu A_i yra įvykis, kuris reiškia, kad i-asis tipo (1.3.33) intervalas uždengia parametrą β_i ir tegu $\mathbf{P}\{A_i\} = 1 - \alpha_i$. Tada

$$\mathbf{P}\{\cap_{i=1}^m A_i\} = 1 - \mathbf{P}\{\cup_{i=1}^m \bar{A}_i\} \ge 1 - \sum_{i=1}^m \mathbf{P}\{\bar{A}_i\} = 1 - (\alpha_1 + ... + \alpha_m).$$

Jeigu parinksime $\alpha_i = \alpha/m,\ i = 1, ..., m$, tai intervalų rinkinys

$$\hat{\beta}_i - s\sqrt{b_{ii}}t_{\alpha/(2m)}(n-m), \quad \hat{\beta}_i + s\sqrt{b_{ii}}t_{\alpha/(2m)}(n-m), \quad i = 1, ..., m. \tag{1.3.34}$$

uždengs visus parametrus $\beta_1, ..., \beta_m$ su tikimybe ne mažesne už $Q = 1 - \alpha$.

1.3.3 pavyzdys. (**1.2.1** pavyzdžio tęsinys). Priėmę normalumo prielaidą pagal **1.2.1** pratimo duomenis, sudarysime parametrų α, β_1, β_2 pasikliovimo intervalų rinkinius, kad jie uždengtų visus parametrus su tikimybe, ne mažesne už $Q = 0,95$.

Remdamiesi pasikliovimo elipsoidu (1.3.27) ir **1.2.2** pavyzdyje surastomis dispersijos išraiškomis gauname intervalus (1.3.32):

$(\underline{\alpha}; \overline{\alpha}) = (1,5215;\ 2,7185)$, $(\underline{\beta}_1; \overline{\beta}_1) = (0,3228;\ 1,3344)$, $(\underline{\beta}_2; \overline{\beta}_2) = (2,1990;\ 3,6296)$.

Naudodami Bonferonio nelygybę, gauname tokius intervalus:

$(\underline{\alpha}; \overline{\alpha}) = (1,5163;\ 2,7237)$, $(\underline{\beta}_1; \overline{\beta}_1) = (0,3184;\ 1,3388)$, $(\underline{\beta}_2; \overline{\beta}_2) = (2,1927;\ 3,6359)$.

Matome, kad šie intervalai yra kur kas ilgesni už intervalus, sudarytus **1.3.2** pavyzdyje kiekvienam parametrui atskirai.

1.4. Pratimai

1.1. Tarkime, kad turime pilno rango (1.1.2) modelį. Pažymėkime $\hat{\boldsymbol{Y}} = (\hat{Y}_1, ..., \hat{Y}_n)^T = \boldsymbol{A}\hat{\boldsymbol{\beta}}$. Įrodykite, kad

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0, \quad \sum_{i=1}^{n}\hat{Y}_i(Y_i - \hat{Y}_i) = 0.$$

1.2. Tegu $Y_1 = \alpha + e_1$, $Y_2 = 2\alpha - \beta + e_2$, $Y_3 = \alpha + 2\beta + e_3$; čia $\{e_i\}$ nepriklausomi a. d., $\mathbf{E}(e_i) = 0$, $\boldsymbol{V}(e_i) = \sigma^2$. Raskite parametrų α ir β mažiausiųjų kvadratų įvertinius ir jų dispersijas.

1.3. Turime tiesinį modelį

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_2(3x_i^2 - 2), \quad i = 1, 2, 3;$$

čia $x_1 = -1$, $x_2 = 0$, $x_3 = 1$. Raskite parametrų $\beta_0, \beta_1, \beta_2$ įvertinius. Įrodykite, kad parametrų β_0, β_2 įvertiniai modelyje, kuriame $\beta_1 = 0$, turi tą patį pavidalą.

1.4. Parametrams α ir β įvertinti turime stebėjimus: m stebėjimų a. d. Y_1, kurio $\mathbf{E}(Y_1) = \alpha$; m stebėjimų a. d. Y_2, kurio $\mathbf{E}(Y_2) = \alpha + \beta$, ir n stebėjimų a. d. Y_3, kurio $\mathbf{E}(Y_3) = \alpha - 2\beta$. Stebėjimų paklaidos nekoreliuotos ir turi vienodas dispersijas. Įrodykite, kad mažiausiųjų kvadratų įvertiniai $\hat{\alpha}$ ir $\hat{\beta}$ nekoreliuoti, kai $m = 2n$.

1.5. Mažiausiųjų kvadratų metodu parenkami pirmojo ir antrojo laipsnio polinomai pagal didumo n imtį $(X_i,\ Y_i)^T$, $i = 1, 2, ..., n$. Tegu ω ir Ω yra šitokios prielaidos:

$$\omega:\ Y_i = \alpha + \beta X_i + e_i,$$

$$\Omega:\ Y_i = \alpha + \beta X_i + \gamma X_i^2 + e_i';$$

čia e_i, e_i' – n. a. d. su nuliniais vidurkiais ir vienodomis dispersijomis σ^2. Sudarykite normaliųjų lygčių sistemas ir raskite parametrų α, β ir α, β, γ mažiausiųjų kvadratų įvertinius.

1.6 (**1.5** tęsinys). Raskite parametrų įvertinių dispersijas. Įrodykite: jeigu prielaidoje ω vietoje $\alpha + \beta X_i$ imsime $\delta + \beta(X_i - \bar{X})$, tai $\mathbf{Cov}\,(\hat{\delta},\ \hat{\beta}) = 0$.

1.7 (**1.5** tęsinys). Sukurkite kriterijų hipotezei $H: \gamma = 0$ tikrinti, kai a. d. e_i' pasiskirstę pagal normalųjį skirstinį.

1.8. $\hat{\theta}_1, ..., \hat{\theta}_k$ yra parametro θ vienmačiai nepaslinktieji įvertiniai ir $\mathbf{Cov}\,(\hat{\theta}_i,\ \hat{\theta}_j) = \sigma_{ij}$. Raskite tiesinę $\hat{\theta}_1, ..., \hat{\theta}_k$ funkciją, kuri būtų nepaslinktasis θ įvertinys ir turėtų minimalią dispersiją. Raskite tos dispersijos reikšmę.

1.9. (**1.8** tęsinys). Tegu $\mathbf{Cov}\,(\hat{\theta}_i,\ \hat{\theta}_j) = 0$, $i \neq j$, $\boldsymbol{V}\hat{\theta}_i = \sigma_i^2$, $i = 1, ..., k$. Įrodykite, kad $\boldsymbol{V}(c_1\hat{\theta}_1 + ... + c_k\hat{\theta}_k)$, $c_1 + ... + c_k = 1$, yra minimali, kai $c_i = \sigma_i^{-2}/(\sigma_1^{-2} + ... + \sigma_k^{-2})$, ir raskite tos dispersijos reikšmę.

1.10. Įrodykite, kad jeigu $\hat{\theta}_1, ..., \hat{\theta}_k$ yra parametrų $\theta_1, ..., \theta_k$ nepriklausomi NMD įvertiniai, tai $c_1\hat{\theta}_1 + ... + c_k\hat{\theta}_k$ yra parametro $\theta = c_1\theta_1 + ... + c_k\theta_k$ NMD įvertinys.

1.11. $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis a. d. X, kurio $\mathbf{E}X = \mu$ ir $\boldsymbol{V}X = \sigma^2$. Užrašykite stebėjimus kaip tiesinį modelį. Raskite parametro μ mažiausiųjų kvadratų įvertinį.

1.12 (**1.11** tęsinys). Tarkime, kad stebėtas normalusis a. d. $X \sim N(\mu,\ \sigma^2)$. Atlikite 1.3.1 teoremos transformaciją ir raskite SS_E skirstinį. Palyginkite su 1.3.1 skyrelio rezultatais.

1.13. Tegu $\mathbf{X} = (X_1, ..., X_n)^T$ ir $\boldsymbol{Y} = (Y_1, ..., Y_m)^T$ yra paprastosios imtys n. a. d. X ir Y, kurių $\mathbf{E}X = \mu_1$, $\mathbf{E}Y = \mu_2$ ir $\boldsymbol{V}X = \boldsymbol{V}Y = \sigma^2$. Užrašykite stebėjimus (1.1.2) pavidalu kaip tiesinį modelį. Raskite parametrų μ_1, μ_2 mažiausiųjų kvadratų įvertinius.

1.14 (**1.13** tęsinys). Tarkime, kad stebėti normalieji a. d. $X \sim N(\mu_1,\ \sigma^2)$ ir $Y \sim N(\mu_2,\ \sigma^2)$. Atlikite 1.2.1 teoremos transformaciją ir raskite SS_E skirstinį.

1.15 (**1.13** tęsinys). Reikia patikrinti sudėtingąją hipotezę $H : \mu_1 = \mu_2$. Užrašykite šią hipotezę matriciniu pavidalu kaip 1.2.2 teoremoje. Pakartokite 1.2.2 teoremos įrodymą šiuo atveju. Raskite SS_{EH} ir $SS_{EH} - SS_E$.

1.16. Tarkime, kad tiesinio modelio (1.1.2) matrica $\boldsymbol{A}^T\boldsymbol{A} = [a_{ij}]_{(m+1)\times(m+1)}$ neišsigimusi ir diagonaliniai elementai a_{ii}, $i = 1, ..., m+1$ fiksuoti. Įrodykite, kad

(a) parametro β_i įvertinio (1.2.6) dispersija tenkina nelygybę $\boldsymbol{V}\hat{\beta}_i \geq 1/a_{ii}$;

(b) įvertinių $\hat{\beta}_i$ dispersijos minimalios, kai plano matricos A stulpeliai ortogonalūs, t. y. $\boldsymbol{A}^T\boldsymbol{A}$ – diagonalioji matrica.

1.17. Turime m objektų, kurių svoriai $\beta_1, ..., \beta_m$ yra nežinomi. Objektų svoris nustatomas sveriant juos lėkštelinėmis svarstyklėmis dviem būdais.

1) Kiekvienas objektas sveriamas r kartų ir jo svorio įvertiniu imamas gautų rezultatų aritmetinis vidurkis.

2) Sveriant keli objektai dedami ant vienos lėkštelės, keli objektai – ant kitos ir pridedamas svarelis y, kad svarstyklės būtų pusiausviros. Tada k-ajam svėrimui aprašyti turime tiesinį modelį:

$$y_k = a_{k1}\beta_1 + a_{k2}\beta_2 + ... + a_{km}\beta_m + e_k,\ k = 1, ..., n;$$

čia plano matrica $A = [a_{kj}]_{n\times m}$ elementas $a_{kj} = +1,\ -1$ arba 0, atsižvelgiant į tai, ar j-asis objektas padėtas ant kairės, dešinės lėkštutės, arba apskritai nedalyvauja sveriant. Tarkime, kad matavimo paklaidos e abiem svėrimo būdais yra vienodai pasiskirstę n. a. d. su ta pačia dispersija σ^2.

Įrodykite, kad antruoju būdu didžiausio tikslumo pasiekiama, kai plano matricos A elementai yra arba $+1$, arba -1 ir jos stulpeliai ortogonalūs.

1.18 (**1.17** tęsinys). Tarkime, kad reikia įvertinti $m = 4$ objektų svorius $\boldsymbol{V}\hat{\beta}_i = \sigma^2/4$ tikslumu. Tada pirmuoju būdu reikėtų atlikti $mr = 16$ svėrimų. Kiek kartų galima sumažinti svėrimų skaičių antruoju būdu? Raskite tokių minimalaus skaičiaus svėrimų plano matricos A pavidalus.

1.19 (**1.18** tęsinys). Sveriant 4 objektus antruoju būdu gauti dviejų nepriklausomų serijų po 4 svėrimus rezultatai

y_k	a_{k1}	a_{k2}	a_{k3}	a_{k4}		y_k	a_{k1}	a_{k2}	a_{k3}	a_{k4}
20,2	+1	+1	+1	+1		19,9	+1	+1	+1	+1
8,1	+1	-1	+1	-1		8,3	+1	-1	+1	-1
9,7	+1	+1	-1	-1		10,2	+1	+1	-1	-1
1,9	+1	-1	-1	+1		1,8	+1	-1	-1	+1

(a) Raskite parametrų $\beta_1, ..., \beta_4$ įverčius pagal vieno ir kito eksperimento rezultatus. Ar galima pagal tas atskiras eksperimentų serijas įvertinti dispersiją σ^2?

(b) Sujunkite šiuos abu eksperimentus ir įvertinkite parametrus $\beta_1, ..., \beta_4,\ \sigma^2$. Kiek kartų reikėtų padidinti svėrimų skaičių naudojant pirmąjį būdą, kad parametrų $\beta_1, ..., \beta_4$ įvertiniai būtų tokio paties tikslumo?

(c) Tarę, kad paklaidų skirstiniai yra normalieji, palyginkite dispersijos σ^2 įvertinių dispersijas pirmuoju ir antruoju būdu, kai parametrų $\beta_1, ..., \beta_4$ įvertinių tikslumas yra vienodas.

1.20. Nagrinėjamas tiesinis modelis (1.1.2), kai stebėjimų skaičius $n = 100$. Kiek kartų parametro β_i pasikliovimo intervalų (1.3.32), (1.3.34) ilgis yra didesnis už intervalo (1.3.33) ilgį, kai $m = 2,\ 5,\ 10$, o pasikliovimo lygmuo $Q = 0,95$.

1.21. Tarkime, $\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{e}$, $\mathbf{E}(\boldsymbol{e}) = \boldsymbol{0}, \boldsymbol{V}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Lambda}$; čia $\mathrm{Rang}(\boldsymbol{A}^T\boldsymbol{A}) = m$, o $\boldsymbol{\Lambda} = [\lambda_{ij}]_{n\times n}$ – žinoma teigiamai apibrėžta matrica. Raskite parametro $\boldsymbol{\beta}$ mažiausiųjų kvadratų įvertinį ir jo kovariacinę matricą.

1.22 (**1.21** tęsinys). Tegu $Y_i, i = 1, ..., n$, yra nepriklausomi a. d., kurių $\mathbf{E}(Y_i) = \theta$, $\boldsymbol{V}(Y_i) = \sigma^2/\omega_i$; ω_i – žinomi. Raskite parametro θ tiesinį nepaslinktąjį įvertinį su minimalia dispersija. Raskite šios dispersijos išraišką.

1.23 (**1.21** tęsinys). Tegu $Y_1, ..., Y_n$ yra nepriklausomi a.d. ir $Y_i \sim N(i\theta, i^2\sigma^2)$, $i = 1, ..., n$. Raskite parametro θ NMD įvertinį ir įrodykite, kad jo dispersija lygi σ^2/n.

1.24 (**1.21** tęsinys). Tarkime, kad **1.21** pratimo sąlygomis $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda})$. Raskite parametrų $\boldsymbol{\beta}$ ir σ^2 įvertinius ir jų skirstinius.

1.25 (**1.21** tęsinys). Reikia įvertinti skysčio tankį d atliekant nepriklausomus įvairaus tūrio skysčio svėrimus. Tegu Y_i yra gautas tūrio X_i skysčio svoris; $\mathbf{E}(Y_i) = dX_i$, $\boldsymbol{V}(Y_i) = \sigma^2 f(X_i)$, $i = 1, ..., n$. Raskite parametro d mažiausiųjų kvadratų įvertinį, kai a) $f(X_i) = 1$; b) $f(X_i) = X_i$; c) $f(X_i) = X_i^2$.

1.26 (**1.21** tęsinys). Tegu $Y_i = \beta_0 + \beta_1 X_i + e_i$, $i = 1, 2, 3$; $\mathbf{E}(\boldsymbol{e}) = \boldsymbol{0}$, $\boldsymbol{V}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Lambda}$; čia

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

o ρ yra žinomas. Raskite parametrų β_0 ir β_1 mažiausiųjų kvadratų įvertinius ir jų dispersijas.

1.27. Tegu imties elementai $Y_1, ..., Y_n$ aprašomi tiesiniu modeliu su normaliosiomis paklaidomis, turi vienodas dispersijas $\boldsymbol{V}(Y_i) = \sigma^2$ ir vienodas kovariacijas $\mathbf{Cov}\,(Y_i, Y_j) = \rho\sigma^2, i \neq j$. Atliekame ortogonalią tiesinę transformaciją, pervedančią a.v. $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ į vektorių $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$, kai $Z_1 = (Y_1 + ... + Y_n)/n$. Įrodykite, kad vektoriaus $\boldsymbol{Z}_2 = (Z_2, ..., Z_n)^T$ koordinatės nekoreliuotos ir turi vienodas dispersijas $\sigma^2(1-\rho)$. Įrodykite, kad parametrinių funkcijų nepaslinktieji tiesiniai įvertiniai su minimalia dispersija yra tiesinio modelio $\boldsymbol{Z}_2 = \boldsymbol{U\beta} + \boldsymbol{e}, \boldsymbol{e} \sim N_{n-1}(\boldsymbol{0}, \sigma^2(1-\rho)\boldsymbol{I})$ mažiausiųjų kvadratų įvertiniai.

1.28. Tarkime, kad į modelį (1.1.2) įtraukiame papildomai r kovariančių. Tada gauname išplėstą tiesinį modelį

$$\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{B\gamma} + \boldsymbol{e} = (\boldsymbol{A}\dot{:}\boldsymbol{B}) \begin{pmatrix} \boldsymbol{\beta} \\ \cdots \\ \boldsymbol{\gamma} \end{pmatrix} = \boldsymbol{W\delta} + \boldsymbol{e}.$$

Tarę, kad $\mathrm{Rang}(\boldsymbol{W}) = m + r$, gauname išplėstinio modelio parametro $\boldsymbol{\delta}$ įvertinį

$$\hat{\boldsymbol{\delta}} = \begin{pmatrix} \boldsymbol{\beta}^* \\ \cdots \\ \boldsymbol{\gamma}^* \end{pmatrix} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Y},$$

kuris yra sistemos, susidedančios iš $m + r$ lygčių, sprendinys. Įrodykite, kad

a) $\boldsymbol{\beta}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{Y} - \boldsymbol{B\gamma}^*)$, $\boldsymbol{\gamma}^* = (\boldsymbol{B}^T\boldsymbol{RB})^{-1}\boldsymbol{B}^T\boldsymbol{RY}$, $\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$, t.y. galima spręsti dvi sistemas, susidedančias iš m ir r lygčių;

b) $SS_E = (\boldsymbol{Y} - \boldsymbol{B\gamma}^*)^T\boldsymbol{R}(\boldsymbol{Y} - \boldsymbol{R\gamma}^*) = \boldsymbol{Y}^T\boldsymbol{RY} - (\boldsymbol{\gamma}^*)^T\boldsymbol{B}^T\boldsymbol{RY}$;

c) $\boldsymbol{V}(\boldsymbol{\beta}^*) = \sigma^2[(\boldsymbol{A}^T\boldsymbol{A})^{-1} + \boldsymbol{LML}^T]$, $\boldsymbol{V}(\boldsymbol{\gamma}^*) = \sigma^2\boldsymbol{M}$,

$$\mathbf{Cov}\,(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) = -\sigma^2\boldsymbol{LM}; \quad \boldsymbol{L} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}, \quad \boldsymbol{M} = (\boldsymbol{B}^T\boldsymbol{RB})^{-1}.$$

1.29 (**1.28** tęsinys). Įrodykite, kad $\boldsymbol{V}(\hat{\beta}_i) \leq \boldsymbol{V}(\beta_i^*)$.

1.30 (**1.28** tęsinys). Tegu a.d. $Y_1, ..., Y_n$ nepriklausomi ir $Y_i \sim N(\theta, \sigma^2)$. Raskite parametro θ mažiausiųjų kvadratų įvertinį. Remdamiesi **1.27** pratimo rezultatais, raskite parametrų θ ir β įvertinius išplėstiniame modelyje

$$Y_i = \theta + \beta X_i + e_i, \quad e_i \sim N(0, \sigma^2), \quad i = 1, ..., n.$$

1.31 (**1.28** tęsinys). Pateikite parametrų perskaičiavimo algoritmą, kai modelis (1.1.2) išplečiamas nuosekliai pridedant po vieną naują kovariantę.

1.5. Atsakymai ir nurodymai

1.2. $\hat{\alpha}=(Y_1+2Y_2+Y_3)/6$, $\hat{\beta}=(2Y_3-Y_2)/5$, $\boldsymbol{V}\hat{\alpha}=\sigma^2/6, \boldsymbol{V}\hat{\beta}=\sigma^2/5$, $\mathbf{Cov}\,(\hat{\alpha},\hat{\beta})=0$. **1.3.** $\hat{\beta}_0=(Y_1+Y_2+Y_3)/3$, $\hat{\beta}_1=(Y_3-Y_1)/2$, $\hat{\beta}_2=(Y_1-2Y_2+Y_3)/6$. **1.4.** Matricos $\boldsymbol{A}^T\boldsymbol{A}=[c_{ij}]_{2\times 2}$ elementas $c_{12}=c_{21}=m-2n$. **1.5.** Atveju ω parametro $\boldsymbol{\beta}=(\alpha,\beta)^T$ įvertinys $\hat{\boldsymbol{\beta}}$ gaunamas sprendžiant dviejų lygčių sistemą: $\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}}=\boldsymbol{A}^T\boldsymbol{Y}$; čia matrica $\boldsymbol{A}^T\boldsymbol{A}=[a_{ij}]_{2\times 2}$, $a_{11}=n, a_{12}=a_{21}=\sum_i x_i, a_{22}=\sum_i x_i^2$, o vektorius $\boldsymbol{A}^T\boldsymbol{Y}=(\sum_i Y_i,\sum_i Y_i x_i)^T$. Atveju Ω parametro $\boldsymbol{\beta}=(\alpha,\beta,\gamma)^T$ įvertinys $\tilde{\boldsymbol{\beta}}$ gaunamas sprendžiant trijų lygčių sistemą: $\boldsymbol{A}^T\boldsymbol{A}\tilde{\boldsymbol{\beta}}=\boldsymbol{A}^T\boldsymbol{Y}$; čia matrica $\boldsymbol{A}^T\boldsymbol{A}=[a_{ij}]_{3\times 3}$, $a_{13}=a_{31}=\sum_i x_i^2, a_{23}=a_{32}=\sum_i x_i^3$, $a_{33}=\sum_i x_i^4$, o vektorius $\boldsymbol{A}^T\boldsymbol{Y}=(\sum_i Y_i,\sum_i Y_i x_i,\sum_i Y_i x_i^2)^T$. **1.6.** $\sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}$. **1.7.** $SS_E=\sum_i(Y_i-\tilde{\alpha}-\tilde{\beta}x_i-\tilde{\gamma}x_i^2)^2$, $SS_{EH}=\sum_i(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2$. Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai $(SS_{EH}-SS_E)(n-3)/SS_E>F_\alpha(1,n-3)$. **1.8.** Pažymėkime $\hat{\boldsymbol{\theta}}=(\hat{\theta}_1,...,\hat{\theta}_k)^T$ ir $\boldsymbol{\Sigma}^{-1}$ kovariacinės matricos $\boldsymbol{\Sigma}=\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ atvirkštinę matricą. Tada minimalią dispersiją turi tiesinė forma $\delta=\boldsymbol{L}^T\hat{\boldsymbol{\theta}}$, kai $\boldsymbol{L}^T=\mathbf{1}^T\boldsymbol{\Sigma}^{-1}/(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})$; čia $\mathbf{1}^T=(1,\ 1,...,1)$; $\boldsymbol{V}(\boldsymbol{L}^T\hat{\boldsymbol{\theta}})=1/(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})$. **1.9.** $\boldsymbol{V}(c_1\hat{\theta}_1+...+c_k\hat{\theta}_k)=1/(\sigma_1^{-2}+...+\sigma_k^{-2})$. **1.11.** $\hat{\mu}=\bar{X}$. **1.12.** $SS_E=\sum_i(X_i-\bar{X})^2\sim\sigma^2\chi^2_{n-1}$. **1.13.** $\hat{\mu}_1=\bar{X},\hat{\mu}_2=\bar{Y}$. **1.14.** $SS_E=\sum_i(X_i-\bar{X})^2+\sum_j(Y_j-\bar{Y})^2\sim\sigma^2\chi^2_{n+m-2}$. **1.15.** $SS_{EH}=\sum_i(X_i-\bar{Z})^2+\sum_j(Y_j-\bar{Z})^2$, $\bar{Z}=(n\bar{X}+m\bar{Y})/(m+n)$; $SS_{EH}-SS_E=mn(\bar{X}-\bar{Y})^2/(m+n)$. **1.18.** Keturis kartus. **1.19.** a) Pagal pirmo eksperimento rezultatus parametrų įverčiai: $\bar{\beta}_1=(Y_1+Y_2+Y_3+Y_4)/4=9,975$, $\bar{\beta}_2=(Y_1-Y_2+Y_3-Y_4)/4=4,975$, $\bar{\beta}_3=(Y_1+Y_2-Y_3-Y_4)/4=4,175$, $\bar{\beta}_4=(Y_1-Y_2-Y_3+Y_4)/4=1,075$. Analogiškai pagal antro eksperimento rezultatus: $\bar{\beta}_1=10,050$, $\bar{\beta}_2=5,000$, $\bar{\beta}_3=4,050$, $\bar{\beta}_4=0,800$; b) $\bar{\beta}_1=10,0125$, $\bar{\beta}_2=4,9875$, $\bar{\beta}_3=4,1125$, $\bar{\beta}_4=0,9375$, $\hat{\sigma}^2=s^2=0,04875$; keturis kartus; c) $\boldsymbol{V}(\hat{\sigma}^2)=2\sigma^4/8$; taikant pirmąjį būdą reikėtų 32 svėrimų. **1.20.** Kai $m=2$, atitinkamai 1,2525 ir 1,1471 kartų; kai $m=5$, atitinkamai 1,7120 ir 1,3241 kartų; kai $m=10$, atitinkamai 2,2157 ir 1,4486 kartų. **1.21.** Parinkime kvadratinę matricą $\boldsymbol{B}$, kad $\boldsymbol{\Lambda}=\boldsymbol{B}\boldsymbol{B}^T$, ir atlikime transformaciją $\boldsymbol{Z}=\boldsymbol{B}^{-1}\boldsymbol{Y}$. Tada a.v. $\boldsymbol{Z}$ tenkina tiesinį modelį: $\boldsymbol{Z}=\boldsymbol{C}\boldsymbol{\beta}+\boldsymbol{\theta}$; čia $\boldsymbol{C}=\boldsymbol{B}^{-1}\boldsymbol{A}$, o $\boldsymbol{V}(\boldsymbol{\theta})=\sigma^2\boldsymbol{I}$. Gauname $\hat{\boldsymbol{\beta}}=(\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{Z}=(\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{Y}$, $\boldsymbol{V}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{C}^T\boldsymbol{C})^{-1}=\sigma^2(\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{A})^{-1}$. **1.22.** $\hat{\theta}=\sum_i(\omega_i Y_i)/\sum_i\omega_i$, $\boldsymbol{V}(\hat{\theta})=\sigma^2/\sum_i\omega_i$. **1.23.** $\hat{\theta}=[\sum_i(Y_i/i)]/n$, $\boldsymbol{V}(\hat{\theta}=\sigma^2/n$. **1.24.** $\hat{\boldsymbol{\beta}}\sim N_m(\boldsymbol{\beta},\sigma^2(\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{A})^{-1}), \hat{\sigma}^2=s^2=SS_E/(n-m)$, $SS_E=(\boldsymbol{B}^{-1})^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{A}\hat{\boldsymbol{\beta}})\boldsymbol{B}^{-1}\sim\sigma^2\chi^2_{n-m}$. **1.25.** a) $\hat{d}=[\sum_i(X_iY_i)]/\sum_i X_i^2$; b) $\hat{d}=(\sum_i Y_i)/\sum_i X_i$; c) $\hat{d}=[\sum_i Y_i/X_i]/n$. **1.26.** $\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$, $\hat{\beta}_1=\sum_i Y_i(X_i-\bar{X})/\sum_i(X_i-\bar{X})^2$; $\boldsymbol{V}\hat{\beta}_1=\sigma^2(1-\rho)/\sum_i(X_i-\bar{X})^2$, $\boldsymbol{V}\hat{\beta}_0=\sigma^2[(1+2\rho)/3+\bar{X}^2(1-\rho)/\sum_i(X_i-\bar{X})^2]$. **Nurodymas.** Nagrinėkime tiesinį modelį $Y_i=\alpha+\beta_1(X_i-\bar{X})+e_i,\ \ i=1,2,3; \alpha=\beta_0+\beta_1\bar{X}$. Tada

$$[\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{A}]^{-1}=\begin{pmatrix}(1+2\rho)/3 & 0\\ 0 & (1-\rho)/\sum_i(X_i-\bar{X})^2\end{pmatrix},\ \boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{Y}=\begin{pmatrix}\sum_i Y_i/(1+2\rho)\\ \sum_i Y_i(X_i-\bar{X})/(1-\rho)\end{pmatrix}$$

ir lieka pasinaudoti **1.21** pratimo sprendimu.

2 skyrius

Dispersinė analizė

2.1. Vienfaktorė dispersinė analizė

2.1.1. Statistinis modelis

Tarkime, kad a. d. Y skirstinys gali priklausyti nuo tam tikro faktoriaus A, kuris gali būti I skirtingo lygmens $A_1, ..., A_I$. Tegu Y skirstinys, kai faktoriaus lygmuo A_i, yra normalusis su vidurkiu μ_i ir dispersija σ^2.

Tarkime, kad turime I paprastųjų nepriklausomų imčių; i-ją imtį sudaro J_i elementų

$$Y_{i1}, ..., Y_{iJ_i}, \quad i = 1, ..., I,$$

gautų, kai faktoriaus A lygmuo yra A_i. Bendrą elementų skaičių pažymėsime $n = J_1 + \cdots + J_I$. Imčių elementus surašykime į tokią lentelę

2.1.1 lentelė. Imčių elementai

A_i	Y_{ij}	μ_i	$\hat{\mu}_i$
A_1	$Y_{11}, Y_{12}, ..., Y_{1J_1}$	μ_1	$\bar{Y}_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	$Y_{i1}, Y_{i2}, ..., Y_{iJ_i}$	μ_i	$\bar{Y}_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	$Y_{I1}, Y_{I2}, ..., Y_{IJ_I}$	μ_I	$\bar{Y}_{I.}$

A. d. $Y_{ij} \sim N(\mu_i, \sigma^2)$ galima užrašyti tokiu pavidalu:

$$Y_{ij} = \mu_i + e_{ij}, \quad j = 1, ..., J_i, \quad i = 1, ..., I; \tag{2.1.1}$$

čia $\mu_i = \mathbf{E}Y_{ij}$ – nežinomi parametrai, o e_{ij} – nepriklausomi a. d., pasiskirstę pagal normalųjį dėsnį su nuliniu vidurkiu ir vienoda dispersija σ^2. Sujungę Y_{ij} į vieną bendrą vektorių

$$\boldsymbol{Y} = (Y_{11}, ..., Y_{1J_1}, Y_{21}, ..., Y_{2J_2}, ..., Y_{I1}, ..., Y_{IJ_I})^T$$

ir pažymėję $\boldsymbol{\beta} = (\mu_1, ..., \mu_I)^T$ nežinomų parametrų (vidurkių) vektorių, modelį (2.1.1) galima užrašyti matriciniu pavidalu kaip tiesinį modelį:

$$\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{e}, \quad \mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A\beta}, \quad \boldsymbol{V}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}, \quad \boldsymbol{Y} \sim N_n(\boldsymbol{A\beta}, \sigma^2 \boldsymbol{I}).$$

Matrica $\boldsymbol{A}$ turi $n = J_1 + \cdots + J_I$ eilučių ir I stulpelių. Pirmosios J_1 eilutės turi pavidalą $(1,\ 0, ..., 0)$, paskui J_2 eilučių turi pavidalą $(0,\ 1, ..., 0)$, pagaliau paskutinės J_I eilutės turi pavidalą $(0,\ 0, ..., 1)$.

Šiame modelyje matrica $\boldsymbol{A}^T\boldsymbol{A}$ yra diagonalioji su diagonaliniais elementais $J_1, ..., J_I$; jeigu visi $J_i \geq 1$, tai matricos $\boldsymbol{A}$ rangas lygus I.

2.1.2. Mažiausiųjų kvadratų įvertiniai

Mažiausiųjų kvadratų (MK) įvertiniai gaunami minimizuojant kvadratinę formą

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{A\beta})^T(\boldsymbol{Y} - \boldsymbol{A\beta}) = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu_i)^2.$$

Gauname MK įvertinius $\hat{\boldsymbol{\beta}} = (\hat{\mu}_1, ..., \hat{\mu}_I)^T$,

$$\hat{\mu}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} = \bar{Y}_{i\cdot}, \quad i = 1, 2, ..., I,$$

ir liekamąją kvadratinę formą

$$SS_E = SS(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

Čia ir toliau raidė su brūkšniu viršuje ir taškais vietoje indeksų reiškia aritmetinį vidurkį, kai indeksai perbėga visas galimas reikšmes. Pagal 1.2.2 teoremą dispersijos σ^2 nepaslinktasis įvertinys (tariama, kad $n > I$) yra

$$s^2 = \frac{SS_E}{n - I} = \frac{1}{n - I}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

Remdamiesi 1.3.1 skyreliu, galime tvirtinti, kad

$$\frac{(n-I)s^2}{\sigma^2} \sim \chi^2(n - I), \quad \sqrt{J_i}\frac{\hat{\mu}_i - \mu_i}{s} \sim S(n - I),$$

ir bet kuriai tiesinei funkcijai $\theta = \boldsymbol{L}^T\boldsymbol{\beta}$, $\boldsymbol{L} \in \mathbf{R}^I$, įvertinys $\hat{\theta} = \boldsymbol{L}^T\hat{\boldsymbol{\beta}}$ tenkina sąlygą

$$\frac{\hat{\theta} - \theta}{s\, b} \sim S(n - I), \quad b^2 = \boldsymbol{L}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{L},$$

tai analogiškai kaip ir 1.3.2 skyrelyje sudaromi pasikliovimo intervalai ir sukuriami kriterijai hipotezėms dėl parametrų σ^2, μ_i, θ reikšmių tikrinti. Tačiau svarbiausieji dispersinės analizės uždaviniai yra vidurkių $\mu_1, ..., \mu_I$ palyginimo hipotezių tikrinimas.

2.1.3. Vidurkių lygybės hipotezės tikrinimas

Vienas iš svarbiausių vienfaktorės dispersinės analizės uždavinių yra patikrinti hipotezę

$$H_A : \mu_1 = \cdots = \mu_I,$$

kad stebėjimų vidurkiai nepriklauso nuo faktoriaus A lygmenų $A_1, ..., A_I$. Alternatyva $\bar{H}_A : \mu_i \neq \mu_j$ bent vienai porai $i \neq j$. Hipotezę galima užrašyti ekvivalenčia forma

$$H_A : \mu_1 - \mu_2 = \cdots = \mu_1 - \mu_I = 0$$

arba $H_A : \boldsymbol{H\beta} = \mathbf{0}$; čia matrica $\boldsymbol{H}$ turi $I-1$ eilutę ir I stulpelių. Šios matricos i-oji eilutė turi pavidalą $(1,\ 0..., 0, -1,\ 0, ..., 0)$, skaičius -1 yra $(i+1)$-oje pozicijoje. Aišku, kad $Rang(\boldsymbol{H}) = I - 1$.

Remiantis 1.3.2 teorema hipotezė H_A atmetama, kai

$$F_A = \frac{(SS_{EH} - SS_E)/(I-1)}{SS_E/(n-I)} > F_\alpha(I-1, n-I); \tag{2.1.2}$$

čia

$$SS_E = \min_{\boldsymbol{\beta}} SS(\boldsymbol{\beta}) = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2, \quad SS_{EH} = \min_{\boldsymbol{\beta}:\boldsymbol{H\beta}=\mathbf{0}} SS(\boldsymbol{\beta}).$$

Lieka surasti kvadratinės formos $SS_{EH} - SS_E =: SS_A$ išraišką.

Hipotezę suformuluosime kitokiu būdu. Suskaidykime vidurkį μ_i į komponentę, nepriklausančią nuo A_i, ir komponentę, apibūdinančią lygmens A_i poveikį:

$$\mu_i = \bar{\mu}_{.} + (\mu_i - \bar{\mu}_{.}) = \mu + \alpha_i, \quad \mu = \bar{\mu}_{.}, \quad \alpha_i = \mu_i - \bar{\mu}_{.};$$

$$\bar{\mu}_{.} = \frac{1}{n}\sum_{i=1}^{I} J_i\mu_i, \quad \sum_{i=1}^{I} J_i\alpha_i = 0. \tag{2.1.3}$$

Tada modelis užrašomas tokiu pavidalu:

$$Y_{ij} = \mu + \alpha_i + e_{ij}. \tag{2.1.4}$$

Parametrų μ_i, μ, α_i nepaslinktieji įvertiniai yra

$$\hat{\mu}_i = \bar{Y}_{i.}, \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{I} J_i\hat{\mu}_i = \bar{Y}_{..}, \quad \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{Y}_{i.} - \bar{Y}_{..}. \tag{2.1.5}$$

2.1.1 teorema. *Kvadratinė forma SS_A turi pavidalą*

$$SS_A = \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})^2, \tag{2.1.6}$$

o jos skirstinys

$$\frac{SS_A}{\sigma^2} \sim \chi^2(I-1;\lambda), \quad \lambda = \frac{1}{\sigma^2}\sum_{i=1}^{I} J_i\alpha_i^2. \tag{2.1.7}$$

Įrodymas. Hipotezę H_A galime užrašyti ekvivalenčia forma:

$$H_A : \alpha_1 = \cdots = \alpha_I = 0.$$

Kvadratinių formų SS_A ir SS_E išraiškų ieškosime tokiu būdu: tapatybės

$$Y_{ij} - \mu_i = (Y_{ij} - \hat{\mu}_i) + (\hat{\mu} - \mu) + (\hat{\alpha}_i - \alpha_i)$$

abi puses keliame kvadratu, sumuojame pagal i ir j ir gauname

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \mu_i)^2 = SS_E + n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{I} J_i(\hat{\alpha}_i - \alpha_i)^2.$$

Iš čia

$$SS_{EH} = \min_{\boldsymbol{H\beta}=\mathbf{0}} SS(\boldsymbol{\beta}) = \min_{\alpha_i=0} SS(\boldsymbol{\beta}) = SS_E + \sum_{i=1}^{I} J_i(\hat{\alpha}_i)^2,$$

$$SS_A = SS_{EH} - SS_E = \sum_{i=1}^{I} J_i(\hat{\alpha}_i)^2 = \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

Remiantis 1.3.2 teorema, SS_A/σ^2 turi necentrinį chi kvadrato skirstinį (primename, kad šio skirstinio necentriškumo parametras šiuo konkrečiu atveju gaunamas į kvadratinės formos išraišką vietoje a. d. $\bar{Y}_{i.} - \bar{Y}_{..}$ įstačius jų vidurkius α_i):

$$SS_A/\sigma^2 \sim \chi^2(I-1;\lambda), \quad \lambda = \frac{1}{\sigma^2} \sum_{i=1}^{I} J_i \alpha_i^2. \tag{2.1.8}$$

▲

2.1.1 išvada. Statistika F_A užrašoma taip:

$$F_A = \frac{(n-I)\sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})^2}{(I-1)\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2}. \tag{2.1.9}$$

Kai hipotezė H_A teisinga, tai pagal (1.3.29) statistika F_A pasiskirsčiusi pagal Fišerio dėsnį su $I-1$ ir $n-I$ laisvės laipsniais.

Kriterijus hipotezei H_A tikrinti. Hipotezė H_A atmetama reikšmingumo lygmens α kriterijumi, kai

$$F_A > F_\alpha(I-1, n-I); \tag{2.1.10}$$

čia $F_\alpha(I-1, n-I)$ – Fišerio skirstinio α kritinė reikšmė. Kriterijaus galia išreiškiama necentrinio Fišerio skirstinio pasiskirstymo funkcija:

$$\beta(\lambda) = \mathbf{P}\{F_{I-1,n-I;\lambda} > F_\alpha(I-1, n-I)\},$$

čia $F_{I-1,n-I;\lambda}$ – atsitiktinis dydis, turintis necentrinį Fišerio skirstinį su $I-1$ ir $n-I$ laisvės laipsnių ir necentriškumo parametru λ. Įvedus vidutines kvadratų sumas

$$MS_A = \frac{SS_A}{I-1} = \sum_{i=1}^{I} J_i \frac{(\bar{Y}_{i.} - \bar{Y}_{..})^2}{I-1}, \quad MS_E = \frac{SS_E}{n-I} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{(Y_{ij} - \bar{Y}_{i.})^2}{n-I},$$

statistika F_A yra tiesiog jų santykis

$$F_A = \frac{MS_A}{MS_E}. \tag{2.1.11}$$

Remiantis (1.3.10) $SS_A/\sigma^2 \sim \chi^2(I-1,\lambda)$, todėl $\mathbf{E}(SS_A) = (I-1+\lambda)\sigma^2$. Gauname

$$\mathbf{E}(MS_A) = \sigma^2 + \sigma_A^2, \quad \mathbf{E}(MS_E) = \sigma^2;$$

čia

$$\sigma_A^2 = \frac{\sigma^2}{I-1}\lambda = \frac{1}{I-1}\sum_{i=1}^{I} J_i \alpha_i^2.$$

Taigi $\mathbf{E}(MS_A) = \mathbf{E}(MS_E) = \sigma^2$, kai teisinga hipotezė H_A, bet $\mathbf{E}(MS_A) > \mathbf{E}(MS_E)$, kai ji neteisinga. Kuo daugiau vidurkiai $\mu_i = \mu + \alpha_i$ skiriasi tarpusavyje, tuo didesnis santykis

$$\frac{\mathbf{E}(MS_A)}{\mathbf{E}(MS_E)} = 1 + \frac{\sigma_A^2}{\sigma^2}.$$

Taigi santykis $F_A = MS_A/MS_E$ turi tendenciją įgyti tuo didesnes reikšmes, kuo didesni skirtumai tarp vidurkių μ_i. Todėl pateiktoji hipotezės atmetimo taisyklė natūrali. Šios taisyklės taikymą pagrįsime dar vienu būdu.

Visų duomenų sklaida

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{..})^2$$

apie bendrą empirinį vidurkį $\bar{Y}_{..}$ vadinama *pilnąją kvadratų suma* (angl. *total sum of squares*) ir gali būti išskaidyta į du dėmenis

$$SS_T = SS_E + SS_A = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

SS_E yra stebėjimų Y_{ij} sklaidų apie savo empirinius vidurkius $\bar{Y}_{i.}$ kiekvienoje imtyje suma ir vadinama *likutine arba paklaidų kvadratų suma* (angl. *residual or error sum of squares*). Šios sklaidos priežastis – atsitiktinės paklaidos e_{ij}.

SS_A yra stebėjimų empirinių vidurkių $\bar{Y}_{i.}$, atitinkančių įvairius faktoriaus lygmenis, sklaida apie bendrą empirinį vidurkį $\bar{Y}_{..}$ ir vadinama *faktoriumi A apibūdinama kvadratų suma* (angl. *factor A attributed sum of squares*).

Jei vidurkių lygybės hipotezė yra teisinga, tai sumos SS_A, apibūdinančios sklaidą tarp grupių, dalis visoje sumoje SS_T turėtų būti maža, o sumos SS_E, apibūdinančios sklaidą grupių viduje, turėtų būti didelė. Atvirkščiai, kuo vidurkiai labiau skiriasi, tuo SS_A dalis turėtų būti didesnė. Taigi natūralu atmesti hipotezę, kai santykis SS_A/SS_E didelis.

Skaičiavimo rezultatai paprastai pateikiami tokioje lentelėje

2.1.2 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	$MS = SS/\nu$	$\mathbf{E}(MS)$
A	$SS_A = \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$I-1$	MS_A	$\sigma^2 + \sigma_A^2$
E	$SS_E = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2$	$n-I$	MS_E	σ^2
T	$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{..})^2$	$n-1$	–	–

Jeigu stebėjimai neprieštarauja prielaidai H_A apie vidurkių $\mu_1, ..., \mu_I$ lygybę, analizę galima ir užbaigti. Tokiu atveju visus stebėjimus galime sujungti į vieną didumo n imtį, gautą stebint normalųjį a. d. su bendru vidurkiu μ ir dispersija σ^2.

2.1.4. Kontrastų analizė

Jei vidurkių lygybės hipotezė atmetama, tai natūraliai kyla klausimas, kaip stebėjimai priklauso nuo faktoriaus A lygmenų. Pavyzdžiui, gal galima suskaidyti faktorių lygmenis į grupes taip, kad skirtumas tarp vidurkių būtų nulemtas skirtumų tarp šių faktoriaus A lygmenų grupių, o viduje grupių vidurkiai skirtųsi nereikšmingai. Suskaidant faktorių lygmenis į grupes dažnai gali padėti turima papildoma informacija.

Jeigu papildomos informacijos nepakanka, tai galima taikyti statistinius stebėjimų prie įvairių faktorių lygmenų palyginimo metodus. Dažniausiai taikomi *daugialypio palyginimo S ir T metodai*, pasiūlyti Šefės ir Tjukio (žr.[14]).

2.1.1 apibrėžimas. Parametrų $\mu_1, ..., \mu_I$ *kontrastu* vadinama tiesinė funkcija

$$\psi = \sum_{i=1}^{I} c_i\mu_i, \quad \sum_{i=1}^{I} c_i = 0, \quad c_i \in \mathbf{R}, \quad i = 1, ..., I. \tag{2.1.12}$$

Naudojant kontrastus, galima patikrinti vidurkių tiesinių darinių lygybės nuliui hipotezes, atskiru atveju palyginti grupių, atitinkančių kai kuriuos faktoriaus A lygmenis, vidurkius.

Hipotezė $H_A : \mu_1 = \cdots = \mu_I$ ekvivalenti teiginiui, kad visi kontrastai ψ lygūs 0. Jei ši hipotezė neteisinga, atsiras kontrastų, kurie nelygūs nuliui. Pavyzdžiui, jei kontrastas $\psi = \mu_1 - \mu_2 \neq 0$, o kontrastai $\mu_1 - \mu_i = 0$, $i = 3, \ldots, I$, tas reiškia, kad hipotezė H_A nėra teisinga ir priežastis ta, kad antrasis vidurkis skiriasi nuo kitų.

Visų kontrastų erdvę žymėsime $\mathcal{L}$.

Tiesinės funkcijos ψ mažiausiųjų kvadratų įvertinys

$$\hat{\psi} = \sum_{i=1}^{I} c_i \hat{\mu}_i = \sum_{i=1}^{I} c_i \bar{Y}_{i\cdot}, \quad \boldsymbol{V}(\hat{\psi}) = \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{J_i}.$$

Dispersijos įvertinys

$$\hat{\boldsymbol{V}}(\hat{\psi}) = s^2 \sum_{i=1}^{I} \frac{c_i^2}{J_i}, \quad s^2 = \frac{SS_E}{n-I} = MS_E. \tag{2.1.13}$$

2.1.2 teorema. (S metodas). *Hipotezė $H_A : \mu_1 = \cdots = \mu_I$ atmetama reikšmingumo lygmens α kriterijumi (2.1.10) tada ir tik tada, kai egzistuoja kontrastas ψ, kad intervalas*

$$(\hat{\psi} - \Delta\sqrt{\hat{\boldsymbol{V}}(\hat{\psi})},\ \hat{\psi} + \Delta\sqrt{\hat{\boldsymbol{V}}(\hat{\psi})}) \tag{2.1.14}$$

neuždengia 0; čia $\Delta^2 = (I-1)F_\alpha(I-1,\ n-I)$.

Įrodymas. Pažymėkime $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_I)^T$. Tada hipotezę $H_A : \mu_1 = \cdots = \mu_I$ galima užrašyti ekvivalenčia forma $H_A : \mu_1 - \mu_2 = \ldots = \mu_1 - \mu_I = 0$ arba $H_A : \boldsymbol{H\mu} = \boldsymbol{0}$; čia $\boldsymbol{H}$ – matrica, turinti $I-1$ eilutę ir I stulpelių, kurios i-oji eilutė turi pavidalą $(1, 0, \ldots, 0, -1, 0, \ldots, 0)$; skaičius -1 yra $(i+1)$-oje pozicijoje. Taigi $Rang(\boldsymbol{H}) = I-1$.

Remiantis 1.3.2 teorema hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai

$$F_A > F_\alpha(I-1, n-I), \tag{2.1.15}$$

čia

$$F_A = \frac{(SS_{EH} - SS_E)(n-I)}{SS_E(I-1)} \sim F(I-1, n-I).$$

Pagal (1.3.19) formulę statistiką F galima užrašyti pavidalu

$$F_A = \frac{\hat{\boldsymbol{\theta}}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\theta}}}{(I-1)s^2}, \quad \hat{\boldsymbol{\theta}} = \boldsymbol{H}\hat{\boldsymbol{\mu}}, \quad \boldsymbol{\Sigma} = \boldsymbol{H}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{H}^T.$$

Taigi hipotezė priimama, kai

$$\frac{\hat{\boldsymbol{\theta}}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\theta}}}{(I-1)s^2} \le F_\alpha(I-1,\ n-I). \tag{2.1.16}$$

Įvykiui {egzistuoja kontrastas ψ, kad intervalas (2.1.14) neuždengia 0} priešingas įvykis yra

$$\hat{\boldsymbol{\psi}} - \Delta\sqrt{\hat{\boldsymbol{V}}(\hat{\psi})} \le 0 \le \hat{\psi} + \Delta\sqrt{\hat{\boldsymbol{V}}(\hat{\psi})}, \quad \forall\, \psi \in \mathcal{L}. \tag{2.1.17}$$

Visų vektorių $\boldsymbol{c} = (c_1, \ldots, c_I)^T$, $\sum_{i=1}^{I} c_i = 0$ erdvės matavimas yra $I-1$, todėl matricos $\boldsymbol{H}$ eilutės sudaro tokių vektorių erdvės bazę. Taigi bet kuris kontrastas

$\boldsymbol{\psi} = \boldsymbol{c}^T\boldsymbol{\mu}$ užrašomas pavidalu $\psi = \boldsymbol{d}^T\boldsymbol{H}\boldsymbol{\mu} = \boldsymbol{d}^T\boldsymbol{\theta}$, o jo nepaslinktasis įvertinys $\hat{\psi} = d^T\boldsymbol{H}\hat{\boldsymbol{\mu}} = d^T\hat{\boldsymbol{\theta}}$. Kadangi matrica $\boldsymbol{A}^T\boldsymbol{A}$ yra diagonalioji su įstrižainės elementais $J_1, \ldots, J_I$, tai

$$\boldsymbol{d}^T\boldsymbol{\Sigma}\boldsymbol{d} = \boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c} = \sum_{i=1}^{I}\frac{c_i^2}{J_i} = \hat{\boldsymbol{V}}(\hat{\boldsymbol{\psi}})/s^2 \quad \Longrightarrow \quad \hat{\boldsymbol{V}}(\hat{\boldsymbol{\psi}}) = s^2\boldsymbol{d}^T\boldsymbol{\Sigma}\boldsymbol{d}.$$

Taigi įvykis (2.1.17) užrašomas bet kuriuo iš toliau užrašytų ekvivalenčių pavidalų:

$$\boldsymbol{d}^T\hat{\boldsymbol{\theta}} - \Delta s\sqrt{\boldsymbol{d}^T\boldsymbol{\Sigma}\boldsymbol{d}} \le 0 \le \boldsymbol{d}^T\hat{\boldsymbol{\theta}} + \Delta s\sqrt{\boldsymbol{d}^T\boldsymbol{\Sigma}\boldsymbol{d}}, \quad \forall\, \boldsymbol{d} \in \mathbf{R}^{I-1},$$

$$|d^T\hat{\boldsymbol{\theta}}| \le \Delta s\sqrt{\boldsymbol{d}^T\boldsymbol{\Sigma}\boldsymbol{d}}, \quad \forall\, \boldsymbol{d} \in \mathbf{R}^{I-1},$$

$$d^T(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T - \Delta^2 s^2\boldsymbol{\Sigma})\boldsymbol{d} \le 0, \quad \forall\, \boldsymbol{d} \in \mathbf{R}^{I-1},$$

$$\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T \le \Delta^2 s^2\boldsymbol{\Sigma}. \tag{2.1.18}$$

Rašome $\boldsymbol{A} \le \boldsymbol{B}$, jeigu matrica $\boldsymbol{A} - \boldsymbol{B}$ neteigiamai apibrėžta.

Kadangi $\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\theta}} > 0$ beveik tikrai, tai padauginę abi (2.1.18) nelygybės puses iš $\hat{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}$ iš kairės bei iš $\hat{\boldsymbol{\theta}}$ iš dešinės, o paskui padaliję abi puses iš $\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\theta}}$, gausime, kad (2.1.18) nelygybė ekvivalenti nelygybei

$$\hat{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}} \le \Delta^2 s^2,$$

o tai ekvivalentu (2.1.16). Kartu ir (2.1.17) ekvivalentu (2.1.16). ▲

S metodas remiasi 2.1.2 teorema: jei hipotezė H_A atmetama, tai ieškoma tų kontrastų, kurie atsakingi už hipotezės atmetimą, t. y. kuriems (2.1.14) intervalas nepadengia nulio.

2.1.1 pastaba. Jeigu kontrastą ψ padauginsime iš konstantos $k \ne 0$, tai $\hat{\boldsymbol{V}}(k\hat{\psi}) = k^2\hat{\boldsymbol{V}}(\hat{\psi})$. Todėl $0 \in (\hat{\psi} - \Delta\sqrt{\hat{\boldsymbol{V}}(\hat{\psi})},\ \hat{\psi} + \Delta\sqrt{\hat{\boldsymbol{V}}(\hat{\psi})}) \iff 0 \in (k\hat{\psi} - \Delta\sqrt{\hat{\boldsymbol{V}}(k\hat{\psi})},\ k\hat{\psi} + \Delta\sqrt{\hat{\boldsymbol{V}}(k\hat{\psi})})$. Kadangi $\boldsymbol{V}(\hat{\psi})$ turi pavidalą $\sigma^2 C$ (čia C – konstanta), tai parinkus $k = 1/\sqrt{C}$ galima pasiekti, kad $\boldsymbol{V}(k\hat{\psi}) = \sigma^2$. Taigi vietoje visų kontrastų klasės $\mathcal{L}$ pakanka nagrinėti *normuotų* kontrastų klasę $\mathcal{L}'$, t. y. aibę tokių kontrastų, kurių $\boldsymbol{V}(\hat{\psi}) = \sigma^2$. Gavome, kad *kriterijus* (2.1.10) *atmeta hipotezę* H_A *tada ir tik tada, kai egzistuoja kontrastas* $\psi \in \mathcal{L}'$, *kad* (2.1.14) *intervalas neuždengia* 0.

Pažymėkime

$$\hat{\psi}_{\max} = \max\{|\hat{\psi}| : \psi \in \mathcal{L}'\}.$$

2.1.3 teorema. *Kvadratinė forma* $SS_A = SS_{EH} - SS_E$ *iš 2.1.2 teoremos beveik tikrai tenkina lygybę*

$$SS_A = \hat{\psi}^2_{\max}. \tag{2.1.19}$$

Įrodymas. Teoremą 2.1.2 galima performuluoti šitaip: su bet kuriuo $\alpha \in (0,1)$

$$F_A = \frac{SS_A}{(I-1)s^2} > F_\alpha(I-1, n-I) \quad \Longleftrightarrow \quad \frac{\hat{\psi}_{\max}}{s\Delta} > 1$$

$$\Longleftrightarrow \quad \frac{\hat{\psi}^2_{\max}}{(I-1)s^2} > F_\alpha(I-1, n-I).$$

Kai α kinta nuo 0 iki 1, tai $x = F_\alpha(I-1, n-I)$ kinta nuo 0 iki ∞. Taigi neneigiamų a.d. $SS_A/(s^2(I-1))$ ir $\hat{\psi}^2_{\max}/(s^2(I-1))$ skirstiniai sutampa. Tada sutampa ir a.d. SS_A ir $\hat{\psi}^2_{\max}$ skirstiniai.

▲

2.1.2 pastaba. Nenuliniai poriniai kontrastai, turintys pavidalą $\mu_i - \mu_j$, geriau pastebimi naudojant vadinamąjį *Tjukio kriterijų* (T metodą). Šis kriterijus naudojamas, kai imčių didumai J_i vienodi.

Tarkime, kad $J_i = J$ su visais $i = 1, \ldots, I$. Metodas grindžiamas pasikliovimo intervalais kontrastams $\mu_i - \mu_j$.

Pažymėkime $V_i = \sqrt{J}(\hat{\mu}_i - \mu_i)$, $V_{(I)} = \max V_i$, $V_{(1)} = \min V_i$. Turėjome, kad

$$V_i/\sigma \sim N(0,1), \quad \nu\, s^2/\sigma^2 \sim \chi^2(\nu), \quad \nu = n - I.$$

A. d. $V_1/\sigma, \ldots, V_I/\sigma$ ir $\nu\, s^2/\sigma^2$ nepriklausomi ir jų skirstiniai nuo nežinomų parametrų nepriklauso. Todėl a. d. $(V_{(I)} - V_{(1)})/\sigma$ skirstinys nepriklauso nuo nežinomų parametrų, be to, a. d. $V_{(I)} - V_{(1)}$ ir s nepriklausomi. Taigi santykio

$$T(I, \nu) = \frac{V_{(I)} - V_{(1)}}{s}$$

skirstinys nepriklauso nuo nežinomų parametrų. Šis santykis vadinamas *stjudentizuotu imties pločiu*. Jo α kritinę reikšmę žymėsime $T_\alpha(I, \nu)$.

2.1.4 teorema. (T metodas). Tikimybė, kad su visais $i, j = 1, \ldots, I$ teisingos nelygybės

$$\hat{\mu}_i - \hat{\mu}_j - sT_\alpha(I,\nu)/\sqrt{J} \le \mu_i - \mu_j \le \hat{\mu}_i - \hat{\mu}_j + sT_\alpha(I,\nu)/\sqrt{J} \qquad (2.1.20)$$

lygi $1 - \alpha$.

Įrodymas. Su visais $i, j = 1, \ldots, I$ nelygybės (2.1.20) teisingos tada ir tik tada, kai

$$\frac{\sqrt{J}\ |\ \hat{\mu}_i - \mu_i - (\hat{\mu}_j - \mu_j)\ |}{s} \le T_\alpha(I,\ \nu) \iff \frac{|\ V_i - V_j\ |}{s} \le T_\alpha(I,\ \nu).$$

Visos šios nelygybės ekvivalenčios vienai nelygybei

$$\frac{V_{(I)} - V_{(1)}}{s} \le T_\alpha(I, \nu),$$

o ji teisinga su tikimybe $1-\alpha$.

▲

2.1.2 išvada. Jei teisinga hipotezė $H_A : \mu_1 = \cdots = \mu_I$, tai tikimybė, kad su visais $i, j = 1, \ldots, I$ intervalai

$$(\hat{\mu}_i - \hat{\mu}_j - sT_\alpha(I,\nu)/\sqrt{J},\ \hat{\mu}_i - \hat{\mu}_j + sT_\alpha(I,\nu)/\sqrt{J}) \tag{2.1.21}$$

padengia 0, lygi $1-\alpha$.

2.1.3 pastaba. Remiantis 2.1.4 teorema ir 2.1.2 išvada galima naudoti tokį kriterijų hipotezei H_A tikrinti: hipotezė atmetama reikšmingumo lygmens α kriterijumi, jei egzistuoja porinis kontrastas $\mu_i - \mu_j$, kuriam intervalas (2.1.21) neuždengia 0. Šiuo kriterijumi ir grindžiamas *Tjukio* arba T metodas: *ieškomi poriniai kontrastai $\mu_i - \mu_j$, kurie atsakingi už hipotezės atmetimą, t. y. intervalai* (2.1.21) *neuždengia* 0.

2.1.4 pastaba. T metodą galima apibendrinti ir visiems, ne vien poriniams kontrastams, bet neporiniams kontrastams naudotinas S metodas, nes jį naudojant gaunami trumpesni pasikliovimo intervalai, negu gauti T metodu.

2.1.1 pavyzdys. Lentelėje pateikti aštuonių vadų (faktorius A) paršiukų svoriai svarais [14].

2.1.3 lentelė. Statistiniai duomenys

A_i	Y_{ij}	$\bar{Y}_{i\cdot}$	s_i^2
A_1	2,0 2,8 3,3 3,2 3,8 1,6 3,6 1,9 3,3 2,8	2,83	0,58
A_2	3,3 3,6 2,6 3,1 3,2 3,3 2,9 3,4 3,2 3,2	3,18	0,08
A_3	3,2 3,3 3,2 2,9 3,3 2,5 2,6 2,8	2,98	0,10
A_4	3,5 2,8 3,2 3,5 2,3 2,4 2,0 1,6	2,66	0,50
A_5	2,6 2,6 2,9 2,0 2,0 2,1	2,37	0,15
A_6	3,1 2,9 3,1 2,5	2,90	0,08
A_7	2,6 2,2 2,5 1,2 1,2	1,94	0,48
A_8	2,5 2,4 3,0 1,5	2,35	0,39

1) priėmę normalumo prielaidą atliksime dispersinę analizę ir reikšmingumo lygmens $\alpha = 0,05$ kriterijumi patikrinsime hipotezę $H_A : \mu_1 = \cdots = \mu_8$;

2) tarkime, kad yra žinoma tokia papildoma informacija: pirmosios trys vados gautos iš vienos paršiavedės, o likusios penkios – iš kitos. Tada natūralu tikrinti prielaidą, kad skirtumą tarp vidurkių lėmė tai, kad buvo dvi skirtingos paršiavedės, t.y. patikrinti hipotezę $H'_A : \mu_1 = \mu_2 = \mu_3;\ \mu_4 = \cdots = \mu_8$.

1) Atlikę skaičiavimus gauname dispersinės analizės lentelę

2.1.4 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS
A	7,247	7	1,035
E	14,106	47	0,300
T	21,353	54	-

Statistika F_A įgyja reikšmę

$$F_A = \frac{MS_A}{MS_E} = 3,45.$$

Kadangi P-reikšmė $pv = \mathbf{P}\{F_{7,\,47} > 3,45\} = 0,0046$, tai hipotezė $H_A : \mu_1 = \cdots = \mu_8$ atmetama kriterijumi, kai reikšmingumo lygmuo $\alpha > 0,0046$.

Remiantis 2.1.2 teorema egzistuoja toks kontrastas $\psi = \sum_i c_i\mu_i$, kad reikšmingumo lygmens $Q = 0,95$ pasikliovimo intervalas $(\underline{\psi}, \bar{\psi})$ neuždengio nulio. Imkime kontrastą, kuriame

$c_2 = 1, c_6 = -1, c_i = 0, i \neq 2,\ 6$. Gauname $\hat{\psi} = 1,24$, $\hat{\boldsymbol{V}}(\hat{\psi}) = 0,09$, ir pasikliovimo intervalą $(\underline{\psi}; \bar{\psi}) = (0,06; 2,42)$. Matome, kad gautasis intervalas nedengia nulio.

2) Kvadratinės formos $SS(\boldsymbol{\mu})$ sąlyginis minimumas, kai teisinga H'_A yra

$$SS'_A = \sum_{i=1}^{3} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}^{(1)})^2 + \sum_{i=4}^{8} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}^{(2)})^2;$$

$$\bar{Y}_{\cdot\cdot}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{3}\sum_{j=1}^{J_i} Y_{ij},\ \ \bar{Y}_{\cdot\cdot}^{(2)} = \frac{1}{n_2}\sum_{i=4}^{8}\sum_{j=1}^{J_i} Y_{ij},$$

$$n_1 = J_1 + J_2 + J_3, \quad n_2 = J_4 + \cdots + J_8.$$

Apskaičiavę gauname $SS'_A = 3,148$, $MS'_A = SS'_A/6 = 0,5247$. Jeigu H'_A teisinga, tai statistika $F'_A = MS'_A/MS_E$ turi Fišerio skirstinį su 6 ir 47 laisvės laipsniais. Gauname $F'_A = 1,749$. Kadangi $pv = \mathbf{P}\{F_{6,\,47} > 1,749\} = 0,1305$, tai atmesti hipotezę H'_A nėra pagrindo.

2.2. Dvifaktorė dispersinė analizė

2.2.1. Statistinis modelis

Tarkime, a. d. Y skirstinys gali priklausyti nuo faktoriaus A, kurio lygmenys yra $A_1, ..., A_I$, ir nuo faktoriaus B, kurio lygmenys yra $B_1, ..., B_J$. Pavyzdžiui, Y gali reikšti kviečių derlingumą, faktorius A – kviečių veislę, o faktorius B – jų auginimo metodiką.

Tarkime, kad stebėjimai atliekami imant visus galimus skirtingus faktorių lygmenų rinkinius (A_i, B_j), be to, kiekvienu atveju matavimai kartojami vienodą skaičių kartų $K > 1$. Toks eksperimentų planas vadinamas *visiškai subalansuotu planu*. Jo analizė kur kas paprastesnė negu tuo atveju, kai stebėjimų skaičiai K_{ij}, gauti esant faktorių lygmenų rinkiniui $(A_i,\ B_j)$, yra skirtingi.

Imties elementus žymėsime $Y_{ijk},\ i = 1, ..., I,\ j = 1, ..., J,\ k = 1, ..., K$; i – faktoriaus A lygmens numeris; j – faktoriaus B lygmens numeris; k – kartotinumo numeris; visų stebėjimų skaičius $n = IJK$.

Dvifaktorės dispersinės analizės modelis. Sakykime, kad a. d. Y_{ijk} aprašomi taip:

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, ..., I, \quad j = 1, ..., J, \quad k = 1, ..., K, \tag{2.2.1}$$

čia μ_{ij} – nežinomi parametrai, e_{ijk} – n. a. d., pasiskirstę pagal normalųjį dėsnį $N(0,\ \sigma^2)$.

Taigi imties skirstinys visiškai nusakomi vidurkiais μ_{ij}, $i = 1, ..., I$, $j = 1, ..., J$, ir viena bendra dispersija σ^2. Analizės tikslas – ištirti stebimojo a. d. Y priklausomybę nuo faktorių A ir B. Sujungus Y_{ijk} į vieną bendrą vektorių

$$\boldsymbol{Y} = (Y_{111}, ..., Y_{11K}, ..., Y_{IJ1}, ..., Y_{IJK})^T$$

ir pažymėjus $\boldsymbol{\mu} = (\mu_{11}, ..., \mu_{1J}, ..., \mu_{I1}, ..., \mu_{IJ})^T$ nežinomų parametrų vektorių, modelį (2.2.1) galima užrašyti matriciniu pavidalu:

$$\boldsymbol{Y} = \boldsymbol{A\mu} + \boldsymbol{e}, \quad \mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A\mu}, \quad \boldsymbol{V}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}, \quad \boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Matrica $\boldsymbol{A}$ turi $n = IJK$ eilučių ir IJ stulpelių. Pirmosios K eilučių turi pavidalą $(1,\ 0, ..., 0)$, paskui K eilučių turi pavidalą $(0,\ 1, ..., 0)$, pagaliau paskutinės K eilutės turi pavidalą $(0,\ 0, ..., 1)$.

Šiame modelyje matrica $\boldsymbol{A}^T\boldsymbol{A}$ yra diagonali su vienodais diagonaliniais elementais, lygiais K, taigi neišsigimusi.

Kad vaizdžiau suformuluotume dispersinės analizės hipotezes, įveskime kitus parametrus.

Skirtingai nuo vienfaktorės analizės faktoriaus A lygmens A_i įtaka a. d. Y vidurkiui gali priklausyti ir nuo faktoriaus B reikšmės.

Visų pirma įvesime parametrą, apibūdinantį „tiesioginę“ faktoriaus A lygmens A_i įtaką, kai „eliminuojama“ faktoriaus B įtaka. Populiacijos daliai, kuriai faktoriaus A reikšmė lygi A_i, kintamojo Y vidurkį $\bar{\mu}_{i.}$ apibrėšime formule

$$\bar{\mu}_{i.} = \frac{1}{J}\sum_{j=1}^{J}\mu_{ij},$$

o bendrą visos populiacijos vidurkį formule

$$\bar{\mu}_{..} = \frac{1}{I}\sum_{i=1}^{I}\mu_{i.} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}.$$

Aišku, šie apibrėžimai natūralūs tik su ta prielaida, kad fiksavus A_i populiacijos dalys, kurioms $B = B_1, \ldots, B_J$ yra lygios (ir, atvirkščiai, fiksavus B_j, populiacijos dalys, kurioms $A = A_1, \ldots, A_I$ yra lygios) su visais $i = 1, \ldots, I; j = 1, \ldots, J$. Modeliai, kai ši prielaida nėra patenkinta, nagrinėjami 2.2.4 skyrelyje.

Lygmens A_i įtaką nuokrypiui nuo bendro vidurkio, eliminavus faktoriaus B įtaką, apibūdina skirtumas

$$\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}.$$

Pagal apibrėžimą tai tiesiog a. d. Y vidurkio populiacijos dalies, kurios faktoriaus A reikšmė lygi A_i, ir a. d. Y vidurkio visai populiacijai skirtumas.

Analogiškai įvedame parametrus, kurie apibūdina faktoriaus B lygmenų B_j įtaką nuokrypiui nuo bendro vidurkio, eliminavus faktoriaus A įtaką:

$$\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..}, \quad \bar{\mu}_{.j} = \frac{1}{I}\sum_{i=1}^{I}\mu_{ij}, \quad j = 1, ..., J.$$

Gauname tokį vidurkio μ_{ij} skaidinį į komponentes:

$$\begin{aligned}\mu_{ij} = \bar{\mu}_{..} + (\bar{\mu}_{i.} - \bar{\mu}_{..}) + (\bar{\mu}_{.j} - \bar{\mu}_{..}) + (\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}) = \\ = \mu + \alpha_i + \beta_j + \gamma_{ij}. \qquad (2.2.2)\end{aligned}$$

Naujai įvesti parametrai tenkina tokias papildomas sąlygas:

$$\sum_i \alpha_i = \sum_j \beta_j \equiv \sum_i \gamma_{ij} \equiv \sum_j \gamma_{ij} \equiv 0. \qquad (2.2.3)$$

Jeigu $\gamma_{ij} = 0$ su visais $i = 1, ..., I$ ir $j = 1, ..., J$, tai galioja lygybės

$$\mu_{ij} = \mu + \alpha_i + \beta_j. \tag{2.2.4}$$

Sakome, kad turime *adityvųjį modelį*. Šiame modelyje, esant fiksuotai faktoriaus B reikšmei B_j, faktoriaus A reikšmės A_i įtaką nuokrypiui nuo bendro vidurkio $\mu_{ij} - \mu$ nusako $\alpha_i = \bar{\mu}_{i.} - \mu$, taigi faktoriaus A įtaka nepriklauso nuo faktoriaus B reikšmės. Ir atvirkščiai, esant fiksuotai faktoriaus A reikšmei A_i, faktoriaus B reikšmės B_j įtaką nusako $\beta_j = \bar{\mu}_{.j} - \mu$, taigi faktoriaus B įtaka nepriklauso nuo faktoriaus A reikšmės.

Adityviajame modelyje yra $I+J-1$ nežinomas parametras, nes naujai įvesti parametrai tenkina sąlygas $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$.

Kai $\gamma_{ij} \neq 0$, tai net kai faktoriaus B reikšmė B_j fiksuota, nuokrypį nuo bendro vidurkio $\mu_{ij} - \mu$ nusako ne tik $\alpha_i = \bar{\mu}_{i.} - \mu$, bet ir γ_{ij}, taigi faktoriaus A reikšmės A_i įtaka priklauso nuo faktoriaus B reikšmės. Analogiškai faktoriaus B įtaka priklauso nuo faktoriaus A reikšmės. Taigi komponentės γ apibūdina faktorių A ir B *sąveiką*. Kartais ir pačius parametrus γ vadinsime faktorių sąveika.

Iš sąlygų $\sum_{i=1}^{I} \alpha_i = 0$ ir $\sum_{j=1}^{J} \beta_j = 0$ išplaukia, kad yra $I-1$ nežinomų parametrų α ir $J-1$ nežinomų parametrų β. Iš sąlygų $\sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$ išplaukia, kad

$$\gamma_{iJ} = -\sum_{j=1}^{J-1} \gamma_{ij} \quad (i = 1, \ldots, I-1), \quad \gamma_{Ij} = -\sum_{i=1}^{I-1} \gamma_{ij} \quad (j = 1, \ldots, J-1),$$

$$\gamma_{IJ} = -\sum_{i=1}^{I-1} \gamma_{iJ} = \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} \gamma_{ij},$$

taigi yra $(I-1)(J-1)$ nežinomų parametrų γ, nes visi jie išreiškiami $\gamma_{ij}, i = 1, \ldots, I-1, j = 1, \ldots, J-1$.

2.2.2. Mažiausiųjų kvadratų įvertiniai

Mažiausiųjų kvadratų įvertiniai randami minimizuojant kvadratinę formą

$$SS(\boldsymbol{\mu}) = \sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2.$$

MK įvertiniai $\hat{\mu}_{ij}$ ir kvadratinės formos minimumas yra:

$$\hat{\mu}_{ij} = \bar{Y}_{ij.}, \quad SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2. \tag{2.2.5}$$

Kartu gauname parametrų $\mu = \bar{\mu}_{..}$, α_i, β_j, γ_{ij} įvertinius

$$\hat{\mu} = \bar{Y}_{...}, \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \hat{\gamma}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}, \tag{2.2.6}$$

ir dispersijos σ^2 įvertinį

$$\hat{\sigma}^2 = s^2 = MS_E = SS_E/(IJ(K-1)), \quad SS_E/\sigma^2 \sim \chi^2(IJ(K-1)). \tag{2.2.7}$$

2.2.3. Faktorių įtakos hipotezių tikrinimas

Pagrindinės dvifaktorės dispersinės analizės hipotezės yra:

$$H_A : \alpha_1 = \ldots = \alpha_I, \;\; H_B : \beta_1 = \ldots = \beta_J, \;\; H_{AB} : \gamma_{11} = \ldots = \gamma_{IJ}.$$

Tai sudėtinės hipotezės. Jas tikriname remdamiesi 1.3.2 teorema. Kiekvienas iš parametrų $\alpha_i, \beta_j, \gamma_{ij}$ yra tiesinė parametrų μ_{ij} funkcija. Taigi visos trys hipotezės užrašomos forma $\boldsymbol{H\mu} = \mathbf{0}$, duota 1.3.2 teoremoje.

Hipotezės H_A ekvivalenti tvirtinimui $H_A : \alpha_1 - \alpha_2 = \cdots = \alpha_1 - \alpha_I = 0$ arba $H_A : \bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{1\cdot} - \bar{\mu}_{I\cdot} = 0$. Taigi šią hipotezę galima užrašyti tai$\boldsymbol{H}_A\boldsymbol{\mu} = \mathbf{0}$; čia $\boldsymbol{H}_A$ yra $(I-1) \times IJ$ rango $I-1$ matrica, kurios i-oji eilutė $(i = 1, \ldots, I-1)$ turi pavidalą

$$(1, \ldots, 1, 0, \ldots, 0, -1, \ldots, -1, 0, \ldots, 0);$$

čia 1 yra $1, \ldots, J$ pozicijose, o -1 yra $iJ+1, \ldots, (i+1)J$ pozicijose.

Analogiškai, hipotezės H_B atveju $\boldsymbol{H}_B$ yra $(J-1) \times IJ$ rango $J-1$ matrica.

Hipotezės H_{AB} ekvivalenti forma $H_{AB} : \gamma_{1j} - \gamma_{ij} = 0$, $i = 2, \ldots, I; j = 1, \ldots, J-1$ arba

$$H_{AB} : (J-1)(\mu_{1j} - \mu_{ij}) - \sum_{l \neq j} (\mu_{1l} - \mu_{il}) = 0, \quad i = 2, \ldots, I; j = 1, \ldots, J-1.$$

Taigi šią hipotezę galima užrašyti forma $\boldsymbol{H}_{AB}\boldsymbol{\mu} = \mathbf{0}$; čia $\boldsymbol{H}_{AB}$ yra $(I-1)(J-1) \times IJ$ rango $(I-1)(J-1)$ matrica, kurios $((I-1)(j-1)+i)$-oji eilutė $(i = 1, \ldots, I-1,\ j = 1, \ldots, J-1)$ turi pavidalą

$$(-1, \ldots, -1, J-1, -1, \ldots, -1, 0, \ldots, 0, -1, \ldots, 1, -(J-1), 1, \ldots, 1, 0, \ldots, 0);$$

dviejų serijų iš nenulinių elementu ilgiai yra J, $\pm(J-1)$ yra j-osiose serijų pozicijose, o antra serija prasideda nuo $iJ+1$.

Žymėsime

$$SS_A = SS_{EH_A} - SS_E, \quad SS_B = SS_{EH_B} - SS_E, \quad SS_{AB} = SS_{EH_{AB}} - SS_E,$$

1.3.2 teoremoje apibrėžtas kvadratines formas.

2.2.1 teorema. *Kvadratinės formos SS_A, SS_B ir SS_{AB} turi tokį pavidalą:*

$$SS_A = JK \sum_i (\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2, \quad SS_B = IK \sum_j (\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2,$$

$$SS_{AB} = K \sum_i \sum_j (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2. \tag{2.2.8}$$

Šios kvadratinės formos nepriklauso nuo kvadratų sumos SS_E. Be to,

$$SS_A/\sigma^2 \sim \chi^2(I-1, \lambda_A), \quad SS_B/\sigma^2 \sim \chi^2(J-1, \lambda_B),$$

$$SS_{AB}/\sigma^2 \sim \chi^2((I-1)(J-1), \lambda_{AB}), \tag{2.2.9}$$

necentriškumo parametrai

$$\lambda_A = \frac{JK}{\sigma^2}\sum_i \alpha_i^2, \quad \lambda_B = \frac{IK}{\sigma^2}\sum_j \beta_j^2, \quad \lambda_{AB} = \frac{K}{\sigma^2}\sum_i\sum_j \gamma_{ij}^2.$$

Jeigu hipotezės H_A, H_B, H_{AB} teisingos, tai atitinkami skirstiniai yra centriniai.

Įrodymas. Tapatybės

$$Y_{ijk} - \mu_{ij} = (Y_{ijk} - \hat\mu_{ij}) + (\hat\mu - \mu) + (\hat\alpha_i - \alpha_i) + (\hat\beta_j - \beta_j) + (\hat\gamma_{ij} - \gamma_{ij})$$

abi puses pakėlę kvadratu, susumavę visoje indeksų i, j, k kitimo srityje ir pasinaudoję (2.2.4) sąryšiais ir analogiškais įvertinių (2.2.6) sąryšiais, gauname

$$SS(\boldsymbol{\mu}) = SS_E + IJK(\hat\mu - \mu)^2 + JK\sum_i(\hat\alpha_i - \alpha_i)^2 +$$

$$+IK\sum_j(\hat\beta_j - \beta_j)^2 + K\sum_i\sum_j(\hat\gamma_{ij} - \gamma_{ij})^2.$$

Taigi

$$SS_{EH_A} = \min_{\mu,\alpha_i=0,\beta_j,\gamma_{ij}} SS(\boldsymbol{\mu}) = SS_E + JK\sum_i \hat\alpha_i^2 = SS_E + JK\sum_i(\bar Y_{i\cdot\cdot} - \bar Y_{\cdots})^2.$$

Įrodėme pirmąją (2.2.8) formulę. Analogiškai gauname kitas dvi (2.2.8) formules. Tvirtinimai (2.2.9) tiesiogiai išplaukia iš 1.3.2 teoremos.

▲

Pagal (1.3.28) hipotezės H_A, H_B, H_{AB} atmetamos reikšmingumo lygmens α kriterijais, jeigu atitinkamai tenkinamos nelygybės:

$$F_A > F_\alpha(I-1, IJ(K-1)), \quad F_B > F_\alpha(J-1, IJ(K-1)),$$

$$F_{AB} > F_\alpha((I-1)(J-1, IJ(K-1))); \tag{2.2.10}$$

čia

$$F_A = \frac{IJ(K-1)SS_A}{(I-1)SS_E} = \frac{MS_A}{MS_E}, \quad F_B = \frac{MS_B}{MS_E}, \quad F_{AB} = \frac{MS_{AB}}{MS_E}.$$

Iš (2.2.8) išplaukia, kad

$$\mathbf{E}(MS_A) = \sigma^2 + JK\sigma_A^2, \quad \mathbf{E}(MS_B) = \sigma^2 + IK\sigma_B^2, \quad \mathbf{E}(MS_{AB}) = \sigma^2 + K\sigma_{AB}^2;$$

čia

$$\sigma_A^2 = \frac{1}{I-1}\sum_i \alpha_i^2, \quad \sigma_B^2 = \frac{1}{J-1}\sum_j \beta_j^2, \quad \sigma_{AB}^2 = \frac{1}{(I-1)(J-1)}\sum_i\sum_j \gamma_{ij}^2.$$

Statistikos F_A, F_B ir F_{AB} turi tendenciją įgyti didesnes reikšmes, kai atitinkamai hipotezės H_A, H_B ir H_{AB} yra neteisingos, negu tuo atveju, kai jos teisingos. Taigi hipotezių atmetimo taisyklės (2.2.10) neprieštarauja sveikai logikai.

Pažymėkime

$$SS_T = \sum_i \sum_j \sum_k (Y_{ijk}^2 - \bar{Y}_{...})^2.$$

2.2.1 pastaba. Užrašę kvadratų sumas pavidalu

$$SS_A = JK \sum_i \bar{Y}_{i..}^2 - IJK\bar{Y}_{...}^2, \quad SS_B = IK \sum_i \bar{Y}_{.j.}^2 - IJK\bar{Y}_{...}^2,$$

$$SS_{AB} = K \sum_i \sum_j \bar{Y}_{ij.}^2 - JK \sum_i \bar{Y}_{i..}^2 - IK \sum_j \bar{Y}_{.j.}^2 + IJK\bar{Y}_{...}^2.$$

$$SS_E = \sum_i \sum_j \sum_k Y_{ijk}^2 - K \sum_i \sum_j \bar{Y}_{ij.}^2, \quad SS_T = \sum_i \sum_j \sum_k Y_{ijk}^2 - IJK\bar{Y}_{...}^2.$$

ir sudėję panariui, gauname

$$SS_A + SS_B + SS_{AB} + SS_E = SS_T.$$

Skaičiavimo rezultatus surašome į dispersinės analizės lentelę.

2.2.1 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + JK\sigma_A^2$
B	SS_B	$J-1$	MS_B	$\sigma^2 + IK\sigma_B^2$
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}	$\sigma^2 + K\sigma_{AB}^2$
E	SS_E	I J(K-1)	MS_E	σ^2
T	SS_T	I J K -1	-	-

2.2.2 pastaba. Kvadratų sumos SS_A, SS_B, SS_{AB} ne tik nepriklauso nuo SS_E, bet yra nepriklausomos tarpusavyje ir nepriklauso nuo $\bar{Y}_{...}$. Kad tuo įsitikintume, užtenka panagrinėti atsitiktinių dydžių sistemas $\{\bar{Y}_{...}\}$, $\{\bar{Y}_{i..} - \bar{Y}_{...}\}$, $\{\bar{Y}_{.j.} - \bar{Y}_{...}\}$, $\{\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\}$, $\{Y_{ijk} - \bar{Y}_{ij.}\}$. Bet kurios iš šių sistemų a. d. yra nekoreliuotas (normalaus skirstinio atveju ir nepriklausomas (žr.2 priedą 7.4 skyrelį)) su bet kuriuo kitos sistemos a. d. Todėl bet kurie a. d., sudaryti iš skirtingų sistemų elementų, yra nepriklausomi.

Patikrinsime, pavyzdžiui, kad atsitiktiniai dydžiai $\{\bar{Y}_{i..} - \bar{Y}_{...}\}$ ir $\{\bar{Y}_{.j.} - \bar{Y}_{...}\}$ yra nekoreliuoti. Gauname

$$\mathbf{Cov}\,(\frac{1}{JK}\sum_j \sum_k Y_{ijk} - \frac{1}{IJK}\sum_i \sum_j \sum_k Y_{ijk}, \frac{1}{IK}\sum_i \sum_k Y_{ijk} -$$

$$-\frac{1}{IJK}\sum_i \sum_j \sum_k Y_{ijk}) = \frac{1}{JKIK}K\sigma^2 - \frac{1}{JKIJK}JK\sigma^2 -$$

$$-\frac{1}{IJKIK}IK\sigma^2 + \frac{1}{(IJK)^2}IJK\sigma^2 = 0.$$

2.2.1 pavyzdys. Eksperimento metu kelioms pelių grupėms buvo duodama skirtingų rūšių nuodų ir paskui jos buvo gydomos įvairiais metodais. Lentelėje pateiktos pelių gyvenimo trukmės nuo eksperimento pradžios iki mirties logaritmai.

2.2.2 lentelė. Statistiniai duomenys

	Nuodai		
Gydymo metodas	I	II	III
G_1	0,31 0,45 0,46 0,43	0,36 0,29 0,40 0,23	0,22 0,21 0,18 0,23
G_2	0,82 1,10 0,88 0,72	0,92 0,61 0,49 1,24	0,30 0,37 0,38 0,29
G_3	0,43 0,45 0,63 0,76	0,44 0,35 0,31 0,40	0,23 0,25 0,24 0,24
G_4	0,45 0,71 0,66 0,62	0,56 1,02 0,71 0,38	0,30 0,36 0,31 0,33

Atlikę skaičiavimus, gauname dispersinės analizės lentelę.

2.2.3 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	0,9178	3	0,3059	$\sigma^2 + JK\sigma_A^2$
B	1,0249	2	0,5125	$\sigma^2 + IK\sigma_B^2$
$A \times B$	0,2520	6	0,4200	$\sigma^2 + K\sigma_{AB}^2$
E	0,8004	36	0,0222	σ^2
T	2,9951	47	-	-

Hipotezėms H_A, H_B ir H_{AB} tikrinti gauname statistikų realizacijas : $F_A = 13,76$, $F_B = 23,05$ ir $F_{AB} = 1,89$. Kadangi $\mathbf{P}\{F_{3,36} > 13,76\} = 3,9 \cdot 10^{-6}$, $\mathbf{P}\{F_{2,36} > 23,05\} = 3,6 \cdot 10^{-7}$, $\mathbf{P}\{F_{6,36} > 1,89\} = 0,109$, tai atmesti hipotezę H_{AB} nėra pagrindo. Tuo tarpu hipotezė H_A ir H_B atmetamos aukštų reikšmingumo lygmenų kriterijais.

2.2.2 pavyzdys. Viena iš dažniausių anemijos priežasčių yra geležies trūkumas organime. Geležies kiekis (gramais šimte gramų produkto) mėsoje, pupose ir žaliose daržovėse, virtose aliumininiuose, moliniuose ir geležiniuose induose, pateikti lentelėje (kiekvieno produkto kiekviename inde paimta po keturis mėginius).

2.2.4 lentelė. Statistiniai duomenys

Indo tipas	Mėsa	Pupos	Žalios daržovės
Aliumininis	1,77 2,36 1,96 2,14	2,40 2,17 2,41 2,34	1,03 1,53 1,07 1,30
Molinis	2,27 1,28 2,48 2,68	2,41 2,43 2,57 2,48	1,55 0,79 1,68 1,82
Geležinis	5,27 5,17 4,06 4,22	3,69 3,43 3,84 3,72	2,45 2,99 2,80 2,92

Atlikę skaičiavimus, gauname dispersinės analizės lentelę.

2.2.5 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	24,8940	2	12,4470	$\sigma^2 + JK\sigma_A^2$
B	9,2969	2	4,6484	$\sigma^2 + IK\sigma_B^2$
$A \times B$	2,6404	4	0,6601	$\sigma^2 + K\sigma_{AB}^2$
E	3,6425	27	0,1349	σ^2
T	40,4738	35	-	-

Hipotezėms H_A, H_B ir H_{AB} tikrinti gauname statistikų realizacijas : $F_A = 92,26$, $F_B = 34,46$ ir $F_{AB} = 4,89$. Kadangi $\mathbf{P}\{F_{2,27} > 92,26\} = 8,5 \cdot 10^{-13}$, $\mathbf{P}\{F_{2,27} > 34,46\} = 3,7 \cdot 10^{-8}$, $\mathbf{P}\{F_{4,27} > 4,89\} = 0,0043$, tai hipotezės H_A ir H_B atmetamos, o hipotezė H_{AB} atmetama, jeigu kriterijaus reikšmingumo lygmuo viršija 0,0043.

2.2.4. Kontrastų analizė

Kaip ir vienfaktorės analizės atveju, jeigu hipotezės H_A, H_B, H_{AB} neatmetamos, tai analizę galima baigti: visus stebėjimus galima sujungti į vieną didumo

IJK imtį, gautą stebint a. d. $Y \sim N(\mu,\ \sigma^2)$. Priešingu atveju analizė pratęsiama, siekiant išskirti tokias faktorių lygmenų rinkinių grupes (suprantama, jų turėtų būti kuo mažiau), kad grupių viduje skirtumai tarp vidurkių μ_{ij} būtų neesminiai. Skaidant į grupes, tikslinga naudoti turimą papildomą informaciją apie a. d. Y_{ijk}. Jeigu tokios informacijos nepakanka, galima taikyti kontrastų analizės S metodą arba T metodą.

Pavyzdžiui, norėdami nustatyti faktoriaus A lygmenų įtaką, nagrinėjame kontrastus

$$\psi = \sum_{i=1}^{I} c_i\alpha_i = \sum_{i=1}^{I} c_i(\bar{\mu}_{i.} - \bar{\mu}_{..}) = \sum_{i=1}^{I} c_i\bar{\mu}_{i.}, \quad \sum_{i=1}^{I} c_i = 0.$$

Kontrasto ψ įvertinys, jo dispersija ir dispersijos įvertinys yra

$$\hat{\psi} = \sum_{i=1}^{I} c_i Y_{i..}, \quad \boldsymbol{V}(\hat{\psi}) = \frac{\sigma^2}{JK}\sum_{i=1}^{I} c_i^2, \quad \hat{\boldsymbol{V}}(\hat{\psi}) = \frac{s^2}{JK}\sum_{i=1}^{I} c_i^2, \quad s^2 = MS_E.$$

Naudojant S metodą (žr. 2.1.2 teoremą), hipotezė H_A atmetama tada ir tik tada, kai atsiranda kontrastas ψ, kad intervalas (2.1.14) neuždengia 0. Sudarant intervalus (2.1.14) reikia imti $\Delta^2 = (I-1)F_\alpha(I-1, IJ(K-1))$.

Taikant kontrastų palyginimo T metodą (žr. 2.1.4 teoremą) poroms $\mu_{i.}$ ir $\mu_{j.}$, imama $\nu = IJ(K-1)$, o $\hat{\mu}_i$ pakeičiami į $\hat{\mu}_{i.} = \bar{Y}_{i..}$.

Abu metodai leidžia išskirti kontrastus, atsakingus už hipotezės atmetimą.

2.2.5. Vieno stebėjimo langelyje atvejis

2.2.5.1. Adityvus modelis

Jeigu dvifaktorėje analizėje kartotinumo skaičius $K = 1$, tai a. d. Y_{ij} skaičius IJ lygus nežinomų parametrų μ_{ij} skaičiui. Minimizuojant kvadatų sumą $SS(\boldsymbol{\mu})$, gaunami vidurkių įvertiniai $\hat{\mu}_{ij} = Y_{ij}$ ir $SS_E = 0$. Dispersijos įvertinys (2.2.6) neapibrėžtas. Norint sumažinti nežinomų parametrų skaičių, reikia nagrinėti siauresnį modelį. Tarsime, kad nėra faktorių A ir B sąveikos, t. y. visi $\gamma_{ij} = 0$. Tokiu atveju vidurkiai μ_{ij} aprašomi adityviuoju modeliu:

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \quad \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0,$$

nežinomų parametrų skaičius vidurkio išraiškoje yra $I + J - 1$.

Kriterijai hipotezėms $H_A : \alpha_1 = \ldots = \alpha_I$, ir $H_B : \beta_1 = \ldots = \beta_J \equiv 0$ tikrinti sudaromi analogiškai. Apibrėžus kvadratų sumas SS_A, SS_B ir SS_{AB} pagal tas pačias (2.2.8) formules (imant $K = 1$ visose formulėse) 2.2.1 teoremos rezultatai išlieka teisingi. Kadangi negalima panaudoti $SS_E = 0$ kriterijams sudaryti, tai kriterijai hipotezėms H_A ir H_B tikrinti grindžiami statistikomis

$$F_A = \frac{MS_A}{MS_{AB}}, \quad F_B = \frac{MS_B}{MS_{AB}}.$$

Hipotezės atmetamos reikšmingumo lygmens α kriterijais, kai atitinkamai tenkinamos nelygybės

$$F_A > F_\alpha(I-1,(I-1)(J-1)), \quad F_B > F_\alpha(J-1,(I-1)(J-1)). \qquad (2.2.11)$$

Skaičiavimo rezultatus surašome į lentelę analogišką 2.2.1 lentelei. Ji skirsis nuo 2.2.1 tuo, kad bus išbraukta eilutė, atitinkanti faktorių E; vietoje K įrašytas 1; prie raidžių Y nebus taško, atitinkančio indeksą k; vietoje σ^2_{AB} bus įrašyta 0.

2.2.5.2. Modelio adityvumo hipotezės tikrinimas

Kriterijai (2.2.11) nėra korektiški, jeigu yra faktorių sąveika, t. y. kai kurie γ_{ij} nelygūs nuliui. Kyla natūralus uždavinys patikrinti hipotezę $H_{AB} : \gamma_{ij} \equiv 0$ remiantis stebėjimais, kai $K = 1$.

Kaip adityviojo modelio alternatyvą nagrinėkime modelį (2.2.2), kuriame faktorių A ir B sąveika tenkina lygybę $\gamma_{ij} = \gamma\alpha_i\beta_j$, t. y.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j + e_{ij}, \quad \sum_{i=1}^{I}\alpha_i = 0, \quad \sum_{j=1}^{I}\beta_j = 0; \qquad (2.2.12)$$

čia n. a. d. $e_{ij} \sim N(0,\sigma^2)$.

Priėmus tokią prielaidą faktorių sąveikos nebuvimo hipotezė ekvivalenti hipotezei $H_\gamma : \gamma = 0$.

2.2.2 teorema. (Tjukio kriterijus). *Pažymėkime*

$$SS_\gamma = \frac{(\sum_i\sum_j \hat\alpha_i\hat\beta_j Y_{ij})^2}{\sum_i\hat\alpha_i^2\sum_j\hat\beta_j^2}, \quad \hat\alpha_i = \bar Y_{i.} - \bar Y_{..}, \quad \hat\beta_j = \bar Y_{.j} - \bar Y_{..},$$

$$SS_{AB} = \sum_i\sum_j (Y_{ij} - \bar Y_{i.} - \bar Y_{.j} + \bar Y_{..})^2, \quad SS_L = SS_{AB} - SS_\gamma.$$

Kai hipotezė H_γ teisinga, a. d. SS_γ/σ^2 ir SS_L/σ^2 yra nepriklausomi ir pasiskirstę pagal chi kvadrato dėsnius su 1 ir $IJ-I-J$ laisvės laipsnių. Tuo remdamiesi sudarome kriterijų hipotezei H_γ tikrinti: hipotezė atmetama, kai

$$F_\gamma > F_\alpha(1,\ IJ-I-J), \quad F_\gamma = SS_\gamma(IJ-I-J)/SS_L. \qquad (2.2.13)$$

Įrodymas. Iš pradžių tarkime, kad α_i ir β_j žinomos konstantos, tenkinančios sąlygas $\sum_i\alpha_i = \sum_j\beta_j = 0$. Tada turime tiesinį modelį, kuriame vidurkiai μ_{ij} tiesiškai priklauso tiktai nuo dviejų nežinomų parametrų μ ir γ. Minimizuodami pagal μ ir γ kvadratinę formą $\sum_i\sum_j(Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma\alpha_i\beta_j)^2$, randame parametro γ mažiausiųjų kvadratų įvertinį

$$\hat\gamma = \frac{\sum_i\sum_j\alpha_i\beta_j Y_{ij}}{\sum_i\alpha_i^2\sum_j\beta_j^2} \sim N\left(\gamma,\ \frac{\sigma^2}{\sum_i\alpha_i^2\sum_j\beta_j^2}\right).$$

Jeigu hipotezė H_γ teisinga, tai santykis

$$X^2 = X^2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\hat{\gamma}^2 \sum_i \alpha_i^2 \sum_j \beta_j^2}{\sigma^2} = \frac{(\sum_i \sum_j \alpha_i \beta_j Y_{ij})^2}{\sigma^2 \sum_i \alpha_i^2 \sum_j \beta_j^2} \sim \chi^2(1).$$

Taigi santykio X^2 skirstinys nepriklauso nuo parametrų α_i ir β_j reikšmių. Skirstinys išlieka toks pat, jei modelyje (2.2.12) parametrus α_i ir β_j pakeistume kitais parametrais α_i^* ir β_j^*, tenkinančiais sąlygas $\sum_i \alpha_i^* = 0$, $\sum_j \beta_j^* = 0$, nes

$$\sum_i \sum_j \alpha_i \beta_j Y_{ij} = \sum_i \sum_j \alpha_i \beta_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}),$$

o $Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$ skirstinys nepriklauso nuo parametrų α_i^* ir β_j^* reikšmių.

Maža to, santykio X^2 skirstinys išlieka toks pat, jei to santykio išraiškoje α_i ir β_j pakeisime įvertiniais $\hat{\alpha}_i$ ir $\hat{\beta}_j$, surastais modelyje, kai nežinomi α_i ir β_j:

$$\frac{SS_\gamma}{\sigma^2} = \frac{(\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j Y_{ij})^2}{\sigma^2 \sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2} \sim \chi^2(1), \tag{2.2.14}$$

$$\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}, \quad \hat{\beta}_j = \bar{Y}_{.j} - \bar{Y}_{..}, \quad \sum_i \hat{\alpha}_i = \sum_j \hat{\beta}_j = 0.$$

Iš tikrųjų lygybės

$$\mathbf{P}_{\boldsymbol{\alpha},\boldsymbol{\beta}}\{X^2(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \le x\} = \int \mathbf{P}_{\boldsymbol{\alpha},\boldsymbol{\beta}}\{X^2(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \le x | \hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*, \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*\} dF_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

$$= \mathbf{P}_{\boldsymbol{\alpha},\boldsymbol{\beta}}\{X^2(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \le x\} = \mathbf{P}_{\boldsymbol{\alpha},\boldsymbol{\beta}}\{X^2(\boldsymbol{\alpha}, \boldsymbol{\beta}) \le x\} = \mathbf{P}\{\chi_1^2 \le x\},$$

bus įrodytos, jei parodysime, kad a. d. $X^2(\alpha^*, \beta^*)$ nepriklauso nuo $\hat{\alpha}_i$ ir $\hat{\beta}_j$. Bet tai išplaukia iš to, kad $X^2(\alpha^*, \beta^*)$ yra $Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$ funkcija, $\hat{\alpha}_i$ yra $\bar{Y}_{i.} - \bar{Y}_{..}$ funkcija, $\hat{\beta}_j$ yra $\bar{Y}_{.j} - \bar{Y}_{..}$ funkcija, o remiantis 2.2.2 pastaba, atsitiktinių dydžių sistemos $\{\bar{Y}_{i.} - \bar{Y}_{..}\}$, $\{\bar{Y}_{.j} - \bar{Y}_{..}\}$, $\{Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\}$ yra nepriklausomos.

Pagal 2.2.1 teoremą

$$\frac{SS_{AB}}{\sigma^2} = \frac{1}{\sigma^2} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 \sim \chi^2((I-1)(J-1)).$$

Nagrinėkime išdėstymą

$$\frac{SS_{AB}}{\sigma^2} = \frac{SS_\gamma}{\sigma^2} + \frac{SS_{AB} - SS_\gamma}{\sigma^2} = \frac{SS_\gamma}{\sigma^2} + \frac{SS_L}{\sigma^2}. \tag{2.2.15}$$

Lieka pasinaudoti tokia kvadratinių formų nuo nepriklausomų normaliųjų a. d. savybe. Tegu kvadratinė forma $Q \sim \chi^2(\nu)$ ir $Q = Q_1 + Q_2$. Jeigu $Q_1 \sim \chi^2(\nu_1)$, o Q_2 neneigiama, tai $Q_2 \sim \chi^2(\nu - \nu_1)$ ir nepriklauso nuo Q_1.

Kadangi $SS_{AB}/\sigma^2 \sim \chi^2((I-1)(J-1))$ ir $SS_\gamma/\sigma^2 \sim \chi^2(1)$, lieka parodyti, kad $SS_{AB} - SS_\gamma \geq 0$. Tai išplaukia iš Koši nelygybės:

$$(\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j Y_{ij})^2 = (\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}))^2 \leq$$

$$\sum_i \sum_j \hat{\alpha}_i^2 \hat{\beta}_j^2 \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2.$$

Taigi $SS_L/\sigma^2 \sim \chi^2(IJ - I - J)$ ir nepriklauso nuo SS_γ/σ^2.

▲

2.2.5.3. Kontrastų analizė

Kontrastų analizė atliekama analogiškai, kaip ir kai $K > 1$. Pavyzdžiui, tirdami faktoriaus A lygmenų įtaką, nagrinėjame kontrastus

$$\psi = \sum_{i=1}^{I} c_i \alpha_i = \sum_{i=1}^{I} c_i (\mu_i - \bar{\mu}_.) = \sum_{i=1}^{I} c_i \bar{\mu}_{i.}, \quad \sum_{i=1}^{I} c_i = 0.$$

Kontrasto ψ įvertinys, jo dispersija ir dispersijos įvertinys yra

$$\hat{\psi} = \sum_{i=1}^{I} c_i \bar{Y}_{i.}, \quad \boldsymbol{V}(\hat{\psi}) = \frac{\sigma^2}{J} \sum_{i=1}^{I} c_i^2, \quad \hat{\boldsymbol{V}}(\hat{\psi}) = \frac{s^2}{J} \sum_{i=1}^{I} c_i^2, \quad s^2 = MS_{AB}.$$

Jei naudojamas S metodas, tai sudarant intervalus (2.1.17) reikia imti $\Delta^2 = (I-1)F_\alpha(I-1,(I-1)(J-1))$. Naudojant T metodą, sudarant intervalus (2.1.20), reikia imti $\nu = (I-1)(J-1)$.

2.2.3 pavyzdys. Atliekant dvifaktorę dispersinę analizę su dviem pastoviais faktoriais ir vienu stebėjimu ląstelėje, turimi tokie stebiniai

2.2.6 lentelė. Statistiniai duomenys

A_i	B_1	B_2	B_3	B_4	B_5	B_6	B_7
A_1	164	172	177	178	163	163	150
A_2	177	197	184	196	177	193	179
A_3	168	167	187	177	144	176	146
A_4	156	161	169	181	165	172	141
A_5	172	180	179	184	166	176	169
A_6	196	190	197	191	178	178	183

Visų pirma taikydami Tjukio kriterijų patikrinsime modelio adityvumo hipotezę. Gauname,kad statistika F_γ įgijo reikšmę 5,44. Jeigu modelio adityvumo prielaida teisinga, tai ši statistika turi Fišerio skirstinį su 1 ir 29 laisvės laipsniais. Adityvumo hipotezė reikšmingumo lygmens $\alpha = 0,02$ kriterijumi neatmetama.

Atlikę skaičiavimus gauname dispersinės analizės lentelę.

2.2.7 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS
A	3861,8	5	772,4
B	2596,0	6	432,7
$A \times B$	1624,3	30	54,1
P	8082,1	41	-

Statistika $F_A = MS_A/MS_{AB}$ įgijo reikšmę 14,3, o statistika $F_B = MS_B/MS_{AB}$ – reikšmę 8,0. Hipotezės H_A ir H_B yra atmetamos kriterijumi su gana aukštu reikšmingumo lygmeniu. Faktorių A ir B įtaką vidurkių kitimui galima apibūdinti mažesniu parametrų skaičiumi. Pavyzdžiui, suskirsčius faktoriaus A lygmenis į dvi grupes (A_2, A_6) ir (A_1, A_3, A_4, A_5), o faktoriaus B lygmenis į tris grupes (B_1, B_2, B_6), (B_3, B_4) ir (B_5, B_7), hipotezė, kad grupių viduje vidurkiai vienodi, neatmetama.

2.3. Dvifaktorė analizė, kai stebėjimų skaičius langeliuose skirtingas

2.3.1. Statistinis modelis

Tarkime, kad a. d. Y skirstinys gali priklausyti nuo faktoriaus A, kurio lygmenys yra $A_1, ..., A_I$, ir nuo faktoriaus B, kurio lygmenys yra $B_1, ..., B_J$. Skirtingai nei 2.2.3 skyrelyje tarsime, kad stebėjimų skaičius K_{ij}, kai faktorių lygmenys yra A_i ir B_j, gali būti skirtingas, t. y. eksperimento planas gali būti nesubalansuotas. Imties elementus žymėsime Y_{ijk}; $k = 1, ..., K_{ij}$, $i = 1, ..., I$, $j = 1, ..., J$; i – faktoriaus A lygmens numeris, j – faktoriaus B lygmens numeris, k – kartotinumo numeris; bendras stebėjimų skaičius $n = \sum_i \sum_j K_{ij}$.

Imties elementai surašyti 2.3.1 lentelėje, o lentelėje 2.3.2 nurodyti jų kartotinumas langeliuose. Jei $K_{ij} = 0$, tai 2.3.1 lentelės i-osios eilutės ir j-ojo stulpelio sankirtoje esantį langelį reikia pakeisti tuščiu.

2.3.1 lentelė. Imties elementai

i j	1	...	j	...	J	$\sum$
1	$Y_{111}, ..., Y_{11K_{11}}$	...	$Y_{1j1}, ..., Y_{1jK_{1j}}$	...	$Y_{1J1}, ..., Y_{1JK_{1J}}$	$Y_{1..}$
...	...	...	...	...	...	...
i	$Y_{i11}, ..., Y_{i1K_{i1}}$	...	$Y_{ij1}, ..., Y_{ijK_{ij}}$	...	$Y_{iJ1}, ..., Y_{iJK_{iJ}}$	$Y_{i..}$
...	...	...	...	...	...	...
I	$Y_{I11}, ..., Y_{I1K_{I1}}$	...	$Y_{Ij1}, ..., Y_{IjK_{Ij}}$	...	$Y_{IJ1}, ..., Y_{IJK_{IJ}}$	$Y_{I..}$
$\sum$	$Y_{.1.}$	...	$Y_{.j.}$	...	$Y_{.J.}$	$Y_{...}$

2.3.2 lentelė. Stebėjimų kartotinumas langeliuose

i j	1	...	j	...	J	$\sum$
1	K_{11}	...	K_{1j}	...	K_{1J}	$K_{1.}$
...	...	...	...	...	...	...
i	K_{i1}	...	K_{ij}	...	K_{iJ}	$K_{i.}$
...	...	...	...	...	...	...
I	K_{I1}	...	K_{Ij}	...	K_{IJ}	$K_{I.}$
$\sum$	$K_{.1}$	...	$K_{.j}$	...	$K_{.J}$	n

Dvifaktorės dispersinės analizės modelis:

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad k = 1, ..., K_{ij}, \quad i = 1, ..., I, \quad j = 1, ..., J, \tag{2.3.1}$$

čia μ_{ij}, $i = 1, ..., I$, $j = 1, ..., J$ – nežinomi parametrai, e_{ijk} – n. a. d. vienodai pasiskirstę pagal normalųjį dėsnį $N(0,\ \sigma^2)$.

Taigi stebėjimų skirstiniai visiškai nusakomi vidurkiais $\mu_{ij}, i = 1, ..., I, j = 1, ..., J$ ir viena bendra dispersija σ^2. Jeigu kuriame nors langelyje (i, j) stebėjimų skaičius $K_{ij} = 0$, tai, nedarant jokių prielaidų apie vidurkių sąryšį, parametras μ_{ij} negali būti įvertintas.

Pažymėkime $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$ nežinomų parametrų vektorių, kuris gaunamas iš vektoriaus $\boldsymbol{\mu} = (\mu_{11}, ..., \mu_{1J}, ..., \mu_{I1}, ..., \mu_{IJ})^T$ praleidus tuos vidurkius μ_{ij}, kuriems $K_{ij} = 0$, ir tegu $\tilde{K}_i > 0$, $i = 1, ..., m$, – stebėjimų skaičius, atitinkantis koordinatę θ_i; čia m – netuščių langelių skaičius. Tada $n = \sum_{i=1}^{m} \tilde{K}_i = \sum_{i=1}^{I} \sum_{j=1}^{J} K_{ij}$.

Sujungę stebėjimus į vieną bendrą vektorių

$$\boldsymbol{Y} = (Y^*_{11}, ..., Y^*_{1\tilde{K}_1}, ..., Y^*_{m1}, ..., Y^*_{m\tilde{K}_m})^T;$$

čia $Y^*_{l1}, ..., Y^*_{l\tilde{K}_l}$ atitinka faktorių A_i ir B_j reikšmes, kurioms $\theta_l = \mu_{ij}$, modelį (2.3.1) galima užrašyti matriciniu pavidalu

$$\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{e}, \quad \mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A\theta}, \quad \boldsymbol{V}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}, \quad \boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

Matrica $\boldsymbol{A}$ turi n eilučių ir m stulpelių. Pirmosios $\tilde{K}_1$ eilučių turi pavidalą $(1,\ 0, ..., 0)$; tolesnės $\tilde{K}_2$ eilučių turi pavidalą $(0,\ 1, ..., 0)$, ir pagaliau paskutiniosios $\tilde{K}_m$ eilučių turi pavidalą $(0,\ 0, ..., 1)$. Matrica $\boldsymbol{A}^T\boldsymbol{A}$ yra diagonalioji su diagonaliniais elementais $\tilde{K}_1, ..., \tilde{K}_m$; jos rangas $Rang(\boldsymbol{A}^T\boldsymbol{A}) = m$.

2.3.2. Mažiausiųjų kvadratų įvertiniai

Parametrų μ_{ij} MK įvertiniai $\hat{\mu}_{ij}$ gaunami minimizuojant kvadratinę formą

$$SS(\boldsymbol{\theta}) = (\boldsymbol{Y} - \boldsymbol{A\theta})^T(\boldsymbol{Y} - \boldsymbol{A\theta}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K_{ij}} (Y_{ijk} - \mu_{ij})^2.$$

Įvertiniai $\hat{\mu}_{ij}$ randami tiems langeliams, kuriuose stebėjimų skaičius $K_{ij} > 0$. Jie yra nepriklausomi ir normalieji:

$$\hat{\mu}_{ij} = \bar{Y}_{ij.} = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} Y_{ijk} \sim N(\mu_{ij}, \frac{\sigma^2}{K_{ij}}). \tag{2.3.2}$$

Kadangi $\sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 \sim \sigma^2 \chi^2_{K_{ij}-1}$, kai $K_{ij} > 1$, ir $\sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 = 0$, kai $K_{ij} = 1$, tai $\sum_{i,j:K_{ij}\geq 1} (K_{ij} - 1) = n - m$, todėl liekamoji kvadratinė forma

$$SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 \sim \sigma^2 \chi^2_{n-m}. \tag{2.3.3}$$

Dispersijos įvertinys

$$\hat{\sigma}^2 = s^2 = \frac{SS_E}{n - m} = MS_E, \quad \frac{s^2(n - m)}{\sigma^2} \sim \chi^2(n - m). \tag{2.3.4}$$

2.3.3. Faktorių įtakos apibūdinimas

Norint suformuluoti pagrindines dispersinės analizės hipotezes, kaip ir pirmesniuose skyreliuose, reikia įvesti naujus parametrus, kurie apibūdintų faktorių A ir B lygmenų įtaką atsitiktinių dydžių Y_{ijk} skirstiniams.

Apibrėžiant naujus parametrus reikia atsižvelgti į tai, kad stebėjimų skaičius 2.3.2 lentelės langeliuose gali būti skirtingas.

Kai planas subalansuotas populiacijos dalies, kurios faktoriaus A reikšmė lygi A_i, kintamojo Y vidurkis $\bar{\mu}_{i.}$ buvo apibrėžtas formule $\bar{\mu}_{i.} = \frac{1}{J}\sum_{j=1}^{J}\mu_{ij}$, o bendras visos populiacijos vidurkis formule $\bar{\mu}_{..} = \frac{1}{I}\sum_{i=1}^{I}\mu_{i.} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}$. Reikšmės A_i įtaką nuokrypiui nuo bendrojo vidurkio, eliminavus faktoriaus B įtaką, apibūdino skirtumas $\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}$.

Toks vidurkio $\bar{\mu}_{i.}$ ir nuokrypio α_i apibrėžimas natūralus, jei tarp populiacijos elementų, kuriems $A = A_i$, faktoriaus B reikšmes B_1,..., B_J pasiskirstę vienodomis dalimis. Tada vidurkiams $\mu_{i1}, \dots \mu_{iJ}$ suteikiamas svoris $1/J$. Analogiškai buvo apibrėžiama faktoriaus B įtaka.

Kai planas nesubalansuotas, vidurkis $\bar{\mu}_{i.}$ dažnai apibrėžiamas kaip svertinis vidurkis

$$\bar{\mu}_{i.} = \sum_{j=1}^{J}\omega_j\mu_{ij}; \tag{2.3.5}$$

čia $\omega_j = K_{.j}/n$. Toks apibrėžimas natūralus, jei tarp populiacijos elementų, kuriems $A = A_i$, faktoriaus B reikšmės B_1,..., B_J pasiskirčiusios dalimis $\omega_1 = K_{.1}/n$,..., $\omega_J = K_{.J}/n$, $\sum_{j=1}^{J}\omega_j = 1$. Tada vidurkiams $\mu_{i1}, \dots \mu_{iJ}$ suteikiami svoriai $\omega_1, \dots, \omega_J$.

Analogiškai populiacijos dalies, kuriai faktoriaus B reikšmė lygi B_j, kintamojo Y vidurkis

$$\bar{\mu}_{.j} = \sum_{i=1}^{J} v_i\mu_{ij}; \tag{2.3.6}$$

čia $v_i = K_{i.}/n$, $\sum_{i=1}^{I} v_i = 1$.

Bendras visų stebėjimų svertinis vidurkis apibrėžiamas lygybe

$$\mu = \sum_{i=1}^{I}\sum_{j=1}^{J} v_i\omega_j\mu_{ij}. \tag{2.3.7}$$

Analogiškai subalansuoto plano atvejui faktorių A ir B lygmenų A_i ir B_j „tiesioginei“ įtakai apibūdinti įvedame parametrus

$$\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..} = \sum_{j=1}^{J}\omega_j\mu_{ij} - \mu, \quad i = 1, ..., I,$$

$$\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..} = \sum_{i=1}^{I} v_i\mu_{ij} - \mu, \quad j = 1, ..., J. \tag{2.3.8}$$

Pagal savo apibrėžimą tai tiesiog skirtumai tarp a. d. Y vidurkio populiacijos daliai, kuriai faktoriaus A reikšmė lygi A_i (atitinkamai populiacijos daliai, kuriai faktoriaus B reikšmė lygi B_j) ir a. d. Y vidurkio visai populiacijai.

Apskritai, jei turima apriorinė informacija, kad, fiksavus A_i, populiacijos dalys, kurioms faktorius B įgyja reikšmes B_1,..., B_J (fiksavus B_j, populiacijos dalys, kurioms faktorius A įgyja reikšmes A_1,..., A_I) santykiauja ne kaip $K_{.1}/n, \ldots, K_{.J}/n$ (atitinkamai $K_{1.}/n, \ldots, K_{I.}/n$), o, sakykime, kaip $\omega_1, \ldots, \omega_J$ (atitinkamai $v_1, \ldots, v_I$) :

$$v_i \geq 0, \quad \omega_j \geq 0, \quad \sum_{i=1}^{I} v_i = 1, \quad \sum_{j=1}^{J} \omega_j = 1, \tag{2.3.9}$$

tai vidurkiai $\bar{\mu}_{i.}$ ir $\bar{\mu}_{.j}$, taip pat parametrai α_i ir β_j, apibūdinantys faktorių įtaką, apibrėžiami tomis pačiomis formulėmis (2.3.5) – (2.3.8), tiktai naudojant apriorinius svorius (2.3.9).

Subalansuoto plano parametrai α_i ir β_j tenkino sąlygas $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, kai planas nesubalansuotas, jie tenkina sąlygas

$$\sum_{i=1}^{I} v_i \alpha_i = \sum_{i=1}^{I} \sum_{j=1}^{J} v_i \omega_j \mu_{ij} - \mu \sum_{i=1}^{I} v_i = 0, \quad \sum_{j=1}^{J} \omega_j \beta_j = 0. \tag{2.3.10}$$

Gauname tokį vidurkio μ_{ij} skaidinį į komponentes:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\mu_{ij} - \alpha_i - \beta_j - \mu) =$$

$$\mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{2.3.11}$$

Jeigu $\gamma_{ij} = 0$ su visais i ir j, tai galioja lygybės

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \quad i = 1, ..., I, \quad j = 1, ..., J, \tag{2.3.12}$$

Turime *adityvųjį modelį*. Šiame modelyje bet kurio faktoriaus lygmens poveikis a. d. Y vidurkiui, kai kitas faktorius fiksuotas, nepriklauso nuo to fiksuoto faktoriaus reikšmės. Nežinomi parametrai, apibūdinantys vidurkio kitimą, yra α_i, $i = 1, ..., I$; β_j, $j = 1, ..., J$. Iš lygybių (2.3.8) ir (2.3.10) išreiškę, pavyzdžiui, α_I ir β_J, gausime, kad modelis (2.3.11) nusakomas $I+J-1$ parametru: μ; α_i, $i = 1, ..., I-1$; β_j, $j = 1, ..., J-1$.

Kai $\gamma_{ij} \neq 0$, tai bet kurio faktoriaus lygmens poveikis a. d. Y vidurkiui, kai fiksuotas kitas faktorius, priklauso dar ir nuo to fiksuoto faktoriaus reikšmės. Taigi komponentės γ apibūdina faktorių A ir B *sąveiką*. Kartais ir pačius parametrus γ vadinsime faktorių sąveika.

Tiesiškai nepriklausomų komponenčių γ_{ij} skaičius yra $(I-1)(J-1)$, nes jos tenkina sąlygas

$$\sum_{i=1}^{I} v_i \gamma_{ij} = 0, \quad \forall\, j = 1, ..., J; \quad \sum_{j=1}^{J} \omega_j \gamma_{ij} = 0, \quad \forall\, i = 1, ..., I, \tag{2.3.13}$$

todėl analogiškai 3.2.1 skyreliui visus γ_{ij} galima išreikšti γ_{ij}, kai $i = 1, ..., I-1$ ir $j = 1, ..., J-1$.

2.3.1 pastaba. Svorių sistema $v_i = K_{i.}/n$ ir $\omega_j = K_{.j}/n$ vadinama *dažnumine*, o svorių sistema $v_i = 1/I$, $\omega_j = 1/J$ – *vienodų svorių* sistema.

2.3.2 pastaba. Subalansuotų eksperimentų planų, kai stebėjimų skaičiai langeliuose $K_{ij} = K \geq 1$ yra vienodi, dažnuminių ir lygių svorių sistemos sutampa. Galime sakyti, kad 2.2 skyrelyje analizę atlikome prie svorių sistemos $\{v_i\}, \{\omega_j\}$, kai $v_i = 1/I,\ \omega_j = 1/J$.

Naujų parametrų terminais pagrindines dispersinės analizės hipotezes galima suformuluoti analogiškai kaip pirmesniame skyrelyje.

Sąveikos nebuvimo hipotezė

$$H_{AB} : \gamma_{ij} = 0, \quad \forall i = 1, ..., I, \quad \forall j = 1, ..., J, \tag{2.3.14}$$

kuri yra ekvivalenti tvirtinimui, kad modelis yra adityvusis (2.3.11) pavidalo, taip pat faktorių A ir B įtakos nebuvimo hipotezės

$$H_A : \alpha_i = 0, \quad \forall i = 1, ..., I, \quad H_B : \beta_j = 0, \quad \forall j = 1, ..., J. \tag{2.3.15}$$

Hipotezės tikrinamos remiantis 1.3.2 teorema. Priminsime, kad kriterijus kuriamas tokiu būdu: randamas besąlyginis kvadratinės formos $SS(\boldsymbol{\beta})$ minimumas SS_E ir sąlyginis $SS(\boldsymbol{\beta})$ minimumas SS_{EH} su sąlyga, kad tikrinamoji hipotezė H teisinga. Sprendimas priimamas palyginant $SS_{EH} - SS_E$ ir SS_E.

Esminis skirtumas nuo subalansuotų planų yra tas, kad parametrų įvertinių ir liekamosios kvadratinės formos suradimas, kai teisinga viena iš suformuluotų hipotezių, yra kur kas sudėtingesnis. Įvertiniai bendru atveju priklauso nuo svorių sistemos ir negalime jų užrašyti išreikštiniu pavidalu, o tenka apsiriboti nurodant lygčių sistemą, kurios sprendiniais yra MK įvertiniai.

Intuityviai aišku, kad nepriklausomai nuo to, kokiu santykiu pasiskirstę populiacijos elementai pagal faktorių A ir B reikšmes, tie faktoriai arba sąveikauja, arba ne. Taigi sąveikos buvimas arba nebuvimas, neturėtų priklausyti nuo parenkamos svorių sistemos. Šį samprotavimą pagrindžia toliau įrodoma teorema. Ji pravers tikrinant faktorių įtakos nebuvimo hipotezę.

2.3.1 teorema. *(Šefės). 1. Jeigu kokios nors svorių $\{v_i\}, \{w_j\}$ sistemos visos sąveikos γ_{ij} lygios 0, tai jos lygios 0 ir esant bet kuriai kitai svorių sistemai.*
2. Jeigu visi $\gamma_{ij} = 0$, tai efektų α_i ir β_j palyginimo kontrastai nepriklauso nuo svorių sistemos.

Įrodymas. 1. Tarkime, kad kai yra svorių $\{v_i\}, \{w_j\}$ sistema, visos sąveikos $\gamma_{ij} \equiv 0$. Tada turime adityvųjį modelį

$$\mu_{ij} = \mu + \alpha_i + \beta_j. \tag{2.3.16}$$

Tegu $\{v_i^0\}, \{w_j^0\}$ kita svorių sistema $v_i^0 \geq 0, \sum_i v_i^0 = 1,\ w_j^0 \geq 0, \sum_j w_j^0 = 1$, kai turime vidurkio μ_{ij} dėstinį (2.3.12)

$$\mu_{ij} = \mu^0 + \alpha_i^0 + \beta_j^0 + \gamma_{ij}^0. \tag{2.3.17}$$

Įstatę į (2.3.6), (2.3.7) ir (2.3.9) (kuriose visi parametrai perrašyti su viršutiniu indeksu ^{0}, vidurkio (2.3.16) išraiškas, gauname

$$\mu^0 = \sum_i \sum_j v_i^0 \omega_j^0 \mu_{ij} = \mu + \sum_i v_i^0 \alpha_i + \sum_j \omega_j^0 \beta_j,$$

$$\alpha_i^0 = \mu + \alpha_i + \sum_j \omega_j^0 \beta_j - \mu^0 = \alpha_i - \sum_i v_i^0 \alpha_i, \tag{2.3.18}$$

$$\beta_j^0 = \mu + \sum_i v_i^0 \alpha_i + \beta_j - \mu^0 = \beta_j - \sum_j \omega_j^0 \beta_j.$$

Įrašę šias išraiškas į (2.3.17), gauname

$$\gamma_{ij}^0 = \mu_{ij} - \mu^0 - \alpha_i^0 - \beta_j^0 = \mu_{ij} - \mu - \alpha_i - \beta_j = 0;$$

čia rėmėmės (2.3.16) lygybe. Taigi visos sąveikos γ_{ij}^0, kai svorių sistema $\{v_i^0\}, \{\omega_j^0\}$, lygios 0.

2. Remiantis (2.3.18) galima tvirtinti, kad parametrai α_i ir β_j gali skirtis nuo α_i^0 ir β_j^0 tik konstanta: $\alpha_i^0 = \alpha_i + a$, $\beta_j^0 = \beta_j + b$. Todėl kontrastai

$$\psi = \sum_i c_i \alpha_i^0 = \sum_i c_i \alpha_i, \quad \psi' = \sum_j c_j' \beta_j^0 = \sum_j c_j' \beta_j$$

sutampa, nes $\sum_i c_i = 0$, $\sum_j c_j' = 0$. ▲

2.3.4. Faktorių sąveikos nebuvimo hipotezės tikrinimas

Faktorių sąveikos nebuvimo hipotezė $H_{AB} : \gamma_{ij} = 0$, $\forall i = 1, ..., I$, $j = 1, ..., J$, yra ekvivalenti tvirtinimui, kad modelis yra adityvusis:

$$H_{AB} : \mu_{ij} = \mu + \alpha_i + \beta_j, \quad i = 1, ..., I, \quad j = 1, ..., J. \tag{2.3.19}$$

Naudojant dispersinės analizės modelį (2.3.1), 2.3.1 skyrelyje (žr. (2.3.3)) buvo rastas besąlyginis kvadratinės formos $SS(\boldsymbol{\theta})$, priklausančios nuo nežinomų parametrų μ_{ij}, minimumas SS_E, kurį naudosime kriterijaus statistikos vardiklyje:

$$SS_E = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2 \sim \sigma^2 \chi^2_{n-m}, \tag{2.3.20}$$

čia m – skaičius tų 2.3.2 lentelės langelių, kuriuose $K_{ij} > 0$.

Kvadratinės formos $SS(\boldsymbol{\theta})$ sąlyginis minimumas $SS_{EH_{AB}}$, kai teisinga hipotezė H_{AB}, randamas minimizuojant pagal μ, α_i, β_j kvadratinę formą

$$\sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i - \beta_j)^2, \tag{2.3.21}$$

kai parametrai α_i ir β_j susieti tokiomis sąlygomis

$$\sum_{i=1}^{I} v_i\alpha_i = 0, \quad \sum_{j=1}^{J} \omega_j\beta_j = 0; \tag{2.3.22}$$

čia

$$\sum_{i=1}^{I} v_i = \sum_{j=1}^{J} \omega_j = 1, \quad v_i \geq 0, \quad \omega_j \geq 0.$$

Išreiškę iš čia, pavyzdžiui, α_I ir β_J, ir, įstatę į (2.3.21), gausime kvadratinę formą $SS(\boldsymbol{\eta})$ nuo $(I+J-1)$-mačio parametro $\boldsymbol{\eta} = (\eta_1, ..., \eta_{I+J-1})^T = (\mu, \alpha_1, ..., \alpha_{I-1}, \beta_1, ..., \beta_{J-1})^T$.

Tada ieškomasis kvadratinės formos minimumas yra

$$SS_{EH_{AB}} = \min_{\boldsymbol{\eta}} SS(\boldsymbol{\eta}) = SS(\hat{\boldsymbol{\eta}}).$$

Remiantis 1.3.2 teorema galima tvirtinti, kad kai hipotezė H_{AB} teisinga, skirtumas $SS_{AB} = SS_{EH_{AB}} - SS_E$ nepriklauso nuo SS_E ir yra pasiskirstęs kaip a. d. $\sigma^2\chi^2_{m-I-J+1}$.

Ieškant parametro $\boldsymbol{\eta}$ įvertinio $\hat{\boldsymbol{\eta}}$, reikia spręsti $I+J-1$ lygčių sistemą

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{A}^T\boldsymbol{Y},$$

kuri gaunama diferencijuojant $SS(\boldsymbol{\eta})$ pagal parametrus η_i, $i = 1, ..., I+J-1$. Matricos $\boldsymbol{A}$ išraiška gana gremėzdiška ir ji nepateikiama. Tuo labiau kad, pasinaudojus 2.3.2 teorema, pakanka spręsti paprastesnę tiesinių lygčių sistemą, kurioje lygčių skaičius yra lygus $\min(I-1, J-1)$.

2.3.2 teorema. *1. Kvadratinės formos $SS(\boldsymbol{\eta})$ minimumą $SS_{EH_{AB}}$ galima apskaičiuoti tokiu būdu*

$$SS_{EH_{AB}} = \sum_i \sum_j \sum_k Y^2_{ijk} - \sum_i \mathcal{G}_i\hat{\alpha}_i - \sum_j Y^2_{.j.}/K_{.j}, \tag{2.3.23}$$

čia

$$\mathcal{G}_i = Y_{i..} - \sum_j \frac{K_{ij}Y_{.j.}}{K_{.j}},$$

o $\hat{\alpha}_1, ..., \hat{\alpha}_{I-1}$ yra lygčių sistemos

$$c_{i1}\hat{\alpha}_1 + ... + c_{i,I-1}\hat{\alpha}_{I-1} = \mathcal{G}_i, \quad i = 1, ..., I-1,$$

$$c_{ii} = K_{i.} - \sum_j \frac{K_{ij}K_{ij}}{K_{.j}}, \quad c_{ii'} = -\sum_j \frac{K_{i'j}K_{ij}}{K_{.j}}, \quad i \neq i', \tag{2.3.24}$$

sprendinys; $\hat{\alpha}_I = 0$.

2. Kvadratinės formos $SS(\boldsymbol{\eta})$ minimumą $SS_{EH_{AB}}$ galima apskaičiuoti ir kitu būdu

$$SS_{EH_{AB}} = \sum_i \sum_j \sum_k Y_{ijk}^2 - \sum_j \mathcal{H}_j \hat{\beta}_j - \sum_i Y_{i..}^2/K_{i.}, \tag{2.3.25}$$

čia

$$\mathcal{H}_j = Y_{.j.} - \sum_i \frac{K_{ij} Y_{i..}}{K_{i.}},$$

o $\hat{\beta}_1, ..., \hat{\beta}_{J-1}$ yra lygčių sistemos

$$b_{j1}\hat{\beta}_1 + ... + b_{j,J-1}\hat{\beta}_{J-1} = \mathcal{H}_j, \quad j = 1, ..., J-1, \tag{2.3.26}$$

$$b_{jj} = K_{.j} - \sum_i \frac{K_{ij}K_{ij}}{K_{i.}}, \quad b_{jj'} = -\sum_i \frac{K_{ij'}K_{ij}}{K_{i.}}, \quad j \neq j'.$$

sprendinys; $\hat{\beta}_J = 0$.

Įrodymas. 1. Diferencijuodami kvadratinę formą (2.3.21) parametrų μ, α_i, β_j atžvilgiu ir prilyginę išvestines 0, gauname normaliųjų lygčių sistemą MK įvertiniams $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j$ rasti (žymėjimai tie patys kaip 2.3.1 ir 2.3.2 lentelėse):

$$\begin{aligned} n\hat{\mu} + \sum_i K_{i.}\hat{\alpha}_i + \sum_j K_{.j}\hat{\beta}_j &= Y_{...}, \\ K_{i.}\hat{\mu} + K_{i.}\hat{\alpha}_i + \sum_j K_{ij}\hat{\beta}_j &= Y_{i..}, \quad i = 1, ..., I, \\ K_{.j}\hat{\mu} + \sum_i K_{ij}\hat{\alpha}_i + K_{.j}\hat{\beta}_j &= Y_{.j.}, \quad j = 1, ..., J. \end{aligned} \tag{2.3.27}$$

Ši lygčių sistema yra išsigimusi ir turi be galo daug sprendinių. Iš tikrųjų, sudėję vidurines I arba paskutines J lygčių, gauname pirmąją lygtį. Tačiau, remiantis 1.2.1 teorema, galima tvirtinti, kad kvadratinės formos (2.3.21) minimumas yra toks pat su bet kuriuo sistemos (2.3.27) sprendiniu ir, naudojantis (1.4.2), gali būti užrašytas tokiu pavidalu:

$$SS_{EH_{AB}} = \sum_i \sum_j \sum_k Y_{ijk}(Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) =$$

$$\sum_i \sum_j \sum_k Y_{ijk}^2 - \hat{\mu}Y_{...} - \sum_i \hat{\alpha}_i Y_{i..} - \sum_j \hat{\beta}_j Y_{.j.}, \tag{2.3.28}$$

čia $\hat{\mu}; \hat{\alpha}_i, i = 1, ..., I; \hat{\beta}_j, j = 1, ..., J$, yra kuris nors sistemos (2.3.27) sprendinys.

Jei teisinga hipotezė H_{AB}, tai pagal 2.3.1 teoremą bet kuris kontrastas nepriklauso nuo svorių v_i, w_j sistemos, todėl kiekvienas (2.3.27) sistemos sprendinys, kai naudojama viena svorių sistema, yra šios sistemos sprendinys, kai naudojama kita svorių sistema. Taigi norint rasti kvadratinės formos $SS(\boldsymbol{\eta})$

minimumą $SS_{EH_{AB}}$, užtenka imti kokią nors svorių sistemą, rasti bent vieną (2.3.27) sistemos sprendinį ir pasinaudoti formule (2.3.28).

Sistemą (2.3.27) lengviausia spręsti imant svorių sistemą $v_1 = \cdots = v_{I-1} = w_1 = \cdots = w_{J-1} = 0$, $v_I = w_J = 1$. Tada pagal (2.3.6) $\mu = \mu_{IJ}$ ir pagal (2.3.7) bei (2.3.9) $\alpha_I = \mu_{IJ} - \mu = 0$, $\beta_J = 0$. Taigi tereikia įvertinti $I + J - 1$ parametrą $\mu, \alpha_1, \ldots, \alpha_{I-1}, \beta_1, \ldots, \beta_{J-1}$, todėl lygčių sistemoje (2.3.27) imame $I - 1$ ir $J - 1$ vietoje I ir J.

Iš paskutiniųjų (2.3.27) sistemos lygčių išreiškę $\hat{\beta}_j$:

$$\hat{\beta}_j = \frac{Y_{.j.}}{K_{.j}} - \hat{\mu} - \frac{1}{K_{.j}} \sum_{i'} K_{i'j} \hat{\alpha}_{i'} \tag{2.3.29}$$

ir, įstatę šias išraiškas į viduriniąsias lygtis, gauname lygčių sistemą

$$K_{i.}\hat{\mu} + K_{i.}\hat{\alpha}_i + \sum_j K_{ij} \left[\frac{Y_{.j.}}{K_{.j}} - \hat{\mu} - \frac{1}{K_{.j}} \sum_{i'} K_{i'j} \hat{\alpha}_{i'} \right] = Y_{i..}, \quad i = 1, ..., I-1,$$

kurią galime užrašyti naudodami teoremos formuluotės žymėjimus:

$$K_{i.}\hat{\alpha}_i - \sum_{i'} \left(\sum_j \frac{K_{ij} K_{i'j}}{K_{.j}} \right) \hat{\alpha}_{i'} = \mathcal{G}_i, \quad i = 1, ..., I-1.$$

Tai ir yra lygčių sistema (2.3.24).

Įstatę $\hat{\beta}_j$ išraišką (2.3.29) į (2.3.28), įsitikiname, kad $SS_{EH_{AB}}$ turi (2.3.23) pavidalą.

2. Įrodoma analogiškai. Iš viduriniųjų (2.3.27) sistemos lygčių išreiškiame $\hat{\alpha}_i$ ir įstatome į paskutiniąsias lygtis. Gauname lygčių sistemą (2.3.26). Gautas $\hat{\alpha}_i$ išraiškas įstatę į (2.3.28), įsitikiname, kad $SS_{EH_{AB}}$ turi ir (2.3.25) pavidalą. ▲

Grįžtame prie hipotezės H_{AB} tikrinimo. Remdamiesi 1.3.2 teorema ir panaudoję gautas liekamųjų kvadratų sumų išraiškas (2.3.20) ir (2.3.24) arba (2.3.25) gauname, kad hipotezė H_{AB} atmetama reikšmingumo lygmens α kriterijumi, kai

$$F_{AB} = \frac{(SS_{EH_{AB}} - SS_E)(n-m)}{(m-I-J+1)SS_E} = \frac{MS_{AB}}{MS_E} > F_\alpha(m-I-J+1, n-m). \tag{2.3.30}$$

Matome, kad netuščių langelių skaičius m lentelėje 2.3.2 turi tenkinti nelygybę $m > I + J - 1$.

2.3.3 pastaba. Jeigu eksperimentų planas subalansuotas ir $K_{ij} = K > 1$ su visais $i = 1, ..., I$ ir $j = 1, ..., J$, tai iš lygčių sistemos (2.3.24) imdami vienodus svorius, ir naudodami sąlygą $\sum_i \alpha_i = 0$, gauname įvertinius

$$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...},$$

ir kvadratinė forma

$$SS_{EH_{AB}} = \sum_i \sum_j \sum_k (\bar{Y}_{i..} - \bar{Y}_{...})^2 + K \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 = SS_E + SS_{AB}$$

sutampa su gautąja 2.2.3 skyrelyje.

2.3.5. Faktorių įtakos nebuvimo hipotezės adityviajame modelyje tikrinimas

Nagrinėkime *adityvųjį modelį*:

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad k = 1, ..., K_{ij}, \quad i = 1, ..., I, \quad j = 1, ..., J;$$

paklaidos e_{ijk} yra n. a. d., turintys normalųjį skirstinį $e_{ijk} \sim N(0, \sigma^2)$, o vidurkiai μ_{ij} tenkina tiesinį modelį (2.3.21):

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \quad \sum_{i=1}^{I} v_i \alpha_i = 0, \quad \sum_{j=1}^{J} \omega_j \beta_j = 0; \tag{2.3.31}$$

čia $\{v_i\}$ ir $\{\omega_j\}$ – kokios nors svorių sistemos, tenkinančios (2.3.5) sąlygas.

Pagrindinės dispersinės analizės hipotezės

$$H_A : \alpha_1 = ... = \alpha_I, \quad H_B : \beta_1 = ... = \beta_J \tag{2.3.32}$$

yra faktorių A ir B įtakos nebuvimo hipotezės.

Kriterijų vėl kuriame remdamiesi 1.3.2 teorema. Besąlyginį kvadratinės formos

$$\sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i - \beta_j)^2 \tag{2.3.33}$$

minimumą modeliui (2.3.31), kurį naudosime kriterijaus statistikos vardiklyje, suradome 2.3.2 teoremoje, nes, kaip jau buvo pažymėta, jis sutampa su kvadratinės formos $SS(\boldsymbol{\eta})$ minimumu $SS_{EH_{AB}}$, kurio išraiška pateikta formulėse (2.3.23) arba (2.3.25).

Lieka rasti sąlyginį kvadratinės formos (2.3.33) minimumą, kai teisinga hipotezė H_A arba H_B.

2.3.3 teorema. *Kvadratinės formos (2.3.33) sąlyginis minimumas, kai teisinga hipotezė H_B, yra*

$$SS_{EH_B} = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{i..})^2, \tag{2.3.34}$$

o sąlyginis minimumas, kai teisinga hipotezė H_A, yra

$$SS_{EH_A} = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{.j.})^2; \tag{2.3.35}$$

čia $\bar{Y}_{i.}$ *ir* $\bar{Y}_{.j.}$ *yra atitinkamai 2.3.1 lentelės* i*-osios eilutės ir* j*-ojo stulpelio stebėjimų aritmetiniai vidurkiai*

$$\bar{Y}_{i.} = \frac{1}{K_{i.}} Y_{i..}, \quad \bar{Y}_{.j.} = \frac{1}{K_{.j}} Y_{.j.}.$$

Tariama, kad kiekvienoje eilutėje ir kiekviename stulpelyje yra nors po vieną stebėjimą, t. y. $K_{i.} > 0$, $K_{.j} > 0$.

Jeigu hipotezės H_A *arba* H_B *teisingos, tai surastosios statistikos atitinkamai turi tokius skirstinius*

$$SS_{EH_B} \sim \sigma^2 \chi^2_{m-J}, \quad SS_{EH_A} \sim \sigma^2 \chi^2_{m-I}. \tag{2.3.36}$$

Įrodymas. Turime

$$SS_{EH_B} = \min_{\mu,\alpha_i,\beta_j=0} \sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i - \beta_j)^2 = \min_{\mu_i} \sum_i \sum_j \sum_k (Y_{ijk} - \mu_i)^2, \tag{2.3.37}$$

jeigu pažymėsime $\mu_i = \mu + \alpha_i$, $i = 1, ..., I$. Esant teisingai hipotezei H_B, stebėjimai, surašyti 2.3.1 lentelės i-oje eilutėje, turi vienodus vidurkius $\mu_i = \mathbf{E}(Y_{ijk})$, $\forall k = 1, ..., K_{ij}$, ir $\forall j = 1, ..., J$. Taigi 2.3.1 lentelę galima traktuoti kaip viefaktorės dispersinės analizės stebėjimų 2.1.1 lentelę. Skirtumas tik tas, kad stebėjimai kitaip sunumeruoti ir bendras i-osios eilutės stebėjimų skaičius yra $K_{i.}$ (vietoje J_i 2.1.1 lentelėje). Todėl sąlyginis minimumas SS_{EH_B} sutampa su besąlyginiu minimumu, surastu 2.1.1 skyrelyje (buvo žymimas SS_E), atlikus atitinkamus žymėjimų pakeitimus. Gauname išraišką (2.3.35).

Analogiškai, kai teisinga hipotezė H_A, visi stebėjimai, surašyti 2.3.1 lentelės j-ame stulpelyje, turi vienodus vidurkius $\mathbf{E}(Y_{ijk}) = \mu_j = \mu + \beta_j$, $\forall k = 1, ..., K_{ij}$, ir $\forall i = 1, ..., I$. Taigi 2.3.1 lentelės duomenis galime traktuoti kaip duomenis vienfaktorės dispersinės analizės schemoje, kurioje faktoriaus lygmenų skaičius yra J, o stebėjimų skaičius, kai yra faktoriaus j-asis lygmuo, yra $K_{.j}$, $j = 1, ..., J$. Sąlyginis kvadratinės formos minimumas sutampa su besąlyginiu kvadratinės formos minimumu tokioje vienfaktorės dispersinės analizės schemoje. Gauname formulę (2.3.35).

Tvirtinimai (2.3.36) sutampa su atitinkamais 2.1.1 skyrelio tvirtinimais minėtose vienfaktorės dispersinės analizės schemose. ▲

Grįžtame prie hipotezių H_A ir H_B tikrinimo. Remdamiesi 1.3.2 teorema sudarome statistikas

$$F_A = \frac{(SS_{EH_A} - SS_{EH_{AB}})(n - I - J + 1)}{(I-1) SS_{EH_{AB}}},$$

$$F_B = \frac{(SS_{EH_B} - SS_{EH_{AB}})(n - I - J + 1)}{(J-1) SS_{EH_{AB}}}; \tag{2.3.38}$$

čia $SS_{EH_{AB}}$ imame iš 2.3.2 teoremos (formulės (2.3.23) arba (2.3.25)).

Hipotezės H_A arba H_B atmetamos reikšmingumo lygmens α kriterijumi, kai atitinkamai teisingos nelygybės

$$F_A > F_\alpha(I-1, n-I-J+1), \quad F_B > F_\alpha(J-1, n-I-J+1). \tag{2.3.39}$$

Jeigu hipotezė H_A arba H_B atmetama, tai, ieškant kontrastų, kurie atsakingi už hipotezės atmetimą, galima naudoti S metodą.

Nagrinėkime kontrastą

$$\psi = \sum_{j=1}^{J} c_j \beta_j, \quad \sum_{j=1}^{J} c_j = 0.$$

Jo MK įvertinys

$$\hat{\psi} = \sum_{j=1}^{J-1} c_j \hat{\beta}_j = \boldsymbol{c}^T \hat{\boldsymbol{\beta}},$$

čia $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_{J-1})^T$ yra lygčių sistemos (2.3.26) sprendinys; $\boldsymbol{c} = (c_1, ..., c_{J-1})^T$.

Pažymėkime $\boldsymbol{B} = [b_{ij}]_{(J-1)\times(J-1)}$ lygčių sistemos (2.3.26) pagrindinę matricą ir $\boldsymbol{H} = (\mathcal{H}_1, ..., \mathcal{H}_{J-1})^T$ – atsitiktinį vektorių, sudarytą iš dešinėje lygčių sistemos (2.3.26) parašytų atsitiktinių dydžių. Tada lygčių sistemą (2.3.26) galime užrašyti matriciniu pavidalu

$$\boldsymbol{B}\hat{\boldsymbol{\beta}} = \boldsymbol{H}, \quad \hat{\boldsymbol{\beta}} = \boldsymbol{B}^{-1}\boldsymbol{H}.$$

Tiesiogiai įsitikiname, kad a. v. $\boldsymbol{H}$ kovariacinė matrica $\boldsymbol{V}(\boldsymbol{H}) = \sigma^2 \boldsymbol{B}$. Todėl

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{B}^{-1}, \quad \boldsymbol{V}(\hat{\psi}) = \sigma^2 \boldsymbol{c}^T \boldsymbol{B}^{-1} \boldsymbol{c}.$$

Dispersijos σ^2 įvertiniu imdami

$$\hat{\sigma}^2 = s^2 = SS_{EH_{AB}}/(n-I-J+1)$$

gauname kontrasto ψ įvertinio $\hat{\psi}$ dispersijos įvertinį

$$\hat{\boldsymbol{V}}(\hat{\psi}) = s^2 \boldsymbol{c}^T \boldsymbol{B}^{-1} \boldsymbol{c}. \tag{2.3.40}$$

Sudarant pasikliovimo intervalus (2.1.14) reikia imti

$$\Delta^2 = (J-1) F_\alpha(J-1, n-I-J+1).$$

Analogiškai tiriame kontrastus, skirtus palyginti faktoriaus A lygmenų įtaką.

2.3.3 pastaba. Praktiškai (2.3.31) adityvusis modelis yra svarbus dėl tokių priežasčių. Jeigu modelis neadityvusis, tai parametrų, aprašančių vidurkius, skaičius yra IJ. Norint juos visus įvertinti, reikia turėti bent po vieną stebėjimą kiekviename lentelės 2.3.1 langelyje ir pakartotinius stebėjimus nors viename langelyje, kad būtų galima įvertinti dispersiją. Taigi būtinas minimalus stebėjimų

skaičius yra ne mažesnis už $IJ+1$. Kartais iš principo negalima atlikti stebėjimų prie visų faktorių A ir B lygmenų rinkinių (A_i, B_j). Be to, toks eksperimentas gali būti praktiškai negalimas dėl laiko ar lėšų stokos, jeigu eksperimentai yra brangūs arba jiems atlikti reikia daug laiko. Adityviajame modelyje vidurkio kitimą nusako $I+J-1$ parametras. Tinkamai planuojant eksperimentą adityviajame modelyje reikalingas būtinas eksperimentų skaičius yra daug mažesnis ir palyginamas su $I + J$. Tada 2.3.1 lentelės dauguma langelių bus tušti ir tik kai kurie langeliai bus užpildyti matavimais. Tokie eksperimentų planai vadinami *nepilnais* planais. Detaliau nepilni eksperimentų planai nagrinėjami 3.8 skyrelyje.

2.3.6. Faktorių įtakos nebuvimo hipotezės neadityviajame modelyje tikrinimas

Nagrinėsime *neadityvųjį modelį:*

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad k = 1, ..., K_{ij}, \quad i = 1, ..., I, \quad j = 1, ..., J,$$

arba pavidalo

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad k = 1, ..., K_{ij}, \quad i = 1, ..., I, \quad j = 1, ..., J.$$

Atsižvelgiant į sąlygas (2.3.10) ir (2.3.13), abiem atvejais nežinomų parametrų, aprašančių vidurkių kitimą skaičius yra IJ.

Tikrinsime faktorių A ir B lygmenų įtakos nebuvimo hipotezes

$$H_A : \alpha_1 = ... = \alpha_I = 0, \quad H_B : \beta_1 = ... = \beta_J = 0,$$

nedarydami jokių prielaidų apie parametrus γ_{ij}, išskyrus sąlygą (2.3.13). Jeigu nors vienas 2.3.1 lentelės langelis yra tuščias, tai negalime įvertinti visų parametrų μ_{ij}. Tai reiškia, kad negalime daryti išvadų apie visus parametrus α_i arba β_j. Todėl tarsime, kad visi 2.3.1 lentelės langeliai yra užpildyti, t. y. $K_{ij} \geq 1$, $\forall i = 1, ..., I,\ j = 1, ..., J$.

Besąlyginis kvadratinės formos

$$\sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 = \sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2 \qquad (2.3.41)$$

minimumas surastas 2.3.4 skyrelyje (žr. (2.3.3)):

$$SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 \qquad (2.3.42)$$

Kad SS_E nebūtų lygi 0, reikia kad nors viename 2.3.1 lentelės langelyje būtų pakartotini stebėjimai. Tada galima įvertinti dispersiją σ^2:

$$SS_E \sim \sigma^2 \chi^2_{n-IJ}, \quad \hat{\sigma}^2 = s^2 = \frac{SS_E}{n - IJ} = MS_E. \qquad (2.3.43)$$

Remiantis 1.3.2 teorema, tikrinant hipotezę H_A, reikia rasti sąlyginį kvadratinės formos (2.3.41) minimumą SS_{EH_A}, kai teisinga hipotezė H_A, ir skirtumą $SS_A = SS_{EH_A} - SS_E$. Statistikos SS_A skirstinys nepriklauso nuo SS_E ir, jeigu H_A teisinga, tai

$$SS_A \sim \sigma^2 \chi^2_{I-1}. \tag{2.3.44}$$

Analogiškai, tikrinant hipotezę H_B, reikia rasti sąlyginį (2.3.41) minimumą SS_{EH_B}, kai H_B teisinga, ir skirtumą $SS_B = SS_{EH_B} - SS_E$. Statistikos SS_B skirstinys nepriklauso nuo SS_E ir, jeigu H_B teisinga, tai

$$SS_B \sim \sigma^2 \chi^2_{J-1}. \tag{2.3.45}$$

2.3.4 teorema. *1. Statistiką SS_A galima apskaičiuoti šitaip:*

$$SS_A = \sum_{i=1}^{I} \theta_i (\hat{a}_i - \bar{a})^2; \tag{2.3.46}$$

$$\theta_i = \{\sum_{j=1}^{J} (\omega_j^2 / K_{ij})\}^{-1}, \quad \hat{a}_i = \sum_{j=1}^{J} \omega_j \bar{Y}_{ij.}, \quad \bar{a} = \sum_{i=1}^{I} \theta_i \hat{a}_i / \sum_{i=1}^{I} \theta_i.$$

2. Statistiką SS_B galima apskaičiuoti taip:

$$SS_B = \sum_{j=1}^{J} \eta_j (\hat{b}_j - \bar{b})^2; \tag{2.3.47}$$

$$\eta_j = \{\sum_{i=1}^{I} (v_i^2 / K_{ij})\}^{-1}, \quad \hat{b}_j = \sum_{i=1}^{I} v_i \bar{Y}_{ij.}, \quad \bar{b} = \sum_{j=1}^{J} \eta_j \hat{b}_j / \sum_{j=1}^{J} \eta_j.$$

Įrodymas. 1. Remsimės 2.1.1 pastaba, pagal kurią

$$SS_A = \hat{\psi}^2_{\max}; \tag{2.3.48}$$

čia $\hat{\psi}^2_{max}$ yra normuotų kontrastų įvertinių kvadratų maksimumas:

$$\hat{\psi}^2_{\max} = \max\{\hat{\psi}^2 : \hat{\psi} = \sum_{i=1}^{I} c_i \hat{\alpha}_i, \quad \sum_{i=1}^{I} c_i = 0, \quad \boldsymbol{V}(\hat{\psi}) = \sigma^2\}. \tag{2.3.49}$$

Nagrinėkime kontrastą

$$\psi = \sum_{i=1}^{I} c_i \alpha_i = \sum_{i=1}^{I} (\sum_{j=1}^{J} \omega_j \mu_{ij}).$$

Jo MK įvertinys ir įvertinio dispersija yra

$$\hat{\psi} = \sum_{i=1}^{I} c_i (\sum_{j=1}^{J} \omega_j \bar{Y}_{ij.}), \quad \boldsymbol{V}\hat{\psi} = \sigma^2 \sum_{i=1}^{I} c_i^2 (\sum_{j=1}^{J} (\omega_j^2 / K_{ij})).$$

Pažymėję

$$\hat{a}_i = \sum_{j=1}^{J} \omega_j \bar{Y}_{ij.}, \quad \theta_i = \{\sum_{j=1}^{J} (\omega_j^2 / K_{ij})\}^{-1},$$

ir atsižvelgę į (2.3.49), gauname, kad, ieškant $SS_A = \hat{\psi}^2_{\max}$, reikia maksimizuoti

$$\hat{\psi} = \sum_{i=1}^{I} c_i \hat{a}_i \tag{2.3.50}$$

parenkant $c_1, ..., c_I$ taip, kad jie tenkintų apribojimus

$$\sum_{i=1}^{I} c_i = 0, \quad \sum_{i=1}^{I} (c_i^2 / \theta_i) = 1. \tag{2.3.51}$$

Remiantis Lagranžo neapibrėžtinių daugiklių metodu, reikia maksimizuoti

$$f = \sum_{i=1}^{I} c_i \hat{a}_i - \lambda_1 \sum_{i=1}^{I} c_i - \lambda_2 \sum_{i=1}^{I} (c_i^2 / \theta_i).$$

Diferencijuodami pagal c_i ir prilyginę išvestines 0, gauname lygčių sistemą

$$\frac{\partial f}{\partial c_i} = \hat{a}_i - \lambda_1 - 2\lambda_2 \frac{c_i}{\theta_i} = 0, \quad i = 1, ..., I.$$

Išsprendę c_i atžvilgiu ir naudodami (2.3.51) lygybes galime eliminuoti parametrus λ_1 ir λ_2:

$$c_i = \frac{\theta_i(\hat{a}_i - \lambda_1)}{2\lambda_2}, \quad i = 1, ..., I;$$

$$\sum_{i=1}^{I} c_i = \frac{1}{2\lambda_2} \sum_{i=1}^{I} \theta_i(\hat{a}_i - \lambda_1) = 0 \Rightarrow \lambda_1 = \bar{a} = (\sum_{i=1}^{I} \theta_i \hat{a}_i /) \sum_{i=1}^{I} \theta_i;$$

$$\sum_{i=1}^{I} \frac{c_i^2}{\theta_i} = \frac{1}{(2\lambda_2)^2} \sum_{i=1}^{I} \theta_i(\hat{a}_i - \bar{a})^2 = 1 \Rightarrow 2\lambda_2 = \{\sum_{i=1}^{I} \theta_i(\hat{a}_i - \bar{a})^2\}^{1/2}.$$

Galutinai gauname

$$c_i = \theta_i(\hat{a}_i - \bar{a}) / \{\sum_{i=1}^{I} \theta_i(\hat{a}_i - \bar{a})^2\}^{1/2}, \quad i = 1, ..., I.$$

Gautąsias išraiškas įstatome į (2.3.50). Gauname

$$\hat{\psi} = \sum_{i=1}^{I} c_i \hat{a}_i = \sum_{i=1}^{I} c_i(\hat{a}_i - \bar{a}) = \{\sum_{i=1}^{I} \theta_i(\hat{a}_i - \bar{a})^2\}^{1/2}.$$

Tokiu būdu,

$$SS_A = \hat{\psi}^2_{\max} = \sum_{i=1}^{I} \theta_i (\hat{a}_i - \bar{a})^2.$$

2. Įrodoma analogiškai, nagrinėjant kontrastą

$$\psi = \sum_{j=1}^{J} c_j \beta_j = \sum_{j=1}^{J} (\sum_{i=1}^{I} v_i \mu_{ij}).$$

▲

Grįžtame prie hipotezių H_A ir H_B tikrinimo. Remdamiesi 1.3.2 teorema apibrėžiame statistikas

$$F_A = \frac{(n-IJ)SS_A}{(I-1)SS_E} = \frac{MS_A}{MS_E}, \quad F_B = \frac{(n-IJ)SS_B}{(J-1)SS_E} = \frac{MS_B}{MS_E}. \tag{2.3.52}$$

Hipotezės H_A arba H_B atmetamos reikšmingumo lygmens α kriterijumi, kai atitinkamai patenkintos nelygybės

$$F_A > F_\alpha(I-1, n-IJ), \quad F_B > F_\alpha(J-1, n-IJ). \tag{2.3.53}$$

Jei hipotezė H_A atmetama, tai kontrastų, skirtų palyginti faktoriaus A lygmenis, analizei galima pritaikyti S metodą. Kontrasto

$$\psi = \sum_{i=1}^{I} c_i \alpha_i = \sum_{i=1}^{I} c_i (\sum_{j=1}^{J} \omega_j \mu_{ij}) = \sum_{i=1}^{I} c_i a_i$$

įvertinys ir įvertinio dispersijos įvertinys yra

$$\hat{\psi} = \sum_{i=1}^{I} c_i \hat{a}_i, \quad \hat{\boldsymbol{V}}(\hat{\psi}) = s^2 \sum_{i=1}^{I} (c_i^2/\theta_i), \quad s^2 = \frac{SS_E}{n-IJ} = MS_E.$$

Sudarant pasikliovimo intervalus (2.1.17) reikia imti $\Delta^2 = (I-1)F_\alpha(I-1, n-IJ)$.

Analogiškai nagrinėjame kontrastus, skirtus faktoriaus B lygmenims palyginti.

2.3.4 pastaba. Gautosios kvadratų sumos (2.3.46) ir (2.3.47) priklauso nuo parinktos svorių sistemos $\{v_i\}, \{\omega_j\}$. Pavyzdžiui, naudojant lygių svorių sistemą $v_i = 1/I$, $\omega_j = 1/J$, formulėje (2.3.46) imama

$$\hat{a}_i = \sum_j \bar{Y}_{ij.}/J, \quad \theta_i = J^2 / \sum_j (1/K_{ij}).$$

Jeigu svorių sistema dažnuminė, t. y. $v_i = K_{i.}/n$, $\omega_j = K_{.j}/n$, tai

$$\hat{a}_i = \sum_j K_{.j} \bar{Y}_{ij.}/n, \quad \theta_i = n^2 / \sum_j (K_{.j}^2/K_{ij}).$$

Jeigu eksperimentų planas subalansuotas ir $K_{ij} = K > 1$, tai abiejų svorių sistemų atveju

$$\hat{a}_i = \bar{Y}_{i..}, \quad \theta_i = JK.$$

Kvadratų suma (2.3.46) yra

$$SS_A = JK \sum_{i=1}^{I} (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

ir sutampa su ta, kuri gauta 2.2.1 teoremoje.

2.3.1 pavyzdys. Eksperimente hibridinius žiurkiukus maitino hibridinės žiurkių patelės. 2.3.3 lentelėje (2.3.1 lentelės analogas) pateikti žiurkiukų svoriai praėjus 28 dienoms nuo gimimo. Šiame eksperimente faktorius A yra maitinančios žiurkės genotipas, o faktorius B – žiurkiukų vados genotipas.

2.3.3 lentelė. Stebėjimo duomenys

$i\ j$	B_1	B_2	$\sum$
A_1	61,5;68,2;64,0;65,0;59,7	37,0;36,3;68,0	459,7
A_2	55,0;42,0;60,2	56,3;69,8;67,0	350,3
A_3	52,5;61,8;49,5;52,7	39,6;46,0;61,3;55,3;55,7	474,4
A_4	42,0;54,0;61,0;48,2;39,6	50,0;43,8;54,5	393,1
$\sum$	936,9	740,6	1677,5

Stebėjimų skaičiai langeliuose yra skirtingi. Jie pateikti 2.3.4 lentelėje (2.3.2 lentelės analogas).

2.3.4 lentelė. Stebėjimų skaičiai

$i\ j$	1	2	$\sum$
1	5	3	8
2	3	3	6
3	4	5	9
4	5	3	8
$\sum$	17	14	31

Iš pradžių patikrinsime modelio adityvumo hipotezę. Pagal (2.3.20) apskaičiuojame $SS_E = 1725,25$ (laisvės laipsnių skaičius $n - m = 23$). Kadangi $J < I$, tai kvadratų sumą $SS_{EH_{AB}}$ randame pagal (2.3.26). Lygčių sistema (2.3.26) susideda tik iš vienos lygties. Randame $\mathcal{H}_1 = 17,91$, $b_{11} = 7,47$ ir $\hat{\beta}_1 = 2,396$. Kvadratų suma $SS_{EH_{AB}} = 2427,39$. Pagal (2.3.30) randame $F_{AB} = 3,12$. Jeigu hipotezė H_{AB} teisinga, tai F_{AB} turi Fišerio skirstinį su 3 ir 23 laisvės laipsniais. Gauname, kad P reikšmė, t.y. $\mathbf{P}\{F_{3,23} > 3,12\} = 0,0457$. Hipotezė H_{AB} atmetama, jeigu kriterijaus reikšmingumo lygmuo $\alpha > 0,0457$.

Patikrinsime hipotezes H_A ir H_B dėl faktorių A ir B įtakos.

Pradžioje tarsime, kad modelis adityvusis. Pagal (2.3.34) ir (2.3.35) apskaičiuojame statistikų SS_{EH_A} ir SS_{EH_B} realizacijas ir skirtumus $SS_{EH_A} - SS_{EH_{AB}} = 420,250$, $SS_{EH_B} - SS_{EH_{AB}} = 42,907$. Pagal (2.3.38) randame, kad statistikos F_A ir F_B įgijo atitinkamai reikšmes 1,428 ir 0,272. Jeigu H_A ir H_B teisingos ir teisinga adityvumo prielaida, tai statistikos F_A ir F_B turi Fišerio skirstinius atitinkamai su laisvės laipsnių skaičiais 3; 26 ir 1; 26. Hipotezes atmesti nėra pagrindo, nes P reikšmės: $\mathbf{P}\{F_{3,26} > 1,5\} = 0,238$, $\mathbf{P}\{F_{1,26} > 0,460\} = 0,504$.

Kadangi adityvumo prielaida kelia abejonių, tai patikrinsime tas pačias hipotezes nedarydami adityvumo prielaidos. Minėjome, kad tokiu atveju atsakymas iš dalies priklauso nuo parinktos svorių sistemos.

Pirmiausia parinkime lygių svorių sistemą $v_i = 1/I$, $\omega_j = 1/J$. Pagal formules (2.3.46) ir (2.3.47) randame statistikų SS_A ir SS_B realizacijas. Gauname $SS_A = 311,50$, $SS_B = 20,47$. Remdamiesi (2.3.52) apskaičiuojame statistikų F_A ir F_B realizacijas.

Gauname $F_A = 311,50 \cdot 23/(3 \cdot 1725,25) = 1,384$ ir $F_B = 0,273$. Hipotezių atmesti nėra pagrindo; P reikšmės atitinkamai yra 0,273 ir 0,606.

Parinkę dažnuminius svorius $v_i = K_{i.}/n$, $\omega_j = K_{.j}/n$, analogiškai randame $F_A = 1,390$ ir $F_B = 0,662$. Hipotezių atmesti nėra pagrindo.

2.4. Vienfaktorė dispersinė analizė, kai faktorius atsitiktinis

2.4.1. Statistinis modelis

Skyrelyje 2.1.1 nagrinėtas *dispersinės analizės, kai faktorius pastovus*, modelis

$$Y_{ij} = \mu_i + e_{ij}, \quad i = 1, ..., I; \quad j = 1, ..., J_i,$$

naudojamas, kai faktoriaus A lygmenys parenkami eksperimentatoriaus ir imtys imamos iš populiacijų, turinčių fiksuotus faktoriaus lygmenis $A_1, \ldots A_I$. Eksperimento tikslas yra palyginti tų populiacijų požymio Y vidurkius. Pagrindinis uždavinys: patikrinti hipotezę $H_A : \mu_1 = \cdots = \mu_I$, kad, eliminavus atsitiktines paklaidas, požymio Y reikšmės būtų vienodos su visomis I faktoriaus reikšmėmis.

Dispersinės analizės, kai faktorius atsitiktinis, modelis naudojamas, kai keli faktoriaus lygmenys parenkami *atsitiktinai* iš visų galimų lygmenų aibės ir norima padaryti bendresnes išvadas apie visus galimus faktoriaus lygmenis, o ne tik apie parinktus eksperimento metu.

Dispersinės analizės su atsitiktiniais faktoriais modelis turi tą pačią formą

$$Y_{ij} = X_i + e_{ij}, \quad j = 1, ..., J_i, \quad i = 1, ..., I,$$

tiktai komponentės $X_1, \ldots, X_I$ šiame modelyje interpretuojamos kaip vienodai pasiskirstę nepriklausomi a. d. $N(\mu, \sigma_A^2)$, tariant, kad $e_{ij} \sim N(0, \sigma^2)$ nepriklausomi tarpusavyje ir nepriklauso nuo a. d. $X_1, ..., X_I$.

Pagrindinis uždavinys: įsitikinti, ar, eliminavus atsitiktines paklaidas, požymio Y reikšmės būtų vienodos imant visas galimas faktoriaus reikšmes, o ne tik naudotas eksperimente. Tai ekvivalentu teiginiui $H_A : \sigma_A^2 = 0$.

Modelių su fiksuotais ir atsitiktiniais faktoriais skirtumus paaiškinsime pavyzdžiais.

2.4.1 pavyzdys. Tarkime, kad eksperimento tikslas yra palyginti iš penkių konkrečių medelyno vietų A_1, A_2, A_3, A_4 ir A_5 paimtų sodinukų augimo tempus. Imama po kelis sodinukus iš kiekvienos vietos, jie pasodinami vienoje vietoje ir po metų pamatuojamas jų prieaugis. Prieaugį nusako dvi komponentės: jis gali priklausyti nuo sėklos daiginimo vietos, be to, nuo atsitiktinės komponentės, nes netgi vienoje vietoje daiginti sodinukai auga nevienodai, turi paveldėtų skirtumų. Šiuo atveju naudojamas modelis su fiksuotais faktoriais: faktoriaus A reikšmės yra fiksuotos ir lyginamas tiktai tose penkiose vietose daigintų sodinukų augimas.

Modelis su atsitiktiniais faktoriais yra naudojamas, jei norima įsitikinti, ar apskritai augimas priklauso nuo sodinuko daiginimo vietos. Tuo atveju atsitiktinai parenkamos kelios medelyno vietos ir po kelis sodinukus iš kiekvienos vietos.

2.4.2 pavyzdys. Jei tikslas yra palyginti trijose konkrečiose gamyklose pagaminintų "Sony" televizorių patikimumą, naudojamas modelis su fiksuotais faktoriais. Jei norima įsitikinti, ar šios markės televizorių patikimumas apskritai nepriklauso nuo pagaminimo vietos, naudojamas modelis su atsitiktiniais faktoriais.

2.4.3 pavyzdys. Tarkime, iš gaminių aibės atsitiktinai atrenkame I gaminių ir J_i kartų pamatuojamos (su galima atsitiktine paklaida) i-ojo gaminio parametro X reikšmės. Tada Y_{ij} yra matavimo rezultatas, gautas su paklaida e_{ij} pamatavus i-ojo gaminio parametro X reikšmę X_i.

Analizės tikslas nėra tų kelių eksperimente dalyvavusių gaminių palyginimas, o technologinio proceso, kuris generuoja nagrinėjamą parametrą, apibūdinimas, t.y. parametro reikšmių sklaidos nustatymas visų gaminių, kurie galėjo būti pagaminti analogiškomis sąlygomis, aibėje.

Modelyje su fiksuotais faktoriais (I modelyje) Y_{ij} skirstinius visiškai nusako I vidurkių $\mu_1, ..., \mu_I$ ir viena bendra dispersija σ^2, o modelyje su atsitiktiniais faktoriais (II modelyje) – vienas bendras vidurkis μ ir dvi dispersijos σ_A^2 ir σ^2. Taigi visus paskutinio modelio uždavinius galima formuluoti parametrų $\mu, \sigma_A^2, \sigma^2$ terminais.

Kitas šių modelių skirtumas yra tas, kad II modelio atsitiktiniai dydžiai Y_{ij} ir $Y_{ij'}$ yra priklausomi net kai $j \neq j'$:

$$\mathbf{Cov}\,(Y_{ij},\ Y_{i'j'}) = \begin{cases} \sigma_A^2 + \sigma^2, & \text{kai} \quad i = i', \quad j = j', \\ \sigma_A^2, & \text{kai} \quad i = i', \quad j \neq j', \\ 0, & \text{kai} \quad i \neq i'. \end{cases}$$

2.4.2. Parametrų įvertiniai ir hipotezių tikrinimas

2.4.2.1 Kvadratų sumų skirstiniai

Tarsime, kad $J_i = J$ su visais i. Apibrėžkime kvadratų sumas kaip ir I modelyje:

$$SS_E = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2, \quad SS_A = J\sum_{i=1}^{I}(\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

2.4.1 teorema. *Kvadratų sumos SS_E ir SS_A yra nepriklausomos ir turi tokius skirstinius*

$$\frac{SS_E}{\sigma^2} \sim \chi^2(I(J-1)), \quad \frac{SS_A}{\tau^2} \sim \chi^2(I-1), \quad \tau^2 = \sigma^2 + J\sigma_A^2. \tag{2.4.1}$$

Įrodymas. Įrašykime į SS_E ir SS_A a. d. Y_{ij} išraiškas

$$Y_{ij} = \mu + Z_i + e_{ij}, \quad \mu = \mathbf{E}X_i, \quad Z_i = X_i - \mu \sim N(0, \sigma_A^2).$$

Gauname

$$SS_E = \sum_i\sum_j(e_{ij} - \bar{e}_{i.})^2, \quad SS_A = J\sum_i(Z_i + \bar{e}_{i.} - \bar{Z}_{.} - \bar{e}_{..})^2.$$

Kvadratinės formos SS_A ir SS_E yra nepriklausomos, nes pagal prielaidas a. d., žymimi skirtingomis raidėmis, yra nepriklausomi, o a. d. $e_{ij}-\bar{e}_{i.}$ ir $\bar{e}_{i'.}-\bar{e}_{..}$, $i,\ i'=1,...,I,\ j=1,...,J$, yra nekoreliuoti.

Kadangi $e_{ij}\sim N(0,\sigma^2)$, tai vidurinioji SS_E suma yra pasiskirsčiusi kaip a. d. χ^2_{J-1}, padaugintas iš σ^2; todėl $SS_E\sim\sigma^2\chi^2_{I(J-1)}$. A. d. $Z_i+\bar{e}_{i.}$ yra nepriklausomi ir vienodai pasiskirstę pagal $N(0,\sigma_A^2+\sigma^2/J)$. Gauname

$$SS_E\sim\sigma^2\chi^2_{I(J-1)},\quad SS_A\sim J(\sigma_A^2+\sigma^2/J)\chi^2_{I-1},$$

o tai ekvivalentu (2.4.1). ▲

2.4.1 išvada Pažymėkime vidutines kvadratų sumas $MS_A=SS_A/(I-1)$ ir $MS_E=SS_E/(I(J-1))$. Tada

$$\frac{\sigma^2 MS_A}{MS_E(\sigma^2+J\sigma_A^2)}\sim F(I-1,\ I(J-1)),\quad \mathbf{E}(MS_E)=\sigma^2,\quad \mathbf{E}(MS_A)=\tau^2.$$

Remdamiesi išvada, standartiniu būdu galime tikrinti hipotezes apie parametrų σ^2 ir τ^2, $\delta^2=\sigma^2/(\sigma^2+J\sigma_A^2)$ ir $\gamma^2=\sigma_A^2/\sigma^2$ reikšmes ir sudaryti jų pasikliovimo intervalus.

Surašę skaičiavimo rezultatus į dispersinės analizės lentelę, matome, kad ji skiriasi nuo 2.1.2 lentelės tik paskutiniuoju stulpeliu.

2.4.1 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	$MS=SS/\nu$	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2+J\sigma_A^2=\tau^2$
E	SS_E	$I(J-1)$	MS_E	σ^2
T	$SS_T=SS_A+SS_E$	$n-1$	–	–

2.4.2.2 Faktoriaus įtakos nebuvimo hipotezės tikrinimas

Priminsime, kad pagrindinė dispersinės analizės su atsitiktiniu faktoriumi hipotezė yra $H_A:\sigma_A^2=0$. Jei teisinga ši hipotezė, tai pagal 2.4.1 išvadą

$$F_A=\frac{MS_A}{MS_E}$$

pasiskirsčiusi pagal Fišerio dėsnį su $I-1$ ir $I(J-1)$ laisvės laipsnių.

Hipotezė H_A atmetama reikšmingumo lygmens α kriterijumi, kai teisinga nelygybė

$$F_A>F_\alpha(I-1,\ I(J-1)). \tag{2.4.2}$$

Kriterijaus galios funkcija

$$\beta(\sigma_A^2)=\mathbf{P}\{\frac{MS_A}{MS_E}>F_\alpha(I-1,I)(J-1)|\sigma_A^2\}$$

$$=\mathbf{P}\{F_{I-1,I(J-1)}>\frac{\sigma^2+J\sigma_A^2}{\sigma^2}F_\alpha(I-1,I(J-1))\}$$

išreiškiama centrinio Fišerio skirstinio pasiskirstymo funkcija. Taigi abiejuose modeliuose hipotezei H_A tikrinti ir statistika F_A, ir kritinė sritis yra tos pačios, tik I modelyje kriterijaus galia išreiškiama necentriniu Fišerio skirstiniu, o II modelyje – centriniu Fišerio skirstiniu.

2.4.2.3 Parametrų vertinimas

Nepaslinktąjį dispersijos komponentės σ_A^2 įvertinį gauname remdamiesi 2.4.1 lentelės paskutiniuoju stulpeliu:

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_E}{J}, \quad \boldsymbol{V}(\hat{\sigma}_A^2) = \frac{2}{J^2}\left(\frac{\tau^4}{I-1} + \frac{\sigma^4}{I(J-1)}\right).$$

Apytikslį pasikliovimo intervalą gauname aproksimuodami normaliuoju arba Fišerio skirstiniu (žr. (2.5 skyrelį)). Vidurkio μ įvertiniu natūralu imti visų stebėjimų aritmetinį vidurkį

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} = \mu + \bar{Z}_{\cdot} + \bar{e}_{\cdot\cdot} \sim N(\mu, \frac{\tau^2}{IJ}).$$

Kadangi $\hat{\mu}$ ir SS_A yra nepriklausomi, tai išvadas apie vidurkį galima daryti remiantis sąryšiu

$$\sqrt{IJ}\frac{\hat{\mu} - \mu}{\sqrt{MS_A}} \sim S(I-1). \tag{2.4.3}$$

2.4.4 pavyzdys. Po 10 kartų buvo išmatuotos 9 atsitiktinai paimtų kineskopų uždarančiosios įtampos reikšmės. Gauti rezultatai surašyti į pateikiamą lentelę.

2.4.2 lentelė. Statistiniai duomenys

i j	1	2	3	4	5	6	7	8	9	10
1	53	54	56	54	55	53	56	55	59	54
2	48	48	47	50	51	48	47	48	50	47
3	62	65	62	66	62	61	63	62	64	61
4	56	55	54	57	55	56	55	53	57	50
5	68	67	70	68	68	67	67	68	70	68
6	55	54	59	54	56	54	55	58	54	53
7	53	54	53	54	53	53	52	53	54	52
8	52	50	50	55	53	51	51	49	55	49
9	66	65	67	65	66	65	66	68	66	68

Atlikę skaičiavimus, gauname tokią dispersinės analizės lentelę.

2.4.3 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	3729,6	8	466,2	$\sigma^2 + 10\sigma_A^2$
E	218,0	81	2,69	σ^2
T	3947,6			

Statistika F_A įgijo reikšmę 173,31. Hipotezė H_A atmetama aukšto reikšmingumo lygmens kriterijumi. Darome išvadą, kad kineskopai nėra identiški, o uždarančiosios įtampos sklaida σ_A^2 yra gana didelė, palyginti ją su matavimo paklaidos dispersija σ^2.

Taškiniai parametrų įverčiai yra $\hat{\sigma}^2 = MS_E = 2,69$; $\hat{\sigma}_A^2 = (MS_A - MS_E)/10 = 46,35$; $\hat{\mu} = \bar{Y}_{..} = 57,22$.

Sudarome parametrų pasikliovimo intervalus, kai pasikliovimo lygmuo $Q = 0,95$. Remdamiesi (2.4.3) ir 2.4.1 teorema gauname: $(\underline{\mu}; \overline{\mu}) = (51,97; 62,47)$; $(\underline{\sigma}^2; \overline{\sigma}^2) = (2,02; 3,76)$. Dispersijos komponentės σ_A^2 apytikslis pasikliovimo intervalas, gautas aproksimuojant Fišerio skirstiniu (žr. (2.5.9)), yra $(\underline{\sigma}_A^2; \overline{\sigma}_A^2) = (21,00; 170,83)$.

2.5. Dvifaktorė dispersinė analizė, kai faktoriai atsitiktiniai

2.5.1. Statistinis modelis

Atsitiktinio faktoriaus sąvoką nagrinėjome atlikdami vienfaktorę dispersinę analizę. Pateiksime keletą situacijų, kurioms duomenų analizei naudojama dvifaktorė analizė su dviem atsitiktiniais faktoriais.

2.5.1 pavyzdys. Psichologas sudarė 10 skirtingų testų. Atsakant į testo klausimus fiksuojamas teisingų atsakymų skaičius. Psichologas mano, kad teisingų atsakymų skaičius Y galėtų priklausyti nuo testų atlikimo eilės (faktorius A), administratoriaus (faktorius B), pateikiančio testus ir paaiškinančio jų pildymo tvarką, ir nuo faktorių A ir B nepriklausančių kitų nežinomų nestebimų atsitiktinių faktorių. Psichologas atsitiktinai parenka keturias testų pateikimo tvarkas ir keturis administratorius, kurių kiekvienas keturioms žmonių grupėms po penkis žmones kiekvienoje pateikia testus skirtinga tvarka (taigi grupių yra šešiolika). Abu faktoriai yra atsitiktiniai.

2.5.2 pavyzdys. Tarkime, kad televizorių konkretaus parametro reikšmės kiekvienu matavimo prietaisu matuojamos su galima sistemine paklaida ir visiems prietaisams vienoda atsitiktine paklaida. Taigi pamatuota parametro reikšmė Y gali priklausyti nuo televizoriaus ir matavimo prietaiso numerių (faktoriai A ir B) bei nuo atsitiktinės matavimo paklaidos. Atsitiktinai parinkime I gaminių, J matavimo prietaisų ir kiekvieną gaminį matuokime kiekvienu matavimo prietaisu po K kartų. Abu faktoriai A ir B yra atsitiktiniai.

Abiejų pavyzdžių atveju gauname modelį $Y = g(A, B) + e$; čia g yra reali atsitiktinių faktorių funkcija, e – nulinio vidurkio atsitiktinis dydis, nepriklausantis nuo $g(A, B)$. Atsitiktinis dydis $X = g(A, B)$ apibūdina faktorių A ir B įtaką tiriamam požymiui Y.

Tarsime, kad faktoriai A ir B yra nepriklausomi. Tai natūrali prielaida: pirmame pavyzdyje administratoriai ir testų pildymo tvarka, antrajame – televizoriai ir matavimo prietaisai parenkami nepriklausomai vienas nuo kito.

Pažymėkime

$$\mu = \mathbf{E}g(A,B), \quad g(A,.) = \mathbf{E}(g(A,B)|A), \quad g(.,B) = \mathbf{E}(g(A,B)|B).$$

Gauname

$$X = \mu + (g(A,.) - \mu) + (g(.,B) - \mu) + (g(A,B) - g(A,.) - g(.,B) + \mu) =$$

$$= \mu + a(A) + b(B) + c(A,B).$$

Akivaizdu, kad

$$\mathbf{E}(a(A)) = \mathbf{E}(b(B)) = \mathbf{E}(c(A,B)) = 0.$$

Dispersijas pažymėkime

$$\boldsymbol{V}(a(A)) = \sigma_A^2, \quad \boldsymbol{V}(b(B)) = \sigma_B^2, \quad \boldsymbol{V}(c(A,B)) = \sigma_{AB}^2, \quad \boldsymbol{V}(e) = \sigma^2.$$

Visi nariai $a(A)$, $b(B)$ ir $c(A,B)$ nekoreliuoti. Iš tikrųjų, kadangi faktoriai nepriklausomi, tai $a(A)$ ir $b(B)$ nekoreliuotumas akivaizdus. Pasinaudoję lygybę $\mathbf{E}a(A) = 0$ ir sąlyginio vidurkio savybėmis, gauname

$$\mathbf{Cov}\,(a(A), c(A,B)) = \mathbf{E}\{a(A)[g(A,B) - g(A,.) - g(.,B) + \mu]\} =$$

$$\mathbf{E}\{a(A)\mathbf{E}[g(A,B) - g(A,.)|A]\} = 0.$$

Atsitiktiniai dydžiai $a(A)$ ir $b(B)$ dažnai vadinami *pagrindiniais faktoriais*. Atsitiktinis dydis $a(A)$ yra sąlyginio vidurkio $\mathbf{E}(Y|A) = \mathbf{E}(X|A)$ nuokrypis nuo vidurkio $\mathbf{E}(Y) = \mathbf{E}(X) = \mu$. Šiuo požiūriu jis apibūdina faktoriaus A įtaką tiriamam požymiui. Jei teisinga hipotezė $H_A : \sigma_A^2 = 0$, tai $\mathbf{E}(Y|A) = \mu$, t. y. sąlyginiai vidurkiai $\mathbf{E}(Y|A)$ nepriklauso nuo A. Ši hipotezė yra hipotezės $H_A : \mu_1 = \cdots = \mu_I$, nagrinėtos dispersinėje analizėje su fiksuotais faktoriais, analogas. Analogiškai interpretuojamas a. d. $b(B)$ ir hipotezė $H_B : \sigma_B^2 = 0$.

Jei faktoriaus A reikšmė fiksuota, tai faktoriaus B įtaką požymio Y skirstiniui apibūdinančio a. d. $X = \mu + a(A) + b(B) + c(A,B)$ reikšmė priklauso ne tiktai nuo B, bet ir nuo tos fiksuotos faktoriaus A reikšmės, jei $c(A,B) \neq 0$. Taigi narys $c(A,B)$ apibūdina *faktorių sąveiką*. Jei teisinga hipotezė $H_{AB} : \sigma_{AB}^2 = 0$, tai faktorių sąveikos nėra. Ši hipotezė yra hipotezės $H_{AB} : \gamma_{ij} \equiv 0$, nagrinėtos dispersinėje analizėje su fiksuotais faktoriais, analogas.

Atsitiktinį I faktoriaus A ir J faktoriaus B lygmenų parinkimą galime interpretuoti kaip didumo I ir J paprastųjų imčių $A_1, ..., A_I$ ir $B_1, ..., B_J$ parinkimą. Taigi nagrinėsime toliau apibrėžtą modelį.

Dvifaktorės dispersinės analizės su atsitiktiniais faktoriais modelis:

$$Y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}, \tag{2.5.1}$$

$$i = 1, ..., I, \quad j = 1, ..., J, \quad k = 1, ..., K;$$

čia visi dėmenys

$$a_i = a(A_i), \quad b_j = b(B_j), \quad c_{ij} = c(A_i,\ B_j), \quad e_{ijk}$$

yra nulinio vidurkio nekoreliuoti atsitiktiniai dydžiai.

Jeigu priimsime papildomą prielaidą, kad a. d. $a_i, b_j, c_{ij}, e_{ijk}$ yra normalieji:

$$a_i \sim N(0,\sigma_A^2), \quad b_j \sim N(0,\sigma_B^2), \quad c_{ij} \sim N(0,\sigma_{AB}^2), \quad e_{ijk} \sim N(0,\sigma^2),$$

tai visi nariai yra nepriklausomi ir modelį visiškai nusako vidurkis μ ir keturios dispersijos σ_A^2, σ_B^2, σ_{AB}^2, σ^2.

Stebėjimų Y_{ijk} kovariacijos yra tokio pavidalo:

$$\mathbf{Cov}\,(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} \sigma_A^2+\sigma_B^2+\sigma_{AB}^2+\sigma^2, & \text{kai} \quad i=i', \quad j=j', \quad k=k', \\ \sigma_A^2+\sigma_B^2+\sigma_{AB}^2, & \text{kai} \quad i=i', \quad j=j', \quad k\neq k', \\ \sigma_A^2, & \text{kai} \quad i=i', \quad j\neq j', \\ \sigma_B^2, & \text{kai} \quad i\neq i', \quad j=j', \\ 0, & \text{kai} \quad i\neq i', \quad j\neq j'. \end{cases}$$

2.5.2. Kvadratų sumų skirstiniai

Kvadratų sumas SS_A, SS_B ir SS_{AB} apibrėžkime kaip ir modelyje, kuriame faktoriai pastovūs (žr. (2.2.8) formules). Vietoje Y_{ijk} įrašykime (2.5.1) išraiškas. Gauname

$$SS_A = JK\sum_i (\bar{Y}_{i..}-\bar{Y}_{...})^2 = JK\sum_i (a_i-\bar{a}_.+\bar{c}_{i.}-\bar{c}_{..}+\bar{e}_{i..}-\bar{e}_{...})^2,$$

$$SS_B = IK\sum_j (\bar{Y}_{.j.}-\bar{Y}_{...})^2 = IK\sum_j (b_j-\bar{b}_.+\bar{c}_{.j}-\bar{c}_{..}+\bar{e}_{.j.}-\bar{e}_{...})^2,$$

$$SS_{AB} = K\sum_i\sum_j (\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...})^2 =$$

$$= K\sum_i\sum_j (c_{ij}-\bar{c}_{i.}-\bar{c}_{.j}+\bar{c}_{..}+\bar{e}_{ij.}-\bar{e}_{i..}-\bar{e}_{.j.}+\bar{e}_{...})^2,$$

$$SS_E = \sum_i\sum_j\sum_k (Y_{ijk}-\bar{Y}_{ij.})^2 = \sum_i\sum_j\sum_k (e_{ijk}-\bar{e}_{ij.})^2.$$

2.5.1 teorema. *Kvadratų sumos $SS_A, SS_B, SS_{AB}, SS_E$ yra nepriklausomos ir turi tokius skirstinius*

$$SS_A \sim (\sigma^2+K\sigma_{AB}^2+JK\sigma_A^2)\chi^2_{I-1}, \quad SS_B \sim (\sigma^2+K\sigma_{AB}^2+IK\sigma_B^2)\chi^2_{J-1},$$

$$SS_{AB} \sim (\sigma^2+K\sigma_{AB}^2)\chi^2_{(I-1)(J-1)}, \quad SS_E \sim \sigma^2\chi^2_{IJ(K-1)}. \tag{2.5.2}$$

Įrodymas. Visos keturios kvadratų sumos sukomponuotos iš tokių a. d. sistemų: $\{a_i-\bar{a}\}$; $\{b_j-\bar{b}\}$; $\{\bar{c}_{i.}-\bar{c}_{..}\}$, $\{\bar{c}_{.j}-\bar{c}_{..}\}$, $\{c_{ij}-\bar{c}_{i.}-\bar{c}_{.j}+\bar{c}_{..}\}$; $\{\bar{e}_{i..}-\bar{e}_{...}\}$, $\{\bar{e}_{.j.}-\bar{e}_{...}\}$, $\{\bar{e}_{ij.}-\bar{e}_{i..}-\bar{e}_{.j.}+\bar{e}_{...}\}$, $\{e_{ijk}-\bar{e}_{ij.}\}$. Kadangi a. d. žymimi skirtingomis raidėmis yra nepriklausomi, tai tereikia įsitikinti, kad a. d. sistemos, sudarytos iš vienodomis raidėmis žymimų a. d., yra tarpusavyje nepriklausomos.

Kadangi $e_{ijk} \sim N(0, \sigma^2)$ yra vienodai pasiskirstę n. a. d., tai remiantis 2.2.2 pastaba galima tvirtinti, kad a. d. sistemos $\{\bar{e}_{i..} - \bar{e}_{...}\}$, $\{\bar{e}_{.j.} - \bar{e}_{...}\}$, $\{\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\}$, $\{e_{ijk} - \bar{e}_{ij.}\}$ yra nepriklausomos.

A. d. $c_{ij} \sim N(0, \sigma^2_{AB})$ yra vienodai pasiskirstę n. a. d., todėl vėl, remiantis 2.2.2 pastaba (imant $K = 1$), galima tvirtinti, kad a. d. sistemos $\{\bar{c}_{i.} - \bar{c}_{..}\}$, $\{\bar{c}_{.j} - \bar{c}_{..}\}$, $\{c_{ij.} - \bar{c}_{i.} - \bar{c}_{.j} + \bar{c}_{..}\}$ yra nepriklausomos.

Taigi kvadratų sumos $SS_A, SS_B, SS_{AB}, SS_E$ yra nepriklausomos.

Kvadratų suma apibūdinanti atsitiktines paklaidas SS_E tokiu pat būdu kaip ir I modelyje, priklauso tik nuo a. d. e_{ijk}, todėl, kaip ir 2.2.1 teoremoje, $SS_E \sim \sigma^2 \chi^2_{IJ(K-1)}$.

Pažymėję vienodai pasiskirsčiusius n. a. d. $Z_i = a_i + \bar{c}_{i.} + \bar{e}_{i..} \sim N(0,\ \sigma^2_A + \sigma^2_{AB}/J + \sigma^2/JK)$, gauname

$$SS_A = JK \sum_i (Z_i - \bar{Z}_.)^2 \sim JK(\sigma^2_A + \frac{\sigma^2_{AB}}{J} + \frac{\sigma^2}{JK})\chi^2_{I-1}.$$

Analogiškai gauname $SS_B \sim (\sigma^2 + K\sigma^2_{AB} + IK\sigma^2_B)\chi^2_{J-1}$.

Ieškodami SS_{AB} skirstinio prisiminkime dvifaktorės dispersinės analizės I modelį. 2.2.1 teoremoje įrodyta, kad iš lygybės

$$\bar{Y}_{ij.} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \bar{\varepsilon}_{ij.}, \quad \bar{\varepsilon}_{ij.} \sim N(0, \sigma^2/K),$$

išplaukia, kad $K \sum_i \sum_j (Y_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \sim \sigma^2 \chi^2_{(I-1)(J-1);\lambda_{AB}}$, $\lambda_{AB} = \frac{K}{\sigma^2} \sum_i \sum_j \gamma_{ij}$.

Kai faktoriai atsitiktiniai, pažymėję $U_{ij} = c_{ij} + \bar{e}_{ij.}$, gausime modelį

$$U_{ij} = 0 + U_{ij}, \quad U_{ij} \sim N(0,\ \sigma^2_{AB} + \sigma^2/K).$$

Šiame modelyje vietoje $\bar{Y}_{ij.}$ ir $\bar{\varepsilon}_{ij.}$ yra U_{ij}, $\mu = \alpha_i = \beta_j = \gamma_{ij} = 0$, o vietoje σ^2 figūruoja $K\sigma^2_{AB} + \sigma^2$. Taigi $SS_{AB} = K \sum_i \sum_j (U_{ij} - \bar{U}_{i.} - \bar{U}_{.j} + \bar{U}_{..})^2 \sim (\sigma^2 + K\sigma^2_{AB})\chi^2_{(I-1)(J-1)})$, nes $\lambda_{AB} = 0$.

▲

2.5.3. Parametrų įvertiniai ir hipotezių tikrinimas

2.5.3.1. Dispersinės analizės lentelė

Pažymėkime

$$MS_A = \frac{SS_A}{I-1}, \quad MS_B = \frac{SS_B}{J-1},$$

$$MS_{AB} = \frac{SS_{AB}}{(I-1)(J-1)}, \quad MS_E = \frac{SS_E}{IJ(K-1)}.$$

Tada šių vidutinių kvadratų sumų vidurkiai yra lygūs daugikliams, parašytiems (2.5.2) formulėse prieš atitinkamus χ^2 atsitiktinius dydžius.

Skaičiavimo rezultatus surašome į dispersinės analizės lentelę

2.5.1 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	$MS = SS/\nu$	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + K\sigma^2_{AB} + JK\sigma^2_A$
B	SS_B	$J-1$	MS_B	$\sigma^2 + K\sigma^2_{AB} + IK\sigma^2_B$
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}	$\sigma^2 + K\sigma^2_{AB}$
E	SS_E	I J(K-1)	MS_E	σ^2
T	SS_T	I J K -1	-	-

Matome, kad ši lentelė skiriasi nuo 2.2.1 lentelės tik paskutiniuoju stulpeliu. Kadangi kiekviena kvadratų suma tenkina sąlygą

$$\frac{SS}{\mathbf{E}(MS)} \sim \chi^2(\nu), \tag{2.5.3}$$

tai iš čia standartiniu būdu galime gauti bet kurio parametro, pateikto 2.5.1 lentelės paskutiniajame stulpelyje, pasikliovimo intervalą, arba sudaryti kriterijus hipotezėms apie šių parametrų reikšmes tikrinti. Pasirinkę bet kurių dviejų skirtingų trupmenų $MS/\mathbf{E}(MS)$ santykius, gausime Fišerio skirstinius su atitinkamais laisvės laipsnių skaičiais, įrašytais 2.5.1 lentelės 3 stulpelyje. Taigi galime daryti korektiškas išvadas apie bet kuriuos parametrus, kurie gaunami imant bet kurių dviejų skirtingų 2.5.1 lentelės paskutiniojo stulpelio reiškinių santykius.

2.5.3.2. Faktorių įtakos ir sąveikos nebuvimo hipotezių tikrinimas

Hipotezėms $H_A : \sigma^2_A = 0$, $H_B : \sigma^2_B = 0$ ir $H_{AB} : \sigma^2_{AB} = 0$ tikrinti pasirenkamos statistikos

$$F_A = \frac{MS_A}{MS_{AB}}, \quad F_B = \frac{MS_B}{MS_{AB}}, \quad F_{AB} = \frac{MS_{AB}}{MS_E}. \tag{2.5.4}$$

Kai atitinkama hipotezė teisinga, tai iš (2.5.2) gauname, kad šie santykiai turi centrinius Fišerio skirstinius su atitinkamais laisvės laipsnių skaičiais. Hipotezės atmetamos reikšmingumo lygmens α kriterijais, kai atitinkamai yra tenkinamos nelygybės

$$F_A > F_\alpha(I-1,\ (I-1)(J-1)), \quad F_B > F_\alpha(J-1,\ (I-1)(J-1)),$$
$$F_{AB} > F_\alpha((I-1)(J-1),\ IJ(K-1)). \tag{2.5.5}$$

Jeigu suformuluotos hipotezės neteisingos, tai santykiai (2.5.4) skiriasi nuo centrinių Fišerio a. d. tik daugikliu. Todėl šiame modelyje kriterijų galia išreiškiama centrinių Fišerio skirstinių pasiskirstymo funkcijomis. Pavyzdžiui, tikrinant hipotezę H_A, kriterijaus galios funkcija

$$\beta(\sigma^2_A) = \mathbf{P}\{\frac{MS_A}{MS_{AB}} > F_\alpha(I-1,(I-1)(J-1))|\sigma^2_A\}$$

$$= \mathbf{P}\{F_{I-1,(I-1)(J-1)} > \frac{\sigma^2 + K\sigma^2_{AB}}{\sigma^2 + K\sigma^2_{AB} + JK\sigma^2_A} F_\alpha(I-1,(I-1)(J-1))\}.$$

Galios funkcija artėja prie 1, kai santykis $JK\sigma^2_A/(\sigma^2 + K\sigma^2_{AB})$ artėja į ∞.

2.5.3.3. Parametrų vertinimas

Parametro μ nepaslinktasis įvertinys yra

$$\hat{\mu} = \bar{Y}_{\cdots} \sim N\left(\mu,\ \frac{\sigma_A^2}{I} + \frac{\sigma_B^2}{J} + \frac{\sigma_{AB}^2}{IJ} + \frac{\sigma^2}{IJK}\right), \tag{2.5.6}$$

o dispersijos komponenčių nepaslinktuosius įvertinius gauname remdamiesi 2.5.1 lentelės paskutiniuoju stulpeliu:

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_{AB}}{JK}, \quad \hat{\sigma}_B^2 = \frac{MS_B - MS_{AB}}{IK},$$
$$\hat{\sigma}_{AB}^2 = \frac{MS_{AB} - MS_E}{K}, \quad \hat{\sigma}^2 = MS_E. \tag{2.5.7}$$

Pasirėmę tuo, kad kvadratinės formos yra nepriklausomos, iš (2.5.2) gauname

$$\boldsymbol{V}(\hat{\sigma}_A^2) = \frac{2}{J^2K^2}\left(\frac{(\mathbf{E}(MS_A))^2}{I-1} + \frac{(\mathbf{E}(MS_{AB}))^2}{(I-1)(J-1)}\right),$$
$$\boldsymbol{V}(\hat{\sigma}_B^2) = \frac{2}{I^2K^2}\left(\frac{(\mathbf{E}(MS_B))^2}{J-1} + \frac{(\mathbf{E}(MS_{AB}))^2}{(I-1)(J-1)}\right),$$
$$\boldsymbol{V}(\hat{\sigma}_{AB}^2) = \frac{2}{K^2}\left(\frac{(\mathbf{E}(MS_{AB}))^2}{(I-1)(J-1)} + \frac{(\mathbf{E}(MS_E))^2}{IJ(K-1)}\right). \tag{2.5.8}$$

Pažymėsime, kad $\boldsymbol{V}(\hat{\sigma}_A^2)$ arba $\boldsymbol{V}(\hat{\sigma}_B^2)$ artėja prie 0 tik tada, kai atitinkamai $I \to \infty$ arba $J \to \infty$. Išraiškomis (2.5.8) galima naudotis planuojant eksperimentą: taip parinkti I, J, K, kad dominančios dispersijos komponentės būtų įvertinamos reikiamu tikslumu.

Naudodamiesi (2.5.6), (2.5.8) ir taikydami normaliąją aproksimaciją galime sudaryti apytikslius parametrų μ, σ_A^2, σ_B^2, σ_{AB}^2 pasikliovimo intervalus. Tikslesnius dispersijos komponenčių pasikliovimo intervalus gauname aproksimuodami Fišerio skirstiniu.

Tarkime, SS_1 ir SS_2 yra tokios nepriklausomos kvadratų sumos, kad SS_1 pasiskirsčiusi kaip $(\varphi+\psi)\chi^2_{\nu_1}$, o SS_2 – kaip $\psi\chi^2_{\nu_2}$. Apytikslis parametro φ pasikliovimo intervalas $(\underline{\varphi},\ \bar{\varphi})$, kai pasikliovimo lygmuo $Q = 1-2\alpha$, aproksimuojant Fišerio skirstiniu yra tokio pavidalo (žr. [ŠEF]):

$$\underline{\varphi} = MS_2\left(\frac{F}{F_\alpha(\nu_1,\ \infty)} - 1 - \frac{F_\alpha(\nu_1,\ \nu_2)}{F}\left(\frac{F_\alpha(\nu_1,\ \nu_2)}{F_\alpha(\nu_1,\ \infty)} - 1\right)\right),$$
$$\bar{\varphi} = MS_2\left(FF_\alpha(\infty,\ \nu_1) - 1 - \frac{1}{FF_\alpha(\nu_2,\ \nu_1)}\left(1 - \frac{F_\alpha(\infty,\ \nu_1)}{F_\alpha(\nu_2,\ \nu_1)}\right)\right); \tag{2.5.9}$$

čia $F = MS_1/MS_2$, o $F_\alpha(\nu_1,\infty) = \chi^2_\alpha(\nu_1)/\nu_1$, $F_\alpha(\infty,\nu_1) = \nu_1/\chi^2_{1-\alpha}(\nu_1)$.

Pateiktuoju intervalu galime naudotis ieškodami parametrų σ_A^2, σ_B^2, σ_{AB}^2 apytikslių pasikliovimo intervalų, taip pat intervaliniu būdu vertindami dispersijos komponentes kitose schemose. Pavyzdžiui, vertinant komponentę σ_A^2

(2.5.9) formulėse reikia įrašyti $\varphi = JK\sigma_A^2$, $\psi = \sigma^2 + K\sigma_{AB}^2$, $\nu_1 = I - 1$, $\nu_2 = (I-1)(J-1)$, $SS_1 = SS_A$, $SS_2 = SS_{AB}$.

Jeigu turime po vieną stebėjimą kiekvienam faktorių lygmenų rinkiniui, t. y. $K = 1$, tai sumą SS_E atitinkantis laisvės laipsnių skaičius lygus 0. Todėl negalime įvertinti dispersijos komponentės σ^2 ir patikrinti hipotezės H_{AB}. Visos kitos išvados lieka teisingos su ta išimtimi, kad negalima atskirai įvertinti σ^2 ir σ_{AB}^2, galima daryti išvadas tik apie sumą $\sigma^2 + \sigma_{AB}^2$.

2.5.3 pavyzdys Tarkime, kad 2.4.4 pavyzdyje matavimai gauti tokiu būdu. Atsitiktinai atrinktų 9 kineskopų (faktorius A) uždarančiosios įtampos buvo matuojamos po $K = 2$ kartus atsitiktinai atrinktais 5 matavimo prietaisais (faktorius B). Matavimo rezultatai pateikti 2.4.2 lentelėje. Pirmuose dviejuose stulpeliuose yra matavimo rezultatai, gauti matuojant pirmuoju prietaisu; kituose dviejuose stulpeliuose – antruoju prietaisu ir pagaliau paskutiniuose dviejuose stulpeliuose – penktuoju prietaisu. Tarsime, kad abu faktoriai yra atsitiktiniai.

Atlikę skaičiavimus, gauname tokią dispersinės analizės lentelę.

2.5.2 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}$(MS)
A	3729,6	8	466,2	$\sigma^2 + 2\sigma_{AB}^2 + 10\sigma_A^2$
B	9,56	4	2,39	$\sigma^2 + 2\sigma_{AB}^2 + 18\sigma_B^2$
$A \times B$	54,44	32	1,70	$\sigma^2 + 2\sigma_{AB}^2$
E	154	45	3,42	σ^2
T	3947,6			

Statistikos F_{AB} ir F_B įgijo atitinkamai reikšmes 0,50 ir 1,4. Atmesti hipotezes H_{AB} ir H_B nėra pagrindo. Tuo labiau kad taškiniai parametrų σ_{AB}^2 ir σ_B^2 įverčiai, apskaičiuoti pagal (2.5.7) formules, įgijo neigiamas reikšmes. Tai gali nutikti dažnai, jeigu kuri nors dispersijos komponentė lygi 0 arba maža, palyginti su kitomis komponentėmis. Tokiu atveju rekomenduojama dispersinės analizės lentelėje pakeisti tokias komponentes 0 ir sujungti tas dispersinės analizės eilutes, kurių po tokio pakeitimo $\mathbf{E}(MS)$ sutampa. Atlikę tokį keitimą 2.5.2 lentelėje, gausime vienfaktorės dispersinės analizės 2.4.3 lentelę.

Pažymėsime, kad sąlygiškai suskirstėme vienfaktorės dispersinės analizės 2.4.2 lentelės duomenis į penkias grupes, todėl MS_E, MS_B ir MS_{AB} faktiškai yra tos pačios paklaidos dispersijos įverčiai. Tą ir patvirtina gauti rezultatai.

2.6. Mišrusis dvifaktorės dispersinės analizės modelis

2.6.1. Statistinis modelis

Dvifaktorė dispersinė analizė vadinama *mišriąja*, jei vienas faktorius yra pastovus, o kitas – atsitiktinis.

Dvifaktorės mišriosios dispersinės analizės modelio forma nesiskiria nuo dvifaktorės dispersines analizės modelių su fiksuotais (modelis I) ir atsitiktiniais (modelis II) faktoriais:

$$Y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + e_{ijk}. \tag{2.6.1}$$

Pateiksime pavyzdžių.

2.6.1 pavyzdys. Tiriant protinio aktyvumo (faktorius A) įtaką kraujui cirkuliuoti smegenyse (požymis Y), eksperimento dalyviai atliko matematikos ($A = A_1$), perskaityto teksto supratimo ($A = A_2$) ir istorijos ($A = A_3$) testus. Eksperimentatorius manė, kad įtaką požymio Y skirstiniui galėtų turėti ir mokinių klasės (faktorius B) parinkimas. Todėl atsitiktinai buvo parinktos šešios ($J = 6$) penktokų klasės. Kiekvienos klasės mokiniai atliko minėtus tris ($I = 3$) testus. Po testų buvo matuojamas kraujo cirkuliavimo smegenyse lygis. Faktorius A yra fiksuotas, faktorius B – atsitiktinis.

2.6.2 pavyzdys. Sakykime, gaminys apibūdinamas atsitiktiniu parametrų vektoriumi $\mathbf{X} = (X_1, ..., X_I)^T$. Jeigu faktoriaus A lygmenimis laikysime parametrų numerius, tai šio faktoriaus reikšmės fiksuotos. Atsitiktinai atrenkame J gaminių (atsitiktinis faktorius B) ir su galima paklaida matuojame kiekvieną jų parametrą po K kartų.

Nagrinėdami abu pavyzdžius gauname, kad, fiksavus faktoriaus A reikšmę A_i, gaunamas modelis $Y_i = X_i + e_i$; čia $X_i = g_i(B)$ yra reali atsitiktinio faktoriaus B funkcija, e_i – nulinio vidurkio atsitiktinis dydis, nepriklausantis nuo X_i. Atsitiktinis dydis $X_i = g_i(B)$ apibūdina faktoriaus B įtaką tiriamam požymiui Y, kai fiksuota faktoriaus A reikšmė A_i.

Pažymėkime

$$\mu_i = \mathbf{E}Y_i = \mathbf{E}g_i(B), \quad \bar{g}(B) = \frac{1}{I}\sum_{i=1}^{I} X_i = \frac{1}{I}\sum_{i=1}^{I} g_i(B), \quad \mu = \mathbf{E}\bar{g}(B) = \frac{1}{I}\sum_{i=1}^{I}\mu_i.$$

Gauname

$$X_i = b_i(B) = \mu + (\mu_i - \mu) + (\bar{g}(B) - \mu) + (g_i(B) - \mu_i - \bar{g}(B) + \mu)$$

$$= \mu + \alpha_i + b(B) + c_i(B).$$

Akivaizdu, kad

$$\mathbf{E}(b(B)) = \mathbf{E}(c_i(B)) = 0, \quad i = 1, ..., I; \quad \sum_i \alpha_i = \sum_i c_i = 0.$$

Atsitiktinių dydžių $b(B)$, $c_1(B), ..., c_I(B)$ kovariacijos išreiškiamos a. v. $\mathbf{X} = (X_1, \ldots, X_I)^T$, $X_i = g_i(B)$, kovariacinės matricos $\mathbf{\Sigma} = [\sigma_{ii'}]_{I\times I}$ elementais $\sigma_{ii'} = \mathbf{Cov}\,(X_i, X_{i'})$:

$$\sigma_B^2 = \boldsymbol{V}(b(B)) = \bar{\sigma}_{..}, \quad \sigma_{AB;i}^2 = \boldsymbol{V}(c_i(B)) = \sigma_{ii} - 2\bar{\sigma}_{i.} + \bar{\sigma}_{..},$$

$$\mathbf{Cov}\,(b(B),\ c_i(B)) = \bar{\sigma}_{i.} - \bar{\sigma}_{..}, \quad \mathbf{Cov}\,(c_i(B),\ c_{i'}(B)) = \sigma_{ii'} - \bar{\sigma}_{i.} - \bar{\sigma}_{.i'} + \bar{\sigma}_{..}. \tag{2.6.2}$$

Atsitiktinį J faktoriaus B lygmenų parinkimą galima interpretuoti kaip didumo J paprastosios imties $B_1, ..., B_J$ parinkimą. Taigi nagrinėsime toliau apibrėžtą modelį.

Mišrusis dvifaktorės dispersinės analizės modelis:

$$Y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + e_{ijk}, \quad b_j = b(B_j), \quad c_{ij} = c_i(B_j). \tag{2.6.3}$$

$$i = 1, ..., I, \quad j = 1, ..., J, \quad k = 1, ..., K;$$

čia a. v. $(b_j,\ c_{1j}, ..., c_{Ij})^T$ skirstinys yra daugiamatis normalusis, turintis nulinį vidurkių vektorių ir kovariacijas (2.6.2) (su skirtingais j šie vektoriai nepriklausomi), o atsitiktinės matavimo paklaidos $e_{ijk} \sim N(0, \sigma^2)$ nepriklausomos tarpusavyje ir nepriklauso nuo šių vektorių.

Parametrai α_i ir atsitiktiniai dydžiai c_{ij} tenkina sąlygas

$$\sum_i \alpha_i = \sum_i c_{ij} = 0, \quad j = 1, \ldots, J.$$

Kadangi $\mu_i = \mathbf{E}(Y|A_i)$, $\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i$, tai kaip ir I modelio parametrai $\alpha_i = \mu_i - \mu$ apibūdina faktoriaus A „tiesioginę“ įtaką, eliminavus faktoriaus B įtaką (μ_i gauti suvidurkinus $g_i(B)$). Jei teisinga hipotezė $H_A : \mu_1 = \cdots = \mu_I$, tai "tiesioginės" faktoriaus A įtakos nėra.

Atsitiktiniai dydžiai $b_j = \bar{g}(B_j) - \mu$ apibūdina faktoriaus B „tiesioginę“ įtaką, eliminavus faktoriaus A įtaką ($\bar{g}(B_j)$ gauti, imant empirinį $g_1(B_j), \ldots, g_I(B_j)$ vidurkį). Jei teisinga hipotezė $H_B : \sigma_B^2 = 0$, tai „tiesioginės“ faktoriaus B įtakos nėra.

Fiksavus faktoriaus A reikšmę A_i, kartu α_i, a. d. $X_i = g_i(B_j) = \mu + \alpha_i + b(B_j) + c_i(B_j)$ priklauso ne tiktai nuo B_j, bet ir nuo i, jeigu $c_i(B_j) \neq 0$, ir, atvirkščiai, fiksavus atsitiktinio faktoriaus B_j reikšmę, kartu fiksavus $b(B_j)$, a. d. X_i priklauso ne tiktai nuo A_i, bet ir nuo B_j reikšmės, jei $c_i(B_j) \neq 0$. Taigi narys $c_i(B_j)$ apibūdina faktorių sąveiką. Pažymėkime

$$\sigma_{AB}^2 = \frac{1}{I-1} \sum_i \sigma_{AB;i}^2.$$

Jei teisinga hipotezė $H_{AB} : \sigma_{AB}^2 = 0$, tai faktorių sąveikos nėra.

2.6.2. Kvadratų sumų skirstiniai

Sudarykime įprastines kvadratų sumas (2.2.8) ir įrašykime jose vietoje a. d. Y_{ijk} (2.6.2) išraiškas. Gauname

$$SS_A = JK \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 = JK \sum_i (\alpha_i + \bar{c}_{i.} + \bar{e}_{i..} - \bar{e}_{...})^2,$$

$$SS_B = IK \sum_j (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = IK \sum_j (b_j - \bar{b}_{.} + \bar{e}_{.j.} - \bar{e}_{...})^2,$$

$$SS_{AB} = K \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 =$$

$$= K \sum_i \sum_j (c_{ij} - \bar{c}_{i.} + \bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...})^2,$$

$$SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 = \sum_i \sum_j \sum_k (e_{ijk} - \bar{e}_{ij.})^2.$$

2.6.1 teorema. *1. Kvadratų sumos $SS_A, SS_B, SS_{AB}, SS_E$ yra nepriklausomos, išskyrus porą SS_B ir SS_{AB}. Kvadratų sumos SS_E ir SS_B turi tokius skirstinius*

$$SS_E \sim \sigma^2 \chi^2_{IJ(K-1)}, \quad SS_B \sim (\sigma^2 + IK\sigma_B^2)\chi^2_{J-1}. \tag{2.6.4}$$

2. Jeigu modelis adityvusis, t. y. $c_{ij} \equiv 0$ su tikimybe 1, tai nepriklausomos visos keturios kvadratų sumos. Šiuo atveju kvadratų sumos SS_A ir SS_{AB} turi tokius skirstinius:

$$SS_{AB} \sim \sigma^2 \chi^2_{(I-1)(J-1)}, \quad SS_A \sim \sigma^2 \chi^2_{I-1;\lambda_A}, \quad \lambda_A = \frac{JK}{\sigma^2}\sum_i \alpha_i^2. \tag{2.6.5}$$

Įrodymas. 1. Kvadratų sumos sudarytos iš tokių a. d. sistemų: $\{b_j - \bar{b}\}$; $\{\bar{c}_{i.}\}$; $\{c_{ij} - \bar{c}_{i.}\}$; $\{\bar{e}_{i..} - \bar{e}_{...}\}$, $\{\bar{e}_{.j.} - \bar{e}_{...}\}$, $\{\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\}$, $\{e_{ijk} - \bar{e}_{ij.}\}$. Skyrelyje (2.4.2) buvo įrodyta, kad paskutinės keturios sistemos yra tarpusavyje nepriklausomos. Kadangi visi dėmenys (2.6.2) išdėstyme nepriklausomi, tai šios sistemos nepriklauso ir nuo pirmųjų trijų. Patikrinsime pirmųjų trijų sistemų nepriklausomumą. Apskaičiuojame kovariacijas:

$$\mathbf{Cov}\,(\bar{c}_i,\ b_j - \bar{b}_{.}) = \frac{1}{J}\mathbf{Cov}\,(b(B),\ c_i(B)) - \frac{J}{J^2}\mathbf{Cov}\,(b(B),\ c_i(B)) = 0;$$

$$\mathbf{Cov}\,(\bar{c}_i,\ c_{i'j} - \bar{c}_{i'.}) = \frac{1}{J}\mathbf{Cov}\,(c_i(B),\ c_{i'}(B)) - \frac{J}{J^2}\mathbf{Cov}\,(c_i(B),\ c_{i'}(B)) = 0;$$

Likusioji kovariacija

$$\mathbf{Cov}\,(c_{ij'} - \bar{c}_{i.},\ b_j - \bar{b}_{.}) = \mathbf{Cov}\,(c_{ij'},\ b_j - \bar{b}_{.}) = \mathbf{Cov}\,(b(B_j),\ c_i(B_{j'})) -$$

$$\frac{1}{J}\mathbf{Cov}\,(b(B),\ c_i(B)) = \begin{cases} \frac{J-1}{J}\mathbf{Cov}\,(b(B),\ c_i(B)), & \text{kai} \quad j = j', \\ -\frac{1}{J}\mathbf{Cov}\,(b(B),\ c_i(B)), & \text{kai} \quad j \neq j'. \end{cases}$$

Vadinasi, priklausomos tik kvadratų sumos SS_B ir SS_{AB}, nes į jas įeina atsitiktiniai dydžiai, paimti iš sistemų $\{b_j - \bar{b}\}$ ir $\{c_{ij} - \bar{c}_{i.}\}$.

Kvadratų suma SS_E priklauso tik nuo paklaidų e_{ijk} tokiu pat būdu kaip ir 2.2.4 skyrelyje. Todėl

$$SS_E \sim \sigma^2 \chi^2_{IJ(K-1)}.$$

Atsitiktiniai dydžiai $Z_j = b_j + \bar{e}_{.j.}$ yra nepriklausomi ir vienodai pasiskirstę $Z_j \sim N(0, \sigma_B^2 + \sigma^2/(IK))$. Todėl pagal I d. 2.5.1 teoremą

$$SS_B = IK\sum_j (Z_j - \bar{Z}_{.})^2 \sim (IK\sigma_B^2 + \sigma^2)\,\chi^2_{J-1}.$$

2. Jeigu $c_{ij} \equiv 0$, tai $c_{ij} - \bar{c}_{i.} \equiv 0$, todėl ir sumos SS_B bei SS_{AB} yra nepriklausomos.

Nepriklausomi atsitiktiniai dydžiai $U_i = \alpha_i + \bar{e}_{i..} \sim N(\alpha_i, \sigma^2/(JK))$. Todėl

$$SS_A = JK\sum_i (U_i - \bar{U}_{.})^2 \sim \sigma^2\,\chi^2_{I-1;\lambda_A}.$$

Necentriškumo parametras λ_A gaunamas vietoje a. d. U_i įrašant jų vidurkius $\mathbf{E}(U_i) = \alpha_i$. ▲

Kaip ir pirmiau pažymėkime

$$MS_A = \frac{SS_A}{I-1}, \quad MS_B = \frac{SS_B}{J-1}, \quad MS_{AB} = \frac{SS_{AB}}{(I-1)(J-1)}, \quad MS_E = \frac{SS_E}{IJ(K-1)}.$$

Remdamiesi (2.6.3) gauname

$$\mathbf{E}(MS_B) = \sigma^2 + IK\sigma_B^2, \quad \mathbf{E}(MS_E) = \sigma^2.$$

Randame

$$\mathbf{E}(MS_{AB}) = \frac{K}{(I-1)(J-1)}\{\mathbf{E}(\sum_i \sum_j (c_{ij} - \bar{c}_{i.})^2) + \mathbf{E}(\sum_i \sum_j (\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...})^2)\}.$$

Kadangi $\bar{e}_{ij.} - \bar{e}_{.j..} \sim N(0, \frac{I-1}{IK}\sigma^2)$, tai pagal I d. 2.5.1 teoremą

$$\sum_j ((\bar{e}_{ij.} - \bar{e}_{.j..}) - (\bar{e}_{i..} - \bar{e}_{...}))^2 \sim \frac{I-1}{IK}\sigma^2 \chi^2_{J-1}.$$

Taigi antrojo dėmens $\mathbf{E}(MS_{AB})$ išraiškoje vidurkis yra $(I-1)(J-1)\sigma^2/K$. Kadangi

$$\mathbf{E}(c_{ij} - \bar{c}_{i.})^2 = \boldsymbol{V}(c_{ij}) - \frac{1}{J}\boldsymbol{V}(c_{ij}) = \frac{J-1}{J}\sigma^2_{AB;i},$$

tai

$$\mathbf{E}(MS_{AB}) = \frac{K}{(I-1)(J-1)}\{(J-1)\sum_i \sigma^2_{AB:i} + \frac{(I-1)(J-1)}{K}\sigma^2\} = \sigma^2 + K\sigma^2_{AB}.$$

Pagaliau

$$\begin{aligned}\mathbf{E}(MS_A) &= \frac{JK}{I-1}\{\sum_i \alpha_i^2 + \mathbf{E}(\sum_i \bar{c}_{i.}^2) + \mathbf{E}(\sum_i (\bar{e}_{i..} - \bar{e}_{...})^2)\} \\ &= \frac{JK}{I-1}\{\sum_i \alpha_i^2 + \frac{1}{J}\sum_i \sigma^2_{AB;i} + \frac{I-1}{JK}\sigma^2\} \\ &= \sigma^2 + K\sigma^2_{AB} + JK\sigma^2_A, \quad \sigma^2_A = \frac{1}{I-1}\sum_i \alpha_i^2.\end{aligned}$$

Surašę rezultatus į dispersinės analizės lentelę, matome, kad ji skiriasi nuo 2.2.1 ar 2.4.1 lentelių tik paskutiniuoju stulpeliu.

2.6.1 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	$MS = SS/\nu$	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + K\sigma^2_{AB} + JK\sigma^2_A$
B	SS_B	$J-1$	MS_B	$\sigma^2 + IK\sigma^2_B$
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}	$\sigma^2 + K\sigma^2_{AB}$
E	SS_E	I J(K-1)	MS_E	σ^2
T	SS_T	I J K -1	-	-

2.6.3. Faktorių įtakos ir sąveikos nebuvimo hipotezių tikrinimas

Pereiname prie pagrindinių dispersinės analizės hipotezių tikrinimo.

Tikrindami hipotezę $H_B : \sigma_B^2 = 0$ apie atsitiktinio faktoriaus B įtaką, sudarome statistiką

$$F_B = \frac{MS_B}{MS_E}.$$

Pagal 2.6.1 teoremą, kai hipotezė teisinga, statistika F_B turi Fišerio skirstinį su $J-1$ ir $IJ(K-1)$ laisvės laipsniais. Hipotezė atmetama α lygmens kriterijumi, kai galioja nelygybė

$$F_B > F_\alpha(J-1, IJ(K-1)). \tag{2.6.6}$$

Kriterijaus galia išreiškiama centrinio Fišerio skirstinio pasiskirstymo funkcija. Atkreipiame dėmesį, kad nors tikrinama hipotezė apie atsitiktinio faktoriaus įtaką, tačiau statistikos F_B vardiklyje yra ne MS_{AB}, o MS_E (kaip ir modelyje su pastoviais faktoriais).

Parodėme, kad atsitiktiniai dydžiai $SS_A/\mathbf{E}(MS_A)$ ir $SS_{AB}/\mathbf{E}(MS_{AB})$ pasiskirstę pagal chi kvadrato dėsnį tik kai $\sigma_{AB}^2 = 0$. Priešingu atveju šių kvadratų sumų skirstiniai nėra nei centriniai, nei necentriniai chi kvadrato skirstiniai. Jie yra sumos kvadratų priklausomų atsitiktinių dydžių su skirtingomis dispersijomis.

Jeigu hipotezė $H_{AB} : \sigma_{AB}^2 = 0$ yra teisinga, tai pagal 2.5.1 teoremą statistika

$$F_{AB} = \frac{MS_{AB}}{MS_E}$$

turi Fišerio skirstinį su $(I-1)(J-1)$ ir $IJ(K-1)$ laisvės laipsniais. Hipotezė atmetama α lygmens kriterijumi, kai galioja nelygybė

$$F_{AB} > F_\alpha((I-1)(J-1), IJ(K-1)). \tag{2.6.7}$$

Šio kriterijaus galia žinomais paprastais skirstiniais neišreiškiama.

Kai teisinga hipotezė $H_A : \alpha_1 = ... = \alpha_I = 0$, statistikos

$$F_A = \frac{MS_A}{MS_{AB}}$$

skaitiklio ir vardiklio vidurkiai yra vienodi (žr. 2.6.1 lentelės paskutinį stulpelį), tačiau jų skirstiniai nėra χ^2 skirstiniai. Todėl kriterijus: atmesti hipotezę H_A, kai teisinga nelygybė

$$F_A > F_\alpha(I-1, (I-1)(J-1)), \tag{2.6.8}$$

yra apytikslis. Tikslus kriterijus hipotezei H_A tikrinti kuriamas remiantis Hotelingo statistika (žr. 4 dalies 2 skyrių). Jeigu faktorių sąveikos nėra, t. y. teisinga adityvumo hipotezė $H_{AB} : \sigma_{AB}^2 = 0$, tai remiantis (2.6.4) galima tvirtinti, kad statistika F_A turi centrinį Fišerio skirstinį su nurodytais laisvės laipsnių skaičiais, kai H_A teisinga, ir necentrinį Fišerio skirstinį, kai H_A neteisinga. Taigi šiuo atveju kriterijus (2.6.7) yra tikslus, o kriterijaus galia išreiškiama necentrinio Fišerio skirstinio pasiskirstymo funkcija.

2.6.4. Parametrų vertinimas

Taškinius parametrų, kuriais išreiškiami vidutinių kvadratų sumų vidurkiai $\mathbf{E}(MS)$, įvertiniai gaunami remiantis 2.6.1 lentelės paskutiniu stulpeliu. Pavyzdžiui,

$$\hat{\sigma}_B^2 = \frac{MS_B - MS_E}{IK}, \quad \hat{\sigma}_{AB}^2 = \frac{MS_{AB} - MS_E}{K}.$$

Pradinių modelio parametrų: vidurkių vektoriaus $\boldsymbol{\mu} = (\mu_1, ..., \mu_I)^T$ ir kovariacinės matricos $\boldsymbol{\Sigma} = [\sigma_{ij}]_{I\times I}$ elementų taškinius įvertinius gauname remdamiesi imtimi $\bar{\boldsymbol{Y}}_{j.} = (\bar{Y}_{1j.}, ..., \bar{Y}_{Ij.})^T, j = 1, ..., J$. Atsitiktiniai vektoriai $\bar{\boldsymbol{Y}}_{j.}$ yra nepriklausomi, vienodai pasiskirstę ir normalieji $\bar{\boldsymbol{Y}}_{j.} \sim N_I(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, $\boldsymbol{\Lambda} = [\lambda_{ij}]_{I\times I}$, $\lambda_{ij} = \sigma_{ij}$, kai $i \neq j$, ir $\lambda_{ii} = \sigma_{ii} + \sigma^2/K$. Gauname

$$\hat{\boldsymbol{\mu}} = \frac{1}{J}\sum_{j=1}^{J} \bar{\boldsymbol{Y}}_{j.} = (\bar{Y}_{1..}, ..., \bar{Y}_{I..})^T \sim N_I(\boldsymbol{\mu}, \frac{1}{J}\boldsymbol{\Lambda});$$

$$\hat{\sigma}_{ii'} = \frac{1}{J-1}\sum_{j=1}^{J}(\bar{Y}_{ij.} - \bar{Y}_{i..})(\bar{Y}_{i'j.} - \bar{Y}_{i'..}), \quad i \neq i';$$

$$\hat{\sigma}_{ii} = \frac{1}{J-1}\sum_{j=1}^{J}(\bar{Y}_{ij.} - \bar{Y}_{i..})^2 - \frac{\hat{\sigma}^2}{K}, \quad \hat{\sigma}^2 = MS_E.$$

2.6.3 pavyzdys. Atliekant eksperimentą buvo tiriama rodiklio, apibūdinančio pagamintų gaminių kiekį ir jų kokybę (analizuojamas kintamasis Y), priklausomybė nuo staklių tipo (pastovus faktorius A) ir nuo darbininko (atsitiktinis faktorius B). Stebėjimų rezultatai pateikti 2.6.2 lentelėje.

2.6.2 lentelė. Statistiniai duomenys

	A_1	A_2	A_3
B_1	51,8; 52,8	59,7; 60,0	61,5; 61,7
B_2	51,1; 52,3	63,2; 62,8	64,1; 66,2
B_3	50,9; 51,8	64,8; 65,0	72,1; 72,0
B_4	46,4; 44,8	43,7; 44,2	62,0; 61,4

Atlikę skaičiavimus, gauname statistikų realizacijas: $MS_A = 443,434$, $MS_B = 167,624$, $MS_{AB} = 41,075$, $MS_E = 0,465$. Hipotezėms H_A, H_B, ir H_{AB} tikrinti randame santykius $F_A = MS_A/MS_{AB} = 10,80$; $F_B = MS_B/MS_E = 360,16$; $F_{AB} = MS_{AB}/MS_E = 88,26$. Hipotezės H_B ir H_{AB} atmetamos. Hipotezė H_A atmetama, jei pasirinktas reikšmingumo lygmuo viršija P reikšmę $\mathbf{P}\{F_{2,6} > 10,80\} = 0,01$. Primename, kad pastarasis kriterijus yra apytikslis, jei hipotezė H_{AB} neteisinga.

2.7. Daugiafaktorė dispersinė analizė

Faktorių (fiksuotų ar atsitiktinių) skaičius gali būti didesnis už du. Pavyzdžiui, reklamos efektyvumą gali lemti skelbimo lentos apipavidalinimas (faktorius A), skaitančiojo lytis (faktorius B) ir rasė (faktorius C). Jei lyginamos kelios konkrečios skirtingai apipavidalintos skelbimų lentos, tai faktorius A yra fiksuotas,

jei norima ištirti, ar apskritai apipavidalinimas turi įtakos reklamos sėkmei, tai faktorius A yra atsitiktinis (tuo atveju atsitiktinai parenkama keletas skirtingai apipavidalintų skelbimų lentų). Faktoriai B ir C yra fiksuoti.

Palyginus tris dvifaktorės dispersinės analizės modelius nesunku įžvelgti tam tikrą simetriją (palyginkite 2.2.1, 2.4.1 ir 2.5.1 lenteles), kuria remiantis galima nurodyti formalias dispersinės analizės lentelės užpildymo taisykles, kai faktorių skaičius yra didesnis už du ir eksperimento planas subalansuotas. Pateiksime tokias taisykles, iliustruodami jas trifaktoriu modeliu, kuriame du faktoriai A ir B yra pastovūs, o trečiasis faktorius C – atsitiktinis.

2.7.1. Statistinis modelis

Faktorių A, B ir C lygmenis numeruokime indeksais i, j ir l; $i = 1, ..., I$, $j = 1, ..., J$, $l = 1, ..., L$; atsitiktinių paklaidų nestebimo faktoriaus E lygmenis – indeksu k, $k = 1, ..., K$. Taigi stebėjimo rezultatus Y_{ijlk} numeruosime keturiais indeksais; bendras stebėjimų skaičius $n = IJLK$. Stebėjimus Y_{ijlk} išskaidome į komponentes analogiškai kaip pirmiau nagrinėtuose modeliuose:

$$Y_{ijlk} = \mu + \alpha_i^A + \alpha_j^B + a_l^C + \alpha_{ij}^{AB} + a_{il}^{AC} + a_{jl}^{BC} + a_{ijl}^{ABC} + e_{ijlk}. \qquad (2.7.1)$$

Kaip matome, šiame skaidinyje išskirtas bendrasis vidurkis μ ir komponentė e_{ijlk}, apibūdinanti atsitiktines matavimo paklaidas (tarsime, kad a. d. e_{ijlk} nepriklausomi tarpusavyje ir nepriklauso nuo kitų a. d., turi nulinius vidurkius ir vienodas dispersijas σ^2). Kitos (2.7.1) skaidinio komponentės apibūdina atskirų faktorių ir jų rinkinių (skirtingų faktorių dvejetų, trejetų sąveikos) įtaką. Dėl simetrijos šios komponentės turi viršutinius ir apatinius indeksus; viršutiniai indeksai žymi atitinkamą faktorių arba jų rinkinį, apatiniai – tų faktorių lygmenų numerius. Jeigu visi viršuje parašyti faktoriai yra pastovūs, tai atitinkamos komponentės yra konstantos ir žymimos raide α, jei nors vienas faktorius yra atsitiktinis, tai atitinkamos komponentės yra a. d. ir žymimos raide a.

Bet kurių komponenčių, pažymėtų raide α arba a, suma, kai indeksas, atitinkantis pastovų faktorių, perbėga visas galimas reikšmes, lygi nuliui su bet kuriomis kitų indeksų reikšmėmis:

$$\sum_i \alpha_i^A = \sum_j \alpha_j^B = \sum_i \alpha_{ij}^{AB} = \sum_j \alpha_{ij}^{AB} =$$

$$= \sum_i a_{il}^{AC} = \sum_j a_{jl}^{BC} = \sum_i a_{ijl}^{ABC} = \sum_j a_{ijl}^{ABC} = 0.$$

Visi a. d., žymimi raidėmis a, turi nulinius vidurkius. Jų dispersijos nepriklauso nuo indeksų, atitinkančių atsitiktinius faktorius, tačiau priklauso nuo indeksų, atitinkančių pastovius faktorius (jeigu jų yra). Dispersijas žymėsime raidėmis σ^2 su indeksais apačioje – iš pradžių tais, kurie parašyti atitinkamo dėmens viršuje, po to pastovių faktorių indeksais, jeigu jų yra:

$$\boldsymbol{V}(a_l^C) = \sigma_C^2, \quad \boldsymbol{V}(a_{il}^{AC}) = \sigma_{AC;i}^2, \quad \boldsymbol{V}(a_{jl}^{BC}) = \sigma_{BC;j}^2, \quad \boldsymbol{V}(a_{ijl}^{ABC}) = \sigma_{ABC;ij}^2. \qquad (2.7.2)$$

Vidurkiai $\mathbf{E}(MS)$ priklauso nuo σ^2 ir parametrų, kuriuos žymėsime raide σ^2 su indeksais apačioje, sutampančiais su (2.7.1) skaidinio viršutiniais indeksais. Jie gaunami šitaip:

1) jeigu (2.7.1) komponentė priklauso tik nuo atsitiktinių faktorių, tai parametras lygus tos komponentės dispersijai (pvz., $\sigma^2_C = \boldsymbol{V}(a^C_l)$);

2) jeigu komponentėje yra ir pastovių, ir atsitiktinių faktorių, tai parametrai, gaunami sumuojant (2.7.2) dispersijas pagal pastovių faktorių indeksus ir kiekvieną kartą dalijant iš lygmenų skaičiaus be vieneto:

$$\sigma^2_{AC} = \frac{1}{I-1}\sum_i \sigma^2_{AC;i}, \quad \sigma^2_{BC} = \frac{1}{J-1}\sum_j \sigma^2_{BC;j},$$

$$\sigma^2_{ABC} = \frac{1}{(I-1)(J-1)}\sum_i\sum_j \sigma^2_{ABC;ij};$$

3) jeigu komponentės viršuje yra tik pastovūs faktoriai, tai sumuojami atitinkamų komponenčių kvadratai, kiekvieną kartą dalijant iš lygmenų skaičiaus be vieneto:

$$\sigma^2_A = \frac{1}{I-1}\sum_i (\alpha^A_i)^2, \quad \sigma^2_B = \frac{1}{J-1}\sum_j (\alpha^B_j)^2,$$

$$\sigma^2_{AB} = \frac{1}{(I-1)(J-1)}\sum_i\sum_j (\alpha^{AB}_{ij})^2.$$

2.7.2. Kvadratų sumų sudarymas

Kvadratų sumos sudaromos imant kiekvieną (2.7.1) skaidinio komponentę (išskyrus μ); jos žymimos dviem raidėmis SS su indeksais apačioje, sutampančiais su atitinkamo (2.7.1) dėmens viršutiniais indeksais (komponentę e_{ijlk} atitinkanti kvadratų suma žymima SS_E). Šias kvadratų sumas gauname taip: 1) sudarome simbolines sandaugas imdami atitinkamo (2.7.1) dėmens apatinius indeksus be vieneto (pavyzdžiui, komponentę a^{AC}_{il} atitinka simbolinė sandauga $(i-1)(l-1) = il - i - l + 1$; komponentei e_{ijlk} priskiriame simbolinę sandaugą $ijl(k-1) = ijlk - ijl$);

2) sudarome stebėjimų tiesinį darinį imdami juos su tokiais ženklais ir indeksais kaip ir atitinkamoje simbolinėje sandaugoje, trūkstamus indeksus pakeičiame taškais ir brūkšniu viršuje (pavyzdžiui, komponentę a^{AC}_{il} atitinka tiesinis darinys $\bar{Y}_{i.l.} - \bar{Y}_{i...} - \bar{Y}_{..l.} + \bar{Y}_{....}$);

3) kvadratų sumą SS gauname sumuodami gautųjų tiesinių darinių kvadratus, kai visi indeksai perbėga galimas reikšmes. Pateiksime kvadratų sumas nagrinėjamo pavyzdžio atveju:

$$SS_A = JLK\sum_i (\bar{Y}_{i...} - \bar{Y}_{....})^2, \quad SS_B = ILK\sum_j (\bar{Y}_{.j..} - \bar{Y}_{....})^2,$$

$$SS_C = IJK\sum_l (\bar{Y}_{..l.} - \bar{Y}_{....})^2; \quad SS_{AB} = LK\sum_i\sum_j (\bar{Y}_{ij..} - \bar{Y}_{i...} - \bar{Y}_{.j..} + \bar{Y}_{....})^2,$$

$$SS_{AC} = JK \sum_i \sum_l (\bar{Y}_{i.l.} - \bar{Y}_{i...} - \bar{Y}_{..l.} + \bar{Y}_{....})^2,$$

$$SS_{BC} = IK \sum_j \sum_l (\bar{Y}_{.jl.} - \bar{Y}_{.j..} - \bar{Y}_{..l.} + \bar{Y}_{....})^2;$$

$$SS_{ABC} = K \sum_i \sum_j \sum_l (\bar{Y}_{ijl.} - \bar{Y}_{ij..} - \bar{Y}_{i.l.} - \bar{Y}_{.jl.} + \bar{Y}_{i...} + \bar{Y}_{.j..} + \bar{Y}_{..l.} - \bar{Y}_{....})^2;$$

$$SS_E = \sum_i \sum_j \sum_l \sum_k (Y_{ijlk} - \bar{Y}_{ijl.})^2.$$

2.7.1 pastaba. Kvadratų sumas rekomenduojama skaičiuoti taip: sumuoti kiekvieno dėmens kvadratus atskirai, paliekant tuos pačius ženklus kaip ir simbolinėje sandaugoje. Pavyzdžiui,

$$SS_{AC} = JK \sum_i \sum_l \bar{Y}^2_{i.l.} - JLK \sum_i \bar{Y}^2_{i...} - IJK \sum_l \bar{Y}^2_{..l.} + IJLK\, \bar{Y}^2_{....}.$$

2.7.3. Laisvės laipsnių skaičių radimas

Laisvės laipsnių skaičių gauname pakeitę atitinkamoje simbolinėje sandaugoje mažąsias raides didžiosiomis.

2.7.4. Vidurkių $\mathbf{E}(MS)$ radimas

Akivaizdu, kad $\mathbf{E}(MS_E) = \sigma^2$. Kiti vidurkiai $\mathbf{E}(MS)$ yra tiesiniai dariniai σ^2 ir tų raide σ^2 žymimų parametrų, tarp kurių apatinių indeksų yra visi faktoriai, esantys tarp atitinkamos kvadratų sumos SS apatinių indeksų. Pavyzdžiui, $\mathbf{E}(MS_{AC})$ yra σ^2 ir parametrų σ^2_{AC}, σ^2_{ABC} tiesinis darinys; $\mathbf{E}(MS_A)$ yra σ^2, σ^2_A, σ^2_{AB}, σ^2_{AC}, σ^2_{ABC} tiesinis darinys. Koeficientas prie dispersijos σ^2 visada lygus vienetui. Ieškant koeficientų prie kitų parametrų, patogu prieš tai sudaryti pagalbinę lentelę. Lentelės pirmoje eilutėje surašome skaidinio (2.7.1) komponentes (išskyrus pirmąjį ir paskutinį dėmenis), o pirmajame stulpelyje – faktorių indeksus. Paskui lentelę nuosekliai užpildome šitaip:

a) jei stulpelio pavadinime yra pastovus faktorius, tai šio stulpelio sankirtoje su eilute, kurios indeksas žymi to faktoriaus lygmenis, įrašome 0; jei stulpelio pavadinime yra atsitiktinis faktorius, tai to stulpelio ir atitinkamos eilutės sankirtoje įrašome 1;

b) į kitus lentelės langelius įrašome atitinkamos eilutės indekso maksimalią reikšmę (t. y. didžiąją raidę).

2.7.1 lentelė. Pagalbinė lentelė

Indeksai	α_i^A	α_j^B	a_l^C	α_{ij}^{AB}	a_{il}^{AC}	a_{jl}^{BC}	a_{ijl}^{ABC}
i	0	I	I	0	0	I	0
j	J	0	J	0	J	0	0
l	L	L	1	L	1	1	1
k	K	K	K	K	K	K	K
-	σ^2_A	σ^2_B	σ^2_C	σ^2_{AB}	σ^2_{AC}	σ^2_{BC}	σ^2_{ABC}

Kiekvieną pagalbinės lentelės stulpelį atitinka parametras σ^2 su tokiais indeksais, kurie yra stulpelio pavadinime (jie surašyti 2.7.1 lentelės paskutinėje eilutėje).

Koeficientas prie parametro σ^2 su indeksais apačioje $\mathbf{E}(MS)$ išraiškoje gaunamas sudauginus atitinkamo stulpelio elementus, kai eilutės, pažymėtos indeksais, esančiais to stulpelio pavadinime, yra išbrauktos. Pavyzdžiui, $\mathbf{E}(MS_{AC})$ išraiškoje koeficientas prie σ^2_{AC} yra JK, o prie σ^2_{ABC} yra 0 (eilutės, pažymėtos indeksais i ir l, yra išbrauktos).

Lentelėje 2.7.2 pateikti duomenys, reikalingi (2.7.1) modelio dispersinei analizei atlikti.

2.7.2 lentelė. Trifaktorės dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + JK\sigma^2_{AC} + JLK\sigma^2_A$
B	SS_B	$J-1$	MS_B	$\sigma^2 + IK\sigma^2_{BC} + ILK\sigma^2_B$
C	SS_C	$L-1$	MS_C	$\sigma^2 + IJK\sigma^2_C$
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}	$\sigma^2 + K\sigma^2_{ABC} + LK\sigma^2_{AB}$
$A \times C$	SS_{AC}	$(I-1)(L-1)$	MS_{AC}	$\sigma^2 + JK\sigma^2_{AC}$
$B \times C$	SS_{BC}	$(J-1)(L-1)$	MS_{BC}	$\sigma^2 + IK\sigma^2_{BC}$
$A \times B \times C$	SS_{ABC}	$(I-1)(J-1)$ $(L-1)$	MS_{ABC}	$\sigma^2 + K\sigma^2_{ABC}$
E	SS_E	$IJL(K-1)$	MS_E	σ^2
T	SS_T	$IJLK-1$		

2.7.5. Parametrų įvertiniai ir hipotezių tikrinimas

Suformuluosime keletą sudarytos lentelės analizės taisyklių.

1. Nepaslinktuosius parametrų, pažymėtų raidėmis σ^2, įvertinius gauname sudarydami tiesinius SS darinius ir atsižvelgdami į $\mathbf{E}(MS)$ išraiškas. Pavyzdžiui, iš 2.7.2 lentelės gauname

$$\hat\sigma^2 = MS_E, \quad \hat\sigma^2_A = \frac{MS_A - MS_{AC}}{JLK}, \quad \hat\sigma^2_{AB} = \frac{MS_{AB} - MS_{ABC}}{LK}.$$

2. Jeigu $K = 1$, tai kvadratų sumos SS_E laisvės laipsnių skaičius lygus 0, todėl tenka daryti prielaidą, kad kažkurie parametrai σ^2 lygūs 0 (paprastai tariama, kad aukščiausios eilės sąveika lygi 0). Dispersinės analizės lentelė tuo atveju, kai $K = 1$, lengvai gaunama iš jau sudarytos. Pavyzdžiui, jeigu $K = 1$ ir $\sigma^2_{ABC} = 0$, tai 2.7.2 lentelėje reikėtų praleisti eilutę, atitinkančią faktorių E, įrašyti vietoje K vienetą ir vietoje σ^2_{ABC} nulį, praleisti tašką indekso k vietoje.

3. Hipotezių apie parametrų σ^2 su indeksais lygybę 0 tikrinimo kriterijai grindžiami statistikomis, kurios yra dviejų MS santykiai. Jie parenkami taip, kad skaitiklio ir vardiklio vidurkiai $\mathbf{E}(MS)$, kai teisinga hipotezė, būtų vienodi. Tokie santykiai, kai hipotezė teisinga, pasiskirstę pagal Fišerio skirstinį su laisvės laipsniais, pateiktais 3 stulpelyje (mišriajame modelyje kai kurie santykiai skiriasi nuo Fišerio, todėl gaunami kriterijai yra apytiksliai).

Jeigu žinoma, kad kai kurie parametrai σ^2 lygūs 0, tai 2.7.2 lentelę reikia pertvarkyti sujungiant kvadratų sumas ir laisvės laipsnius tų eilučių, kurių $\mathbf{E}(MS)$ yra vienodi.

2.7.2 pastaba. Kartais negalima parinkti tokių dviejų sumų MS, kad jų vidurkiai, kai teisinga tikrinamoji hipotezė, sutaptų. Pavyzdžiui, jeigu visi trys faktoriai A, B, C yra atsitiktiniai, tai iš 2.7.2 lentelės analogo gauname, kad $\mathbf{E}(MS_A) = \tau = \sigma^2 + K\sigma^2_{ABC} + LK\sigma^2_{AB} + JK\sigma^2_{AC}$, kai hipotezė $H_A : \sigma^2_A = 0$ yra teisinga. Nėra kitos kvadratų sumos MS, kurios vidurkis būtų lygus τ. Tokiu atveju statistikos vardiklyje yra įrašoma toks tiesinis kelių sumų MS darinys, kad jo vidurkis sutaptų su skaitiklio vidurkiu. Minėtame pavyzdyje vidurkį τ turi tiesinė funkcija $MS_{AB}+MS_{AC}-MS_{ABC}$. Suprantama, kad tokia funkcija nėra proporcinga atsitiktiniam dydžiui, turinčiam chi kvadrato skirstinį. Apytikslis Fišerio kriterijus gaunamas aproksimuojant vardiklį a. d. $(\tau/\nu)\chi^2_\nu$. Laisvės laipsnių skaičius ν parenkamas iš sąlygos, kad minėto tiesinio kodarinio dispersija sutaptų su a. d. $(\tau/\nu)\chi^2_\nu$ dispersija $2\tau^2/\nu$. Nežinomi parametrai τ išraiškoje pakeičiami jų įvertiniais. Šio metodo iliustraciją žr. 2.33 pratime.

2.7.1 pavyzdys. Konservų gamykloje atliktas eksperimentas pagal pilną trifaktorės dispersinės analizės schemą su vienu stebėjimu langelyje: faktorius A – konservų dėžutės užpildymas, t. y. įdėtų vyšnių svoris; faktorius B – cukraus sirupo koncentracija; faktorius C – gauto produkto spalva (C_1 – šviesus, C_2 – vidutinis, C_3 – tamsus). Lentelėje 2.7.3 sąlyginiais vienetais pateiktas gauto produkto svoris [14].

2.7.3 lentelė. Statistiniai duomenys

	B_1			B_2			B_3		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
A_1	55	95	169	55	69	163	49	88	153
A_2	200	232	223	183	215	207	148	200	245
A_3	233	285	291	236	259	278	233	223	259

Atliekame skaičiavimus tardami, kad visi faktoriai pastovūs ir visų trijų faktorių sąveika lygi nuliui. Gauname statistikų realizacijas: $F_A = 179,41$, $F_B = 3,07$, $F_C = 31,15$, $F_{AB} = 0,38$, $F_{AC} = 4,38$, $F_{BC} = 0,52$, kurias atitinka P reikšmės $2,3 \cdot 10^{-7}$, $0,1024$, $0,0002$, $0,8190$, $0,0362$, $0,7257$. Matome, kad atmesti hipotezes H_B, H_{AB} ir H_{BC} nėra pagrindo. O kitos hipotezės yra atmestinos. Galima daryti išvadą, kad faktorius B neturi įtakos tiriamajam požymiui. Tuo remiantis galima sujungti kvadratų sumas SS_B, SS_{AB} ir SS_{BC} su SS_{ABC} sudedant ir atitinkamus laisvės laipsnius, t. y. nagrinėti dvifaktorės analizės schemą su faktoriais A ir C, o faktoriaus B lygmenys atitiktų stebėjimų kartotinumą. Gautume statistikų realizacijas $F_A = 182,22$, $F_C = 31,64$ ir $F_{AC} = 4,45$, kurioms atitinka P reikšmės $1,1 \cdot 10^{-12}$, $1,3 \cdot 10^{-6}$ ir $0,0112$. Hipotezės atmetamos, jei kriterijų reikšmingumo lygmuo viršija $0,0112$.

2.8. Dispersinės analizės eksperimentų planai

Eksperimento planavimo tikslas – gauti kuo daugiau informacijos minimaliomis sąnaudomis. Pradiniame kiekvieno eksperimento etape turi būti suformuluotas uždavinys, parinktas priklausomas kintamasis (požymis), kuris bus nagrinėjamas, nustatyti pagrindiniai faktoriai, nuo kurių jis turėtų labiausiai priklausy-

ti. Paskui turėtų būti atliktas eksperimento planavimas, kurio metu parenkamas eksperimentų skaičius, dalyvaujančių eksperimente faktorių lygmenų rinkiniai, eksperimento atlikimo tvarka ir numatomas matematinis modelis, kuriuo remiantis bus nagrinėjami eksperimento rezultatai.

Eksperimentų planai skiriasi tiek pagal faktorių lygmenų parinkimo, tiek pagal eksperimento objektų, atitinkančių pasirinktų faktorių lygmenis, parinkimo pobūdį.

Parenkant eksperimentų atlikimo tvarką labai svarbu duomenų *randomizavimas*. Atsižvelgiant į *eksperimento objektų parinkimo pobūdį*, dispersinėje analizėje dažniausiai naudojami du randomizuoti eksperimento planai:

1. *Visiškai randomizuotas eksperimentų planas.* Fiksavus bet kurią eksperimente naudojamą faktorių lygmenų rinkinį kombinaciją, atsitiktinai parenkami vienas ar keli objektai, atitinkantys šį rinkinį, ir kiekvieno iš jų matuojama tiriamo požymio Y reikšmė.
2. *Randomizuotas blokuotųjų duomenų planas.* Jei eksperimente dalyvaujantys objektai nėra homogeniški dėl priežasčių, nesusijusių su tiriamais faktoriais, tai tų objektų skirtumai gali paslėpti faktorių įtaką. Siekiant to išvengti, duomenys skirstomi į blokus. Jei eksperimente naudojamų faktorių lygmenų rinkinių yra M, tai vienam blokui priskiriama M atsitiktinai parinktų kuriuo nors požiūriu homogeniškų (nebūtinai skirtingų) objektų. Bloko i-asis objektas stebimas, kai fiksuotas i-asis faktorių lygmenų rinkinys. Dažniausiai vieną bloką sudaro atsitiktinai parinktas tas pats objektas (jis pats sau homogeniškas), kurio požymio Y reikšmė matuojama M kartų, kiekvieną kartą esant skirtingam faktorių lygmenų rinkiniui. Jei tai įmanoma, faktorių išdėstymo tvarka bloke – atsitiktinė. Matavimų vektorius kiekviename bloke dažniausiai sudarytas iš priklausomų atsitiktinių dydžių, nes visi matavimai susiję.

2.8.1 pavyzdys. Firma gamina kartonines dėžes, naudodama keturias technologijas. Tiriama naudotos technologijos (požymis A) įtaka atitinkančių standartą gaminių, pagamintų per dieną, skaičiui Y.

Visiškai randomizuotas eksperimentų planas: kiekvienu iš keturių metodų apmokoma dirbti po tris atsitiktinai parinktus darbininkus. Stebima, kiek atitinkančių standartą dėžių pagamina kiekvienas iš 12 darbininkų per dieną (kiekvienai technologijai trys skirtingi darbininkai).

Darbininkų gabumai gali būti skirtingi. Tie skirtumai gali paslėpti technologijų, skirtumus, todėl natūralu naudoti kitą eksperimentų planą.

Randomizuotas blokuotųjų duomenų planas: atsitiktinai parenkami trys darbininkai ir apmokomi naudotis visomis keturiomis technologijomis. Kiekvienas iš trijų darbininkų keturias dienas gamina dėžes kiekvieną dieną naudodamas skirtingą technologiją. Technologijų parinkimo tvarka kiekvienam darbininkui – atsitiktinė. Naudojant šį planą darbininkų gabumai nepaslepia technologijų skirtumo. Vieno darbininko rezultatai sudaro keturmatį vektorių (bloką), kurio komponentės yra priklausomi atsitiktiniai dydžiai.

2.8.2 pavyzdys. Reikia nustatyti, ar keturių markių I, II, III, IV padangos dyla vienodu greičiu. Tiriamas požymis Y yra protektoriaus gylio sumažėjimas po T kilometrų ridos nuo eksploatacijos pradžios. Tarkime, kad turime po vieną komplektą iš keturių kiekvienos markės padangų ir keturis automobilius.

Visiškai randomizuotas eksperimentų planas: atsitiktinai parenkamas vienas iš galimų 16-os padangų išdėstymų 16 pozicijų (tikimybė $1/(16!)$). Randomizavimo tikslas – suvidurkinti vi-

sus nuo automobilio priklausančius skirtumus, kurie gali turėti įtakos rezultatui. Automobilių skirtumai bus įtraukti modelyje į atsitiktinę paklaidą. Jeigu automobilio įtaka padangos dilimui didelė, tai paklaidos dispersija gali žymiai padidėti. Tada nepastebėsime padangų markių skirtumo, nors faktiškai jis gal ir yra.

Randomizuotas blokuotųjų duomenų planas: ant kiekvieno automobilio sumontuokime visų keturių markių (I, II, III, IV) padangas. Tai reiškia, kad nuo vieno faktoriaus (padangos markė) dirbtinai pereiname prie dviejų faktorių: pirmas faktorius – padangos markė, antrasis (trukdantysis) – automobilis. Padangos išdėstomos ant automobilio atsitiktinai parenkant vieną iš galimų 4 padangų išdėstymo 4 pozicijose variantą su tikimybe $1/(4!)$. Toks eksperimento planas leidžia daryti išvadas apie padangos markės įtaką jos dilimui eliminuojant automobilio (ir padangos padėties vietos) įtaką. Ant vieno automobilio sumontuotų padangų protektorių gylio sumažėjimai sudaro keturmatį vektorių (bloką).

Šio pavyzdžio galimi ir kiti eksperimento planai. Jeigu pasirodytų, kad padangos pozicija daro didelę įtaką jai dilti, gali būti tikslinga planuoti eksperimentą taip, kad būtų eliminuota ir šio trukdančiojo faktoriaus įtaka. Pavyzdžiui, po $T/4$ ridos perkelti padangas į kitas pozicijas. Tokiu atveju atsiranda randomizacijos apribojimų. Pirmajame etape padangas išdėstome skirtingose automobilio pozicijose atsitiktinai. Padangos montuojamos į kitas pozicijas atsitiktinai, tačiau su sąlyga, kad jos nepateks į tą poziciją, kurioje buvo.

Jei būtų naudojamas planas be randomizacijos: pirmos markės padangos sumontuojamos ant pirmo automobilio (I, I, I, I), antros markės padangos – ant antrojo automobilio (II, II, II, II) ir t.t., tai jis būtų blogas, nes padangų markės susiejamos su automobiliais. Pastebėtas padangų dilimo skirtumas galėtų reikšti ne padangos markės, o automobilio įtaką padangai dilti (automobilio techninė būklė, vairuotojo stažas, eksploatacijos sąlygos ir kt.).

Atsižvelgiant į *faktorių lygmenų parinkimo pobūdį*, dispersinėje analizėje su fiksuotais faktoriais dažniausiai naudojami du eksperimento planai:

1. *Kryžminė klasifikacija:* naudojami visi galimi faktorių lygmenų rinkiniai.
2. *Hierarchinė klasifikacija:* faktoriaus A (vadinamo grupuojančiu) lygmuo A_1 derinamas su J_1 kito faktoriaus B (vadinamo sugrupuotu pagal A) lygmenimis, A_2 su J_2 iš likusių faktoriaus B lygmenų ir t. t.

Hierarchinė klasifikacija gali būti naudojama ir kai vienas ar abu faktoriai yra atsitiktiniai. Šiuo atveju fiksavus faktoriaus A reikšmę A_i, J_i lygmenų parenkama ne iš fiksuoto jo lygmenų poaibio, o iš visų galimų lygmenų aibės.

Kryžminė klasifikacija gali būti nepriimtina, kai yra eksperimentų skaičiaus apribojimų dėl jų didelės kainos ar trukmės. Kai klasifikacija hierarchinė, eksperimentų skaičius mažesnis. Pavyzdžiui, jei faktorius A įgyja keturias reikšmes, o faktorius B – dvidešimt reikšmių, tai kryžminės klasifikacijos atveju yra $4 \times 20 = 80$ faktorių lygmenų rinkinių. Jei $J_1 = \cdots = J_4 = 5$, tai hierarchinės klasifikacijos atveju naudojama $4 \times 5 = 20$ faktorių lygmenų rinkinių.

2.8.3 pavyzdys. Tarkime, kad faktorius A yra kviečių auginimo metodika (sena ir nauja), o B yra kviečių veislė, įgyjanti 10 reikšmių. Kryžminėje klasifikacijoje atveju naudojama $2 \times 10 = 20$ požymių reikšmių rinkinių. Hierarchinei klasifikacijai atsitiktinai parinktos, sakykime, 4 kviečių veislės auginamos naudojant seną, o likusios 6 – pagal naują metodiką. Šiuo atveju naudojama $4 + 6 = 10$ faktorių lygmenų rinkinių. Auginimo metodika A yra grupuojantis faktorius, kviečių veislė B – sugrupuotas pagal auginimo metodiką faktorius.

Kartais iš principo negalima atlikti eksperimento pagal kryžminės klasifikacijos planą ir faktoriaus B reikšmės, atitinkančios faktoriaus A reikšmę A_i, gali būti parinktos tik iš kurio nors visų galimų faktoriaus B lygmenų poaibio.

2.8.4 pavyzdys. Tarkime, kad gaminiai gali būti gaminami esant I skirtingiems technologinio proceso režimams. Iš gaminių, pagamintų i-uoju režimu, aibės atsitiktinai atrenkame J_i gaminių ir j-ojo gaminio tiriamą parametrą pamatuojame K_{ij} kartų. Tegu faktoriaus A lygmenys atitinka skirtingus technologinio proceso režimus, o faktoriaus B lygmenys – gaminių numerius. Kadangi praktiškai tas pats gaminys negali būti pagamintas skirtingais technologinio proceso režimais, tai su kiekvienu faktoriaus A lygmeniu bus matuojamos skirtingų gaminių tiriamo parametro reikšmės. Taigi kryžminė klasifikacija negalima ir turime hierarchinės klasifikacijos planą, kuriame faktoriaus B lygmenys sugrupuoti pagal faktorių A. Šiame pavyzdyje faktorius B atsitiktinis.

Matėme, kad eksperimentų planus galima skirstyti ir pagal eksperimentų skaičių, kai fiksuojami įvairūs faktorių lygmenų rinkiniai: eksperimento planas vadinamas *subalansuotu*, jei su bet kuriuo faktorių lygmenų rinkiniu atliekamas vienodas eksperimentų skaičius; priešingu atveju eksperimento planas vadinamas *nesubalansuotu*.

2.9. Dvifaktorė dispersinė analizė naudojant hierarchinę klasifikaciją

Smulkiau panagrinėkime hierarchinės klasifikacijos duomenis ir juos atitinkantį matematinį modelį.

2.9.1. Modelis, kai faktoriai pastovūs

Tarkime, a. d. Y skirstinys gali priklausyti nuo faktoriaus A, kuris turi I lygmenų, ir nuo faktoriaus B, turinčio J lygmenų.

Hierarchinės klasifikacijos modelis: kai faktoriaus A lygmuo yra A_1, tai stebėjimai atliekami fiksuojant atsitiktinai parinktus J_1 faktoriaus B lygmenis (žymėsime juos $B_1, \ldots, B_{J_1}$); kai faktoriaus A lygmuo yra A_2 – atsitiktinai iš likusių parinktus J_2 lygmenis $B_{J_1+1}, \ldots, B_{J_1+J_2}$; ir t. t., ir pagaliau, kai faktoriaus A lygmuo yra A_I – $B_{J_1+\ldots+J_{I-1}+1}, \ldots, B_J$, $J = J_1 + \ldots + J_I$.

Faktorius A vadinamas *grupuojančiuoju*, o faktorius B – *sugrupuotu* pagal faktorių A.

Šis hierarchinės klasifikacijos modelis yra atskiras dvifaktorės analizės su pastoviais faktoriais, nagrinėtos 2.3.1 skyrelyje, atvejis, kai stebėjimų skaičius K_{ij} kai kuriuose langeliuose lygūs 0. Nepaisant to, šiuo konkrečiu atveju modelyje (2.3.11) parametrus, apibūdinančius faktorių įtaką, įvesime kitaip, apsiribodami faktoriaus B įtaka, kai faktoriaus A lygmuo fiksuotas.

Lentelėje 2.9.1 nurodyti langeliai, atitinkantys faktorių lygmenų rinkinius, su kuriais atliekami stebėjimai. Kai klasifikacija kryžminė, stebėjimai yra atliekami visuose pateiktos lentelės langeliuose.

2.9.1 lentelė. Turimų stebėjimų pozicijos

$A\ B$	B_1	...	B_{J_1}	B_{J_1+1}	...	$B_{J_1+J_2}$	...	B_{J-J_I+1}	...	B_J
A_1	×	×	×							
A_2				×	×	×				
...	...	...	...	...	...	...	...	...	...	...
A_I								×	×	×

Dvifaktorės dispersinės analizės modelis naudojant hierarchinę klasifikaciją: k-asis stebėjimas, gautas, kai faktoriaus A lygmuo A_i ir faktoriaus B lygmuo $B_{J_1+...+J_{i-1}+j}$ (žr. 2.9.1 lentelę), turi pavidalą

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, ..., I, \quad j = 1, ..., J_i \quad k = 1, ..., K_{ij}; \tag{2.9.1}$$

čia $\mu_{ij} = \mathbf{E}(Y_{ijk})$, o paklaidos $e_{ijk} \sim N(0, \sigma^2)$ yra nepriklausomos.

Nežinomų parametrų μ_{ij} skaičius yra $J = J_1 + ... + J_I$.

Norėdami suformuluoti pagrindines hipotezes, parametrus μ_{ij} suskaidykime į komponentes

$$\mu_{ij} = \mu + \alpha_i + \beta_{(i)j}, \tag{2.9.2}$$

$$\alpha_i = \bar{\mu}_{i.} - \mu, \quad \beta_{(i)j} = \mu_{ij} - \bar{\mu}_{i.},$$

$$\bar{\mu}_{i.} = \frac{1}{K_{i.}} \sum_j K_{ij}\mu_{ij} \quad K_{i.} = \sum_j K_{ij}, \quad \mu = \frac{1}{n} \sum_i K_{i.}\mu_{i.}.$$

Nurodant, kad faktorius B sugrupuotas pagal faktorių A, pastarojo indeksas i yra apskliaustas.

Nauji parametrai tenkina papildomas sąlygas

$$\sum_i K_{i.}\alpha_i = 0, \quad \sum_j K_{ij}\beta_{(i)j} = 0, \quad \forall i = 1, ..., I. \tag{2.9.3}$$

Parametras α_i parodo svertinio požymio Y vidurkio, kai fiksuotas faktoriaus A lygmuo A_i, ir bendro svertinio požymio Y vidurkio skirtumą. Taigi parametrai α_i apibūdina faktoriaus A lygmenų įtaką požymio Y skirstiniui. Analogiškai parametras $\beta_{(i)j}$ rodo faktoriaus B įtaką požymio Y vidurkiui, kai fiksuotas faktoriaus A lygmuo A_i. Kadangi su kiekvienu faktoriaus A lygmeniu turime skirtingą modelį, tai natūralu, kad parametrų, apibūdinančių faktorių sąveiką, modelyje (2.9.1) nėra.

2.9.2. Parametrų įvertiniai ir hipotezių tikrinimas

Parametro μ_{ij} mažiausiųjų kvadratų įvertinys yra gaunamas minimizuojant kvadratų sumą $SS(\boldsymbol{\mu}) = \sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2$:

$$\hat{\mu}_{ij} = \bar{Y}_{ij.} = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} Y_{ijk}, \tag{2.9.4}$$

o liekamoji kvadratinė forma

$$SS_E = \min_{\boldsymbol{\mu}} SS(\boldsymbol{\mu}) = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 \sim \sigma^2 \chi^2_{n-J}, \quad n = \sum_i \sum_j K_{ij}. \tag{2.9.5}$$

Tiesinių funkcijų $\alpha_i = \bar{\mu}_{i.} - \mu$ ir $\beta_{(i)j} = \mu_{ij} - \bar{\mu}_{i.}$ mažiausiųjų kvadratų įvertiniai

$$\hat{\alpha}_i = \frac{1}{K_{i.}} \sum_j K_{ij} \bar{Y}_{ij.} - \frac{1}{n} \sum_i \sum_j K_{ij} \bar{Y}_{ij.}, \quad \hat{\beta}_{(i)j} = \bar{Y}_{ij.} - \frac{1}{K_{i.}} \sum_j K_{ij} \bar{Y}_{ij.}. \tag{2.9.6}$$

tenkina analogiškus (2.9.3) apribojimus

$$\sum_i K_{i.} \hat{\alpha}_i = 0, \quad \sum_j K_{ij} \hat{\beta}_{(i)j} = 0, \quad \forall i = 1, ..., I. \tag{2.9.7}$$

Pagrindinės dispersinės analizės hipotezės yra

$$H_A : \alpha_1 = ... = \alpha_I = 0, \tag{2.9.8}$$

faktoriaus A įtakos nėra, ir hipotezė

$$H_{B(A)} : \beta_{(i)j} = 0, \quad \forall j = 1, ..., J_i, \quad \forall i = 1, ..., I, \tag{2.9.9}$$

kad faktorius B a. d. Y_{ijk} skirstiniams neturi įtakos (kai fiksuotas bet kuris faktoriaus A lygmuo).

Hipotezes tikriname remdamiesi 1.3.2 teorema. Besąlyginis kvadratinės formos $SS(\boldsymbol{\mu})$ minimumas, kurį naudosime Fišerio statistikos vardiklyje, pateiktas (2.9.5) lygybėje. Lieka rasti sąlyginius šios kvadratinės formos minimumus $SS_{EH_{B(A)}}$ ir SS_{EH_A}, kai atitikamai teisingos hipotezės (2.9.8) ir (2.9.9), ir skirtumus $SS_{B(A)} = SS_{EH_{B(A)}} - SS_E$, $SS_A = SS_{EH_A} - SS_E$.

2.9.1 teorema. *Kvadratinės formos SS_A ir $SS_{B(A)}$ turi tokį pavidalą:*

$$SS_A = \sum_i K_{i.} \hat{\alpha}_i^2, \quad SS_{B(A)} = \sum_i \sum_j K_{ij} \hat{\beta}^2_{(i)j}; \tag{2.9.10}$$

$\hat{\alpha}_i$ ir $\hat{\beta}_{(i)j}$ pateikti (2.9.6) formulėje. Šios kvadratų sumos nepriklauso nuo SS_E. Be to,

$$\frac{SS_A}{\sigma^2} \sim \chi^2(I-1; \lambda_A), \quad \frac{SS_{B(A)}}{\sigma^2} \sim \chi^2(J-I; \lambda_{B(A)}). \tag{2.9.11}$$

Necentriškumo parametrai

$$\lambda_A = \frac{1}{\sigma^2} \sum_i K_{i.} \alpha_i^2, \quad \lambda_{B(A)} = \frac{1}{\sigma^2} \sum_i \sum_j K_{ij} \beta^2_{(i)j}.$$

Jeigu hipotezės H_A ar $H_{B(A)}$ yra teisingos, tai atitinkami skirstiniai yra centriniai.

Įrodymas. Tapatybės

$$Y_{ijk} - \mu_{ij} = (Y_{ijk} - \bar{Y}_{ij.}) + (\hat{\mu} - \mu) + (\hat{\alpha}_i - \alpha_i) + (\hat{\beta}_{(i)j} - \beta_{(i)j})$$

abi puses pakėlę kvadratu, susumavę pagal visas galimas indeksų reikšmes ir pasinaudoję sąlygomis (2.9.3), (2.9.7), gauname

$$SS(\boldsymbol{\mu}) = SS_E + n(\hat{\mu}-\mu)^2 + \sum_i K_{i.}(\hat{\alpha}_i-\alpha_i)^2 + \sum_i \sum_j K_{ij}(\hat{\beta}_{(i)j}-\beta_{(i)j})^2. \quad (2.9.12)$$

Iš čia randame

$$SS_{EH_A} = \min_{\alpha_i \equiv 0} SS(\boldsymbol{\mu}) = SS_E + \sum_i K_{i.}\hat{\alpha}_i^2, \quad SS_A = SS_{EH_A} - SS_E = \sum_i K_{i.}\hat{\alpha}_i^2;$$

$$SS_{EH_{B(A)}} = \min_{\beta_{(i)j} \equiv 0} SS(\boldsymbol{\mu}) = SS_E + \sum_i \sum_j K_{ij}\hat{\beta}_{(i)j}^2,$$

$$SS_{B(A)} = SS_{EH_{B(A)}} - SS_E = \sum_i \sum_j K_{ij}\hat{\beta}_{(i)j}^2.$$

Kiti teoremos tvirtinimai tiesiogiai išplaukia iš 1.3.2 teoremos. ▲

Pažymėję

$$MS_A = \frac{SS_A}{I-1}, \quad MS_{B(A)} = \frac{SS_{B(A)}}{J-I},$$

gauname

$$\mathbf{E}(MS_A) = \sigma^2 + \frac{\sigma^2\lambda_A}{I-1}, \quad \mathbf{E}(MS_{B(A)}) = \sigma^2 + \frac{\sigma^2\lambda_{B(A)}}{J-I}.$$

Skaičiavimo rezultatus surašome į dispersinės analizės lentelę.

2.9.2 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + \frac{\sigma^2\lambda_A}{I-1}$
B (sugrupuotas pagal A	$SS_{B(A)}$	$J-I$	$MS_{B(A)}$	$\sigma^2 + \frac{\sigma^2\lambda_{B(A)}}{J-I}$
E	SS_E	$n-m$	MS_E	σ^2

Grįžtame prie hipotezių H_A ir $H_{B(A)}$ tikrinimo. Sudarome statistikas

$$F_A = \frac{MS_A}{MS_E}, \quad F_{B(A)} = \frac{MS_{B(A)}}{MS_E}.$$

Hipotezes H_A ir $H_{B(A)}$ atmetame reikšmingumo lygmens α kriterijais, kai atitinkamai teisingos nelygybės

$$F_A > F_\alpha(I-1, n-J), \quad F_{B(A)} > F_\alpha(J-I, n-J). \quad (2.9.13)$$

Kriterijų galia išreiškiama necentrinio Fišerio skirstinio pasiskirstymo funkcija.

2.9.1 pastaba. Jeigu 2.9.1 lentelėje stebėjimų kartotinumai $K_{ij} = 1$, tai kvadratų suma $SS_E = 0$ ir dispersijos σ^2 įvertinys neapibrėžtas. Norėdami sumažinti nežinomų parametrų skaičių, tarkime, kad hipotezė $H_{B(A)}$ teisinga. Tada 2.9.1 lentelę galima interpretuoti kaip vienfaktorės analizės 2.1.1 lentelę.

2.9.3. Modelis, kai faktoriai atsitiktiniai

Dažniausiai aptinkamas atvejis, kai grupuojamasis faktorius B yra atsitiktinis. Pavyzdžiui, 2.9.2 pavyzdyje natūralu tarti, kad pagal kiekvieną technologiją reikia palyginti ne kelis eksperimente dalyvaujančius gaminius, o visus pagal šią technologiją pagamintus gaminius. Taigi faktorių B reikėtų laikyti atsitiktiniu. Grupuojantysis faktorius A gali būti arba pastovus, arba atsitiktinis. Pirmu atveju stebėjimų matematinis modelis yra

$$Y_{ijk} = \mu + \alpha_i + b_{(i)j} + e_{ijk}, \tag{2.9.14}$$

o antruoju

$$Y_{ijk} = \mu + a_i + b_{(i)j} + e_{ijk}. \tag{2.9.15}$$

Parametrai α_i tenkina papildomą sąlygą $\alpha_1 + ... + \alpha_I = 0$. Tarsime, kad a. d. a_i ir $b_{(i)j}$ yra nepriklausomi tarpusavyje, nepriklauso nuo paklaidų $e_{ijk} \sim N(0, \sigma^2)$ ir yra vienodai pasiskirstę pagal normalųjį dėsnį $a_i \sim N(0, \sigma_A^2)$, $b_{(i)j} \sim N(0, \sigma_{B(A)}^2)$. Pasirinkę $J_1 = ... = J_I = J$ ir $K_{ij} = K > 1$ gausime, kad paskutiniame 2.9.2 lentelės stulpelyje abiem atvejais

$$\mathbf{E}(MS_{B(A)}) = \sigma^2 + K\sigma_{B(A)}^2.$$

Pirmoje eilutėje vidurkis bus skirtingas:

$$\mathbf{E}(MS_A) = \sigma^2 + K\sigma_{B(A)}^2 + \frac{JK}{I-1}\sum_i \alpha_i^2,$$

kai faktorius A pastovus, ir

$$\mathbf{E}(MS_A) = \sigma^2 + K\sigma_{B(A)}^2 + JK\sigma_A^2,$$

kai faktorius A atsitiktinis.

Hipotezės $H_{B(A)}$ analogas yra tvirtinimas, kad $\sigma_{B(A)}^2 = 0$. Ši prielaida yra atmetama, kai

$$F_{B(A)} = \frac{MS_{B(A)}}{MS_E} > F_\alpha(I(J-1), IJ(K-1)). \tag{2.9.16}$$

Tikrinant hipotezę $H_A : \alpha_1 = ... = \alpha_I = 0$ arba jos analogą $H_A : \sigma_A^2 = 0$, ji atmetama, kai

$$F_A = \frac{MS_A}{MS_{B(A)}} > F_\alpha(I-1, I(J-1)). \tag{2.9.17}$$

2.9.2 pastaba. Jeigu faktorių skaičius didesnis už du, tai galimi įvairūs atvejai. Vieni faktoriai gali dalyvauti eksperimente pagal kryžminės klasifikacijos schemą, o kiti faktoriai sugrupuoti pagal tam tikrus faktorius, arba jų lygmenų rinkinius.

Pavyzdžiui, tarkime, kad tiriamas oro užterštumas kelių valstybių miestų įvairiose zonose (centras, gyvenamieji kvartalai, pramoninės zonos, poilsio zonos ir pan.). Šioje trifaktorės analizės schemoje faktorius A (valstybė) yra grupuojantysis. Pagal jį sugrupuotas faktorius B (miestai). Savo ruožtu faktorius C (miesto zonos) sugrupuotas pagal faktorių B. Jeigu tiriame oro užterštumą, pavyzdžiui, miesto centre įvairiuose aukščiuose, tai faktorius A (valstybė) yra grupuojantysis, o faktorius B (miestas) ir faktorius C (aukštis) gali dalyvauti eksperimente pagal kryžminės klasifikacijos schemą.

Dispersinės analizės lentelė sudaroma ir kriterijai hipotezėms tikrinti kuriami anlogiškai išnagrinėtiems atvejams.

2.9.1 pavyzdys. Tiriamas kalcio kiekis ropių lapuose (žr.[13]). Atsitiktinai atrinkti keturi augalai (faktorius A) ir kiekvieno jų atsitiktinai atrinkta po tris lapus (faktorius B). Iš kiekvieno lapo paimta po du mėginius ir nustatytas kalcio kiekis (tiriamas požymis Y). Stebėjimo rezultatai pateikti 2.9.3 lentelėje.

2.9.3 lentelė. Statistiniai duomenys

	B_1	B_2	B_3
A_1	3,28; 3,09	3,52; 3,48;	2,88; 2,80
A_2	2,46; 2,44	1,87; 1,92	2,19; 2,19
A_3	2,77; 2,66	3,74; 3,44	2,55; 2,55
A_4	3,78; 3,87	4,07; 4,12	3,31; 3,31

Šiame pavyzdyje faktorius B (lapas) sugrupuotas pagal faktorių A (augalas). Natūralu tarti, kad abu faktoriai atsitiktiniai. Atlikę skaičiavimus gauname $MS_E = 0,006654$, $MS_{B(A)} = 0,328775$, $MS_A = 2,52011$; laisvės laipsnių skaičius atitinkamai yra 12, 8, 3. Hipotezėms $H_{B(A)} : \sigma^2_{B(A)} = 0$ ir $H_A : \sigma^2_A = 0$ tikrinti gauname statistikų realizacijas $F_{B(A)} = MS_{B(A)}/MS_E = 49,41$, $F_A = MS_A/MS_{B(A)} = 7,67$. Kadangi P reikšmės $\mathbf{P}\{F_{8,12} > 49,41\} = 5,1 \cdot 10^{-8}$, $\mathbf{P}\{F_{3,8} > 7,67\} = 0,0097$ yra mažos, tai abi hipotezės atmetamos. Taškiniai parametrų įverčiai yra $\hat{\sigma}^2 = 0,0067$, $\hat{\sigma}^2_{B(A)} = 0,1611$, $\hat{\sigma}^2_A = 0,3652$, $\hat{\mu} = 3,0121$.

2.10. Blokuotųjų duomenų dispersinė analizė

2.10.1. Vienfaktorė blokuotųjų duomenų dispersinė analizė

2.10.1.1 Statistinis modelis

Tarkime, kad tiriama vieno faktoriaus A, turinčio I lygmenų, įtaka nagrinėjamam požymiui Y. Jei eksperimente dalyvauja $J_1 + \cdots + J_I$ nepriklausomų objektų ir su kiekvienu fiksuotu faktoriaus A lygmeniu A_i matuojamos skirtingų J_i objektų požymio Y reikšmės, tai atliekama vienfaktorė dispersinė analizė, nagrinėta ankstesniuose skyreliuose. Pagrindinis tokios analizės uždavinys – palyginti požymio Y vidurkius, atitinkančius skirtingus faktoriaus lygmenis.

Jei atsitiktinai parenkama n objektų ir kiekvieno iš jų požymio Y reikšmės matuojamos I kartų, kai faktoriaus reikšmės perbėga visas I galimas reikšmes, tai atliekama vienfaktorė blokuotųjų duomenų dispersinė analizė. Uždavinys tas pats – palyginti požymio Y vidurkius, atitinkančius skirtingus faktoriaus lygmenis.

Pagal pirmą planą su bet kokia fiksuota faktoriaus reikšme gaunami nepriklausomi požymio stebėjimai, pagal antrą planą jie priklausomi.

Vienfaktorė blokuotųjų duomenų dispersinė analizė naudojama ir šiek tiek bendresnėms situacijoms (žr. 2.8 skyrelį), kai duomenys suskirstomi į blokus, vienam blokui priskiriant I atsitiktinai parinktų pagal kokį tai požymį homogeniškų objektų. Bloko i-asis objektas stebimas, kai fiksuota i-ji faktoriaus reikšmė. Jei tai įmanoma, faktorių išdėstymo bloke tvarka – atsitiktinė.

Pavyzdžiui, 2.9.2 pavyzdyje, tiriant skirtingų padangų markių protektoriaus dėvėjimosi greitį, blokus sudaro $I = 4$ padangos, sumontuotos ant vieno automobilio. Tiriant tam tikro gydymo metodo poveikį pacientų kraujo spaudimui, atliekami kraujo spaudimo matavimai prieš gydymą, po gydymo, praėjus pusmečiui ir metams po gydymo. Blokus sudaro $I = 4$ kraujo spaudimo matavimai, atlikti tam pačiam pacientui. Trimis skirtingais būdais tiriant bulvių krakmolingumą, kiekvienas būdas naudojamas tai pačiai bulvei. Blokus sudaro $I = 3$ atlikti tos pačios bulvės krakmolingumo matavimai.

Šiuose pavyzdžiuose norime ištirti matuojamo kintamojo priklausomybę nuo vieno faktoriaus A (padangos dėvėjimosi greičio priklausomybę nuo jos markės; kraujo spaudimo priklausomybę nuo gydymo metodo; krakmolingumo nustatymo metodų skirtingumą).

Bloko parinkimą galima traktuoti kaip papildomo faktoriaus B, pagal kurio reikšmes sudaromi blokai, įtraukimą. Galima tarti, kad minėtuose pavyzdžiuose faktorius B (automobilis, pacientas, bulvė) atsitiktinai atrenkamas iš pakankamai didelės generalinės aibės, t. y. gali būti traktuojamas kaip *atsitiktinis*.

Pereiname prie mišraus dvifaktorės dispersinės analizės modelio su $K = 1$ stebėjimu langelyje, kuriame pagrindinis yra faktorius A (tiriama būtent jo įtaka nagrinėjamam požymiui), antrasis atsitiktinis faktorius B (apibūdinantis bloką) yra trukdantysis ir jo įtaką reikia pašalinti.

Faktorių išdėstymo bloke tvarka imama atsitiktinė, kad būtų galima pašalinti tvarkos įtaką. Pavyzdžiui, padangų, uždėtų ant to paties automobilio, dėvėjimasis gali priklausyti nuo padangos pozicijos. Lyginant skirtingus psichologinius testus, to paties asmens rezultatai gali priklausyti nuo testų atlikimo tvarkos (nuovargio įtaka), ne tik nuo asmens gabumų. Kartais randomizacija bloko viduje negalima. Pavyzdžiui, kraujo spaudimo matavimų pavyzdyje negalima keisti kraujo spaudimo matavimų tvarkos.

Tarsime, kad nėra faktoriaus A ir trukdančiojo faktoriaus B sąveikos. Taigi naudojame mišraus dvifaktorės dispersinės analizės adityvųjį modelį, kai $K = 1$ (žr. 2.6 skyrelį):

$$Y_{ij} = \mu + \alpha_i + b_j + e_{ij}; \tag{2.10.1}$$

čia i – faktoriaus A lygmens numeris, $i = 1, ..., I$; indeksas j – faktoriaus B (bloko) numeris, $j = 1, ..., J$; parametrai α_i tenkina sąlygą $\sum_i \alpha_i = 0$; visi

atsitiktiniai dydžiai $b_j \sim N(0, \sigma_B^2)$, $e_{ij} \sim N(0, \sigma^2)$ nepriklausomi.

Pagrindinis dispersinės analizės uždavinys – patikrinti hipotezę $H_A : \alpha_1 = ... = \alpha_I = 0$, kad faktoriaus A įtakos nėra.

Eksperimentų duomenis (2.10.1) galime interpretuoti ir kaip atsitiktinio vektoriaus $(Y_1, ..., Y_I)^T \sim N_I(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ paprastąją atsitiktinę imtį

$$(Y_{11,}..., Y_{I1})^T, \quad ... \quad (Y_{1J,}..., Y_{IJ})^T,$$

kurios didumas lygus blokų skaičiui J.

Vidurkių vektorius yra

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_I)^T = (\mu + \alpha_1, ..., \mu + \alpha_I)^T,$$

o kovariacijų matrica

$$\boldsymbol{\Sigma} = [\sigma_{ii'}]_{I \times I}, \quad \sigma_{ii} = \boldsymbol{V} Y_i = \sigma_B^2 + \sigma^2, \quad \sigma_{ii'} = \mathbf{Cov}\,(Y_i, Y_{i'}) = \sigma_B^2, \quad i \neq i'.$$

Taigi visos kovariacijos vienodos, o skirtumai $Y_i - Y_{i'}$, $i \neq i'$ vienodai pasiskirstę: $Y_i - Y_{i'} \sim N(\alpha_i - \alpha_{i'}, 2\sigma^2)$.

Hipotezės H_A ekvivalenti forma: $H_A : \alpha_1 = ... = \alpha_I = 0$, t. y. ši hipotezė yra normalaus atsitiktinio vektoriaus $(Y_1, ..., Y_I)^T$ komponenčių vidurkių lygybės hipotezė.

Primename, kad $I = 2$ atveju, kai stebimas dvimatis normalusis atsitiktinis vektorius su priklausomomis koordinatėmis, vidurkiams palyginti naudojamas Stjudento kriterijus priklausomoms imtims (žr. 1 dalį, 3.7.3 skyrelį). Taigi toliau pateiktą kriterijų hipotezei H_A tikrinti galima interpretuoti kaip Stjudento priklausomų imčių kriterijaus apibendrinimą atvejui, kai stebimo vektoriaus dimensija didesnė už du.

Neparametriniu atveju, kai nežinomas a. v. $(Y_1, ..., Y_I)^T$ skirstinys, hipotezę H_A galima tikrinti naudojant Frydmano kriterijų.

2.10.1.2 Faktorių įtakos nebuvimo hipotezės tikrinimas

Naudosime 2.6 skyrelio pažymėjimus (imame $K = 1$):

$$SS_A = J \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2, \quad SS_B = I \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2,$$

$$SS_{AB} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2,$$

$$MS_A = \frac{SS_A}{I-1}, \quad MS_B = \frac{SS_B}{J-1}, \quad MS_{AB} = \frac{SS_{AB}}{(I-1)(J-1)}.$$

Dispersinės analizės lentelė gaunama iš 2.6.1 lentelės atlikus tokius pakeitimus: vietoje σ_{AB}^2 įrašomas nulis; vietoje K įrašomas 1 ir praleidžiamas taškas, atitinkantis indeksą k; praleidžiama eilutė, atitinkanti faktorių E.

2.10.1 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	$MS = SS/\nu$	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + J\sigma_A^2$
B	SS_B	$J-1$	MS_B	$\sigma^2 + I\sigma_B^2$
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}	σ^2
T	SS_T	I J -1	-	-

Hipotezė $H_A : \alpha_1 = ... = \alpha_I = 0$ yra atmetama (žr. kriterijų (2.6.8)), kai

$$F_A = \frac{MS_A}{MS_{AB}} > F_\alpha(I-1, (I-1)(J-1)). \tag{2.10.2}$$

Norint patikrinti, ar apskritai buvo verta duomenis grupuoti į blokus, kartais patikrinama bloko įtakos nebuvimo hipotezė $H_B : \sigma_B^2 = 0$. Ji atmetama, kai

$$F_B = \frac{MS_B}{MS_{AB}} > F_\alpha(J-1, (I-1)(J-1)). \tag{2.10.3}$$

Šis kriterijus skiriasi nuo (2.6.6) tuo, kad statistikos vardiklyje yra MS_{AB} (vietoje MS_E).

2.10.1 pastaba. Jeigu kiekviename bloke kiekvienam faktoriaus A lygmeniui yra gaunami pakartotiniai matavimai, tai turime pilną mišrios dvifaktorės analizės planą. Tada galima įvertinti ir faktorių sąveiką bei patikrinti sąveikos nebuvimo hipotezę. Iš 2.6 skyrelio analizės išplaukia, kad F kriterijus hipotezei $H_A : \alpha_1 = ... = \alpha_I = 0$ tikrinti, kai $\sigma_{AB}^2 \neq 0$, yra apytikslis. Jeigu sąveikos nebuvimo hipotezė yra atmetama, tai, tikrinant hipotezę H_A, kai kuriuose matematinės statistikos TPP yra numatytos pataisos (Grynhauzo ir Geiserio, Hjuino ir Feldto, arba apatinio rėžio; žr.[13]).

2.10.1 pavyzdys. [14] Tiriama avižų derlingumo (kintamasis Y) priklausomybė nuo jų veislės (faktorius A). Atsitiktinai parenkamos 5 vietovės (faktorius B) ir kiekvienoje iš jų auginamos visos 8 avižų veislės vienodo didumo sklypeliuose. Šiame pavyzdyje faktorius B (blokas) yra trukdantysis ir jo įtaką pageidautina eliminuoti. Derlingumo duomenys pateikti 2.10.2 lentelėje.

2.10.2 lentelė. Statistiniai duomenys

	B_1	B_2	B_3	B_4	B_5
A_1	296	357	340	331	348
A_2	202	390	431	340	320
A_3	437	334	426	320	296
A_4	303	319	310	260	242
A_5	469	405	442	487	394
A_6	345	342	358	300	308
A_7	324	339	357	352	220
A_8	488	274	401	338	320

Atlikę skaičiavimus gauname $MS_A = 10038$, $MS_B = 6292,625$, $MS_{AB} = 2769,225$. Hipotezei H_A tikrinti (eliminuojant bloko įtaką) apskaičiuojame santykį $F_A = MS_A/MS_{AB} = 3,62$. Kadangi P reikšmė $\mathbf{P}\{F_{7,28} > 3,62\} = 0,0066$, tai hipotezė H_A atmetama. Norėdami atsakyti į klausimą, ar vertėjo skirstyti duomenis į blokus, tikriname hipotezę $H_B : \sigma_B^2 = 0$.

Gauname $F_B = MS_B/MS_{AB} = 2,27$. Kadangi $\mathbf{P}\{F_{4,28} > 2,27\} = 0,087$, tai bloko įtaka nėra didelė ir galbūt skirstyti duomenis į blokus nebuvo būtina.

2.10.2. Dvifaktorė blokuotųjų duomenų dispersinė analizė

Dažnai tenka spręsti sudėtingesnius uždavinius, kai tiriama požymio Y priklausomybė nuo keleto faktorių esant vienam ar keletui trukdančiųjų faktorių, kurių įtaką reikia pašalinti. Pateiksime porą pavyzdžių.

2.10.2 pavyzdys. Tiriama kviečių derlingumo (požymis Y) priklausomybė nuo jų veislės (faktorius A, kurio lygmenys $A_1, ..., A_I$) ir nuo auginimo metodikos (faktorius B, kurio lygmenys $B_1, ..., B_J$). Derlingumas gali priklausyti ir nuo kitų faktorių: dirvožemio savybių, klimato sąlygų ir pan. Siekiant pašalinti jų įtaką, įvairiose šalies vietovėse atsitiktinai parenkama L žemės sklypų. Kiekvienas sklypas padalijamas į IJ sklypelių, kurie apsėjami skirtingomis kviečių veislėmis naudojant skirtingas auginimo metodikas pagal kryžminės klasifikacijos schemą. Tam, kad būtų pašalinta kitų neįskaitytų faktorių įtaka, faktorių lygmenų rinkiniai parenkami atsitiktinai iš $I!J!$ galimų variantų. Taigi nuo dvifaktorės dispersinės analizės dirbtinai pereiname prie trifaktorės, kurioje faktorius C (vietovė) yra trukdantysis ir reikia pašalinti jo įtaką. Natūralu tarti, kad faktoriai A ir B yra pastovūs, o faktorius C – atsitiktinis.

2.10.3 pavyzdys. Tiriama produkcijos išeigos (požymis Y) priklausomybė nuo 4 staklių tipų (faktorius A) ir nuo darbininko (faktorius B). Tačiau per vieną pamainą galima atlikti tik 16 eksperimentų: 4 darbininkai, dirba su kiekvienomis iš 4 staklių 1/4 pamainos pagal kryžminės klasifikacijos schemą. Todėl toks eksperimentas kartojamas kelių atsitiktinai parinktų pamainų metu. Ir šiame pavyzdyje nuo dvifaktorės dispersinės analizės pereiname prie trifaktorės pridedant trečią trukdantįjį faktorių C – pamainą. Šiame pavyzdyje faktorius C atsitiktinis, faktorius A pastovus, o faktorius B gali būti pastovus (domina tik 4 eksperimente dalyvaujančių darbininkų įtaka), arba atsitiktinis (darbininkai eksperimentui parenkami atsitiktinai iš visos jų aibės ir mus domina apskritai darbininko įtaka požymiui Y).

Nagrinėsime dvifaktorės dispersinės analizės schemą, kai yra du pastovūs faktoriai A ir B, o jų lygmenys $A_1, ..., A_I$ ir $B_1, ..., B_J$. Pridedamas trečias trukdantysis atsitiktinis faktorius C (blokas), kurio lygmenys $C_1, ..., C_L$. Kai lygmuo C_l fiksuotas, eksperimentas atliekamas pagal kryžminės dvifaktorės dispersinės analizės klasifikacijos,kai yra vienas stebėjimas langelyje, schemą.

Reikia pažymėti, kad turime mišrųjį trifaktorės dispersinės analizės modelį, kai yra du pastovūs faktoriai A ir B bei vienas atsitiktinis faktorius C. Požymio Y stebėjimus numeruosime trimis indeksais: Y_{ijl}, $i = 1, ..., I$, $j = 1, ..., J$, $l = 1, ..., L$; čia i – faktoriaus A lygmens numeris, j – faktoriaus B lygmens numeris, l – faktoriaus C (bloko) lygmens numeris.

Naudodami 2.7 skyrelio žymenis, gauname modelį

$$Y_{ijl} = \mu + \alpha_i^A + \alpha_j^B + a_l^C + \alpha_{ij}^{AB} + a_{il}^{AC} + a_{jl}^{BC} + a_{ijl}^{ABC} + e_{ijl}, \qquad (2.10.4)$$

kuriame raidėmis μ, α žymimos konstantos, o raidėmis a, e – atsitiktiniai dydžiai. Tariama, kad a. d. $e_{ijl} \sim N(0, \sigma^2)$ yra nepriklausomi tarpusavyje ir nepriklauso nuo a. d., žymimų raidėmis a. Atsitiktiniai dydžiai, žymimi raidėmis a, nor-

malieji su nuliniais vidurkiais ir dispersijomis, kurios gali priklausyti nuo a. d. apibrėžime esančių pastovių faktorių indeksų.

Standartinių kvadratų sumų, atitinkančių dėstinio (2.10.4) narius (išskyrus μ ir e_{ijl}), išraiškos pateiktos 2.7 skyrelio iliustraciniame pavyzdyje. Tereikia atlikti tokius pakeitimus: įrašyti vietoje K vienetą ir praleisti tašką, atitinkantį indeksą k.

Vidutinių kvadratų sumų MS vidurkiai $\mathbf{E}(MS)$ yra tiesinės funkcijos parametro σ^2 (su koeficientu 1) ir parametrų (žr. 2.7 skyrelį)

$$\sigma_C^2 = \boldsymbol{V}(a_l^C), \quad \sigma_{AC}^2 = \frac{1}{I-1}\sum_i \boldsymbol{V}(a_{il}^{AC}), \quad \sigma_{BC}^2 = \frac{1}{J-1}\sum_j \boldsymbol{V}(a_{jl}^{BC}),$$

$$\sigma_{ABC}^2 = \frac{1}{(I-1)(J-1)}\sum_i\sum_j \boldsymbol{V}(a_{ijl}^{ABC}) \tag{2.10.5}$$

bei parametrų

$$\sigma_A^2 = \frac{1}{I-1}\sum_i (\alpha_i^A)^2, \quad \sigma_B^2 = \frac{1}{J-1}\sum_j (\alpha_j^B)^2,$$

$$\sigma_{AB}^2 = \frac{1}{(I-1)(J-1)}\sum_i\sum_j (a_{ij}^{AB})^2. \tag{2.10.6}$$

Kadangi $SS_E = 0$, tai nėra galimybės įvertinti parametrą σ^2. Todėl tenka nagrinėti siauresnį modelį, turintį mažesnį parametrų skaičių. Natūralu pasirinkti prielaidą, kad parametras σ_{ABC}^2, apibūdinantis visų trijų faktorių sąveiką, lygus nuliui.

Dispersinės analizės lentelė gaunama iš lentelės 2.7.2, praleidus eilutę, atitinkančią faktorių E, įrašius 0 vietoje σ_{ABC}^2 ir vienetą vietoje K, bei praleidus tašką, atitinkantį indeksą k.

2.10.3 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}(MS)$
A	SS_A	$I-1$	MS_A	$\sigma^2 + J\sigma_{AC}^2 + JL\sigma_A^2$
B	SS_B	$J-1$	MS_B	$\sigma^2 + I\sigma_{BC}^2 + IL\sigma_B^2$
C	SS_C	$L-1$	MS_C	$\sigma^2 + IJ\sigma_C^2$
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}	$\sigma^2 + L\sigma_{AB}^2$
$A \times C$	SS_{AC}	$(I-1)(L-1)$	MS_{AC}	$\sigma^2 + J\sigma_{AC}^2$
$B \times C$	SS_{BC}	$(J-1)(L-1)$	MS_{BC}	$\sigma^2 + I\sigma_{BC}^2$
$A \times B \times C$	SS_{ABC}	$(I-1)(J-1)(L-1)$	MS_{ABC}	σ^2
T	SS_T	$IJLK-1$		

Visų pirma patikrinsime duomenų skirstymo į blokus tikslingumo hipotezę. Jeigu hipotezės $H_{AC} : \sigma_{AC}^2 = 0$, $H_{BC} : \sigma_{BC}^2 = 0$, apibūdinančios faktorių ir bloko sąveiką, yra teisingos, tai atitinkamos kvadratų sumos išreiškiamos paklaidų e_{ijl} terminais:

$$SS_{AC} = J\sum_i\sum_l (\bar{e}_{i.l} - \bar{e}_{i..} - \bar{e}_{..l} + \bar{e}_{...})^2 \sim \sigma^2\chi^2_{(I-1)(L-1)},$$

$$SS_{BC} = I\sum_j\sum_l(\bar{e}_{.jl} - \bar{e}_{.j.} - \bar{e}_{..l} + \bar{e}_{...})^2 \sim \sigma^2\chi^2_{(J-1)(L-1)}.$$

Hipotezės H_{AC} arba H_{BC} atmetamos α lygmens kriterijais, kai atitinkamai teisingos nelygybės

$$F_{AC} = \frac{MS_{AC}}{MS_{ABC}} > F_\alpha((I-1)(L-1),(I-1)(J-1)(L-1)),$$

$$F_{BC} = \frac{MS_{BC}}{MS_{ABC}} > F_\alpha((J-1)(L-1),(I-1)(J-1)(L-1)). \qquad (2.10.7)$$

Pažymėjus $Z_l = a_l^C + \bar{e}_{..l} \sim N(0,\sigma_C^2 + \sigma^2/(IJ))$, kvadratų suma SS_C, apibūdinanti bloko įtaką, turi tokį pavidalą:

$$SS_C = IJ\sum_l(Z_l - \bar{Z}_.)^2 \sim (\sigma^2 + IJ\sigma_C^2)\chi^2_{I-1}.$$

Hipotezė H_C atmetama α lygmens kriterijumi, kai teisinga nelygybė

$$F_C = \frac{MS_C}{MS_{ABC}} > F_\alpha(L-1,(I-1)(J-1)(L-1)). \qquad (2.10.8)$$

Remiantis 2.10.3 lentelės paskutiniuoju stulpeliu, nagrinėjamų faktorių sąveikos hipotezė $H_{AB} : \sigma^2_{AB} = 0$ atmetama α lygmens kriterijumi, kai teisinga nelygybė

$$F_{AB} = \frac{MS_{AB}}{MS_{ABC}} > F_\alpha((I-1)(J-1),(I-1)(J-1)(L-1)). \qquad (2.10.9)$$

Jeigu hipotezės H_{AC} ir H_{BC} atmetamos, tai iš 2.10.3 lentelės matome, kad hipotezių $H_A : \sigma_A^2 = 0$ ir $H_B : \sigma_B^2 = 0$ tikrinimo statistikų vardikliuose reikia imti MS_{AC} ir MS_{BC}. Hipotezės H_A ir H_B, pašalinus bloko ir bloko sąveikos su faktoriais įtaką, atmetamos α lygmens kriterijais, kai teisingos nelygybės

$$F_A = \frac{MS_A}{MS_{AC}} > F_\alpha(I-1,(I-1)(L-1)), \quad F_B = \frac{MS_B}{MS_{BC}} > F_\alpha(J-1,(J-1)(L-1)). \qquad (2.10.10)$$

Primename (žr. 2.6 skyrelį), kad šie kriterijai yra apytikslieji.

Jeigu hipotezės H_{AC} ir H_{BC} dėl bloko ir faktorių sąveikos neatmetamos (kartais tokią prielaidą galima padaryti iš anksto remiantis eksperimento sąlygomis), tai statistikos vardikliu reikėtų imti

$$MS_E^* = \frac{SS_E^*}{(IJ-1)(L-1)}, \quad SS_E^* = SS_{AC} + SS_{BC} + SS_{ABC} =$$

$$\sum_i\sum_j\sum_l Y^2_{ijl} - JL\sum_i\bar{Y}^2_{i..} - IL\sum_j\bar{Y}^2_{.j.} - IJ\sum_l\bar{Y}^2_{..l} + 2\bar{Y}^2_{...} \sim \sigma^2\chi^2_{(IJ-1)(L-1)}.$$

Hipotezės H_A, H_B ir H_{AB} atmetamos, pašalinus adityvią bloko įtaką, α lygmens kriterijais, kai atitinkamai teisingos nelygybės

$$F_A^* = \frac{MS_A}{MS_E^*} > F_\alpha(I-1,(IJ-1)(L-1)), \quad F_B^* = \frac{MS_B}{MS_E^*} > F_\alpha(J-1,(IJ-1)(L-1)),$$

$$F^*_{AB} = \frac{MS_{AB}}{MS^*_E} > F_\alpha((I-1)(J-1),(IJ-1)(L-1)). \qquad (2.10.11)$$

Pagaliau, jeigu ne tik $\sigma^2_{AC} = 0$, $\sigma^2_{BC} = 0$, bet ir adityvioji bloko įtaka $\sigma^2_C = 0$, tai turime dvifaktorės dispersinės analizės modelį, aprašytą 3.2.1 skyrelyje, kuriame kartotinumų skaičius kiekviename lngelyje yra $K = L$. Kriterijai hipotezėms H_A, H_B, H_{AB} tikrinti pateikti (2.2.10) formulėmis.

2.10.2 pastaba. Galimos ir kitos dvifaktorės blokuotųjų duomenų eksperimentų schemos. Pavyzdžiui, abu faktoriai A ir B gali būti atsitiktiniai arba vienas iš jų atsitiktinis, o kitas fiksuotas. Be to, kiekviename bloke eksperimentai gali būti atliekami pagal hierarchinės klasifikacijos schemą. Kriterijai faktorių įtakai apibūdinti eliminuojant bloko įtaką randami anlogiškai išnagrinėtam pavyzdžiui. Panašiai analizuojami blokuotųjų eksperimentų duomenys, kai faktorių skaičius didesnis už du.

2.10.3 pastaba. Blokų parinkimas taip pat gali būti nulemiamas ne vieno, o keleto faktorių, kurie eksperimente dalyvauja pagal kryžminės ar hierarchinės klasifikacijos schemas. Pavyzdžiui, 2.10.2 pavyzdyje galėjo dominti produkcijos išeigos priklausomybė tik nuo staklių tipo, o faktorius B (darbininkas) ir faktorius C (pamaina) yra trukdantieji, kurių įtaką reikia pašalinti. Blokų skaičius yra JL (J – dalyvaujančių eksperimente darbininkų skaičius, L – pamainų skaičius). Kiekviename bloke atliekama po I stebėjimų (I – staklių tipų skaičius); bendras stebėjimų skaičius $n = IJL$. Tokiame eksperimente dauguma stebėjimo rezultatų panaudojama blokų įtakai pašalintiti. Jeigu faktoriaus A ir faktorių B ir C lygmenų skaičius yra vienodas $I = J = L = m$, tai, priėmus adityvumo sąlygas ir išdėsčius eksperimentus pagal lotyniškųjų kvadratų schemą, blokų skaičių galima gerokai sumažinti. Aptariamame pavyzdyje vietoje $n = m^3$, reikėtų atlikti tik $n = m^2$ stebėjimų (detaliau, žr. 2.11 skyrelį).

2.10.4 pastaba. Apskritai kalbant, jeigu bendras faktorių skaičius k yra didelis, tai reikalingų atlikti eksperimentų skaičius tampa didele problema. Pavyzdžiui, jeigu faktorių skaičius $k = 10$, tai netgi parinkus minimalų galimą visų faktorių lygmenų skaičių, lygų 2, kryžminės klasifikacijos schemoje reikėtų atlikti $n = 2^{10} = 1024$ stebėjimus, o tai realizuoti praktiškai gali pasirodyti neįmanoma. Priėmus adityvumo prielaidas ir išdėsčius eksperimentus specialiu būdu į faktorių lygmenų rinkinius, stebėjimų skaičių galima sumažinti iki skaičiaus, nedaug viršijančio faktorių skaičių k (detaliau žr. 4.5 skyrelį).

2.10.3. Nepilni subalansuoti blokai

2.10.3.1 Statistinis modelis

Nagrinėdami atsitiktinius blokus tarėme, kad bloko didumas sutampa su faktoriaus A lygmenų skaičiumi I. Kartais tenka nagrinėti atvejį, kai bloko didumas yra mažesnis už faktoriaus A lygmenų skaičių.

Pavyzdžiui, tiriant septynių padangų markių dilimo greitį, negalima ant vieno automobilio sumontuoti visų septynių markių padangas. Bloko didumą

sudaro keturios padangos, sumontuotos ant vieno automobilio.

Kai bloko didumas k yra mažesnis už faktoriaus A lygmenų skaičių I, tai blokas vadinamas *nepilnu*.

Nagrinėsime konkretų nepilnų blokų sudarymo planą, kuris leidžia surasti paprastus kriterijus pagrindinėms dispersinės analizės hipotezėms tikrinti. Tarsime, kad yra J blokų, visų blokų dydžiai vienodi ir lygūs k, $k < I$, visi faktoriaus A lygmenys bloke skirtingi, o kiekvienas lygmuo priklauso r blokų.

Tada bendras stebėjimų skaičius

$$n = Jk = Ir. \tag{2.10.12}$$

Be to, tarsime, kad naudojamas *subalansuotas nepilnų blokų planas*: bet kuri dviejų faktoriaus A lygmenų pora aptinkama tame pačiame skaičiuje s blokų. Pavyzdžiui, tiriant 7 markių padangų dėvėjimąsi, jos gali būti sumontuotos ant 7 automobilių 2.10.4 lentelėje nurodytu būdu.

2.10.4 lentelė. Subalansuotų blokų eksperimentų planas

III	I	I	I	II	I	II
V	IV	II	II	III	III	IV
VI	VI	V	III	IV	IV	V
VII	VII	VII	VI	VII	V	VI

Šiuo atveju blokų skaičius $J = 7$, bloko didumas $k = 4$, faktoriaus A lygmenų skaičius $I = 7$, kiekvienas lygmuo pasitaiko $r = 4$ blokuose. Pateiktas eksperimentų planas subalansuotas, nes kiekviena faktoriaus A lygmenų pora aptinkama vienodą skaičių $s = 2$.

Parametras s tenkina sąryšį

$$s = \frac{r(k-1)}{I-1} \tag{2.10.13}$$

Iš tikrųjų, konkretus faktoriaus A lygmuo pasirodo r blokuose. Taigi šiuose r blokuose yra $rk-k$ langelių, kuriuose minimas lygmuo nepasirodo. Kita vertus, tokių langelių skaičius lygus $(I-1)s$, t. y. kitų faktoriaus A lygmenų skaičiui, padaugintam iš parametro s. Prilyginę $rk-k = (I-1)s$, gauname (2.10.13).

Šį eksperimentų planą galima interpretuoti kaip atskirą dvifaktorės dispersinės analizės, kai skirtingas stebėjimų skaičius langeliuose, atvejį. Jame $K_{ij} = 1$, kai faktoriaus i-asis lygmuo pasirodo j-ajame bloke, ir $K_{ij} = 0$ priešingu atveju.

Lentelėje 2.10.5 (2.3.2 lentelės analogas) nurodyti tie langeliai, kuriuose $K_{ij} = 1$ pagal eksperimentų planą, pateiktą 2.10.4 lentelėje.

2.10.5 lentelė. Nepilnas subalansuotų blokų eksperimentų planas

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	$\sum$
A_1		1	1	1		1		r
A_2			1	1	1		1	r
A_3	1			1	1	1		r
A_4		1			1	1	1	r
A_5	1		1			1	1	r
A_6	1	1		1			1	r
A_7	1	1	1		1			r
$\sum$	k	k	k	k	k	k	k	n

Stebėjimus Y_{ij} aprašysime adityviuoju dispersinės analizės modeliu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad (i,j) \in \mathcal{D},$$

čia $\mathcal{D}$ – dalyvaujančių eksperimente indeksų (i,j) aibė; a. d. $e_{ij} \sim N(0,\sigma^2)$ yra nepriklausomi.

2.10.3.2 Mažiausiųjų kvadratų įvertiniai ir hipotezių tikrinimas

Pagrindinė dispersinės analizės faktoriaus A įtakos nebuvimo hipotezė yra H_A : $\alpha_1 = ... = \alpha_I = 0$. Mažiau įdomi trukdančiojo faktoriaus B (bloko) įtakos nebuvimo hipotezė $H_B : \beta_1 = ... = \beta_J = 0$.

Hipotezes vėl tikriname remdamiesi 1.3.2 teorema. Visų pirma reikia rasti parametrų μ, α_i, β_j MK įvertinius ir besąlyginę liekamąją kvadratinę formą

$$SS_E = \min_{\mu,\alpha_i,\beta_j} \sum_{(i,j)\in\mathcal{D}} (Y_{ij} - \mu - \alpha_i - \beta_j)^2. \tag{2.10.14}$$

Eksperimentų planą, kai skirtingas stebėjimų skaičius langeliuose, parinkome taip, kad, kitaip nei 2.3.4 skyrelyje, parametrų įvertinius būtų galima parinkti išreikštinio pavidalo.

Pažymėkime

$$\mathcal{E} = \frac{I(k-1)}{k(I-1)}.$$

Koeficientas $\mathcal{E}$ vadinamas plano *efektyvumo daugikliu*. Kai blokai nepilni $k < I$, todėl $\mathcal{E} < 1$. Jeigu blokai pilni, tai $k = I$, todėl $\mathcal{E} = 1$.

2.10.1 teorema. *Minimizuojančius kvadratinę formą (2.10.14) parametrų α_i ir β_j MK įvertinius galima parinkti taip:*

$$\hat{\alpha}_i = \frac{1}{r\mathcal{E}}\mathcal{G}_i, \quad i = 1,...,I; \quad \hat{\beta}_j = \frac{1}{k\mathcal{E}}\mathcal{H}_j, \quad j = 1,...,J; \tag{2.10.15}$$

čia

$$\mathcal{G}_i = Y_{i\cdot} - \frac{1}{k}\sum_j K_{ij}Y_{\cdot j}, \quad \mathcal{H}_j = Y_{\cdot j} - \frac{1}{r}\sum_i K_{ij}Y_{i\cdot}.$$

Kvadratinė forma SS_E apskaičiuojama taip:

$$SS_E = \sum_{(i,j)\in\mathcal{D}} Y_{ij}^2 - r\mathcal{E}\sum_i \hat{\alpha}_i^2 - \frac{1}{k}\sum_j Y_{\cdot j}^2 =$$

$$\sum_{(i,j)\in\mathcal{D}} Y_{ij}^2 - k\mathcal{E}\sum_j \hat{\beta}_j^2 - \frac{1}{r}\sum_i (Y_{i\cdot})^2 \sim \sigma^2\chi^2_{n-I-J+1}. \tag{2.10.16}$$

Įrodymas. Remdamiesi nagrinėjamu subalansuotų blokų eksperimento planu gauname (žr. 2.10.5 lentelę)

$$K_{i\cdot} = \sum_j K_{ij} = r, \quad K_{\cdot j} = \sum_i K_{ij} = k, \quad \sum_j K_{ij}^2 = r, \quad \sum_j K_{ij}K_{i'j} = s, \quad i \neq i'.$$

Pasinaudojus šiomis lygybėmis ir eliminavus $\hat{\beta}_j$, kaip ir įrodant 2.3.2 teoremą, įvertiniams $\hat{\alpha}_i$, $i = 1, ..., I$ rasti gaunama lygčių sistema

$$(1 - \frac{1}{k})\hat{\alpha}_i - \frac{\lambda}{rk}\sum_{i'\neq i}\hat{\alpha}_{i'} = \frac{1}{r}\mathcal{G}_i, \quad i = 1, ..., I.$$

Ši lygčių sistema turi be galo daug sprendinių. Konkretų sprendinį parinkime taip, kad būtų tenkinama sąlyga $\hat{\alpha}_1 + ... + \hat{\alpha}_I = 0$. Tada gauname

$$\hat{\alpha}_i = \frac{1}{r\mathcal{E}}\mathcal{G}_i, \quad i = 1, ..., I, \tag{2.10.17}$$

nes

$$1 - \frac{1}{k} + \frac{\lambda}{rk} = \frac{r(k-1) + r(k-1)/(I-1)}{rk} = \frac{I(k-1)}{k(I-1)} = \mathcal{E}.$$

Analogiškai gauname parametrų β_j įvertinius

$$\hat{\beta}_j = \frac{1}{k\mathcal{E}}\mathcal{H}_j, \quad j = 1, ..., J. \tag{2.10.18}$$

Liekamoji kvadratinė forma SS_E gaunama įstatant į (2.10.14) arba (2.10.16) gautuosius įvertinius (2.10.17) arba (2.10.18). ▲

Grįžtame prie hipotezių H_A ir H_B tikrinimo. Remiantis 1.3.2 teorema, tikrinant hipotezę H_A reikia rasti sąlyginį kvadratinės formos (2.10.14) minimumą SS_{EH_A}, kai visi $\alpha_i = 0$, ir skirtumą $SS_A = SS_{EH_A} - SS_E$. Turime vienfaktorės dispersinės analizės modelį, kuriame faktoriaus lygmenų skaičius yra J, o stebėjimų skaičiai su kiekvienu faktoriaus lygmeniu vienodi ir lygūs k. Analogiškai 2.2.5 teoremai gauname

$$SS_{EH_A} = \sum_{(i,j)\in\mathcal{D}} (Y_{ij} - \bar{Y}_{\cdot j})^2 = \sum_{(i,j)\in\mathcal{D}} Y_{ij}^2 - \frac{1}{k}\sum_j Y_{\cdot j}^2$$

$$SS_A = SS_{EH_A} - SS_E = r\mathcal{E}\sum_i \hat{\alpha}_i^2 \sim \sigma^2\chi^2_{I-1;\lambda_A}; \tag{2.10.19}$$

necentriškumo parametras

$$\lambda_A = \frac{r\mathcal{E}}{\sigma^2}\sum_i \alpha_i^2.$$

Vidutinės kvadratų sumos $MS_A = SS_A/(I-1)$ vidurkis yra

$$\mathbf{E}(MS_A) = \sigma^2 + r\mathcal{E}\sigma_A^2, \quad \sigma_A^2 = \frac{1}{I-1}\sum_i \alpha_i^2. \tag{2.10.20}$$

Sudarome statistiką

$$F_A = \frac{SS_A(n-I-J+1)}{(I-1)SS_E} = \frac{MS_A}{MS_E},$$

kuri, kai teisinga hipotezė H_A, turi Fišerio skirstinį su $I-1$ ir $n-I-J+1$ laisvės laipsniais. Hipotezė H_A atmetama reikšmingumo lygmens α kriterijumi, kai teisinga nelygybė

$$F_A > F_\alpha(I-1, n-I-J+1). \tag{2.10.21}$$

Analogiškai, tikrindami hipotezę H_B, randame

$$SS_B = SS_{EH_B} - SS_E = k\mathcal{E}\sum_j \hat{\beta}_j^2.$$

Hipotezė H_B atmetama reikšmingumo lygmens α kriterijumi, kai teisinga nelygybė

$$F_B = \frac{SS_B(n-I-J+1)}{(J-1)SS_E} = \frac{MS_B}{MS_E} > F_\alpha(J-1, n-I-J+1). \tag{2.10.22}$$

2.10.3.3 Kontrastų analizė

Jeigu hipotezė H_A atmetama, tai atsakingiems už hipotezės atmetimą kontrastams rasti galima naudoti S metodą. Tuo tikslu reikia rasti kontrasto $\psi = \sum_i c_i\alpha_i$, $\sum_i c_i = 0$, įvertinio $\hat{\psi} = \sum_i c_i\hat{\alpha}_i$ dispersiją $\boldsymbol{V}\hat{\psi}$.

2.10.2 teorema. *Kontrasto $\psi = \sum_i c_i\alpha_i$ įvertinio $\hat{\psi}$ dispersija yra*

$$\boldsymbol{V}\hat{\psi} = \frac{\sigma^2}{r\mathcal{E}}\sum_i c_i^2. \tag{2.10.23}$$

Įrodymas. Įvertiniai $\hat{\alpha}_1, ..., \hat{\alpha}_I$ nepaslinktieji, t. y. $\mathbf{E}(\hat{\alpha}_i) = \alpha_i$, ir tenkina sąlygą $\sum_i \hat{\alpha}_i = \sum_i \alpha_i = 0$. Iš simetrijos aišku, kad a. d. $\mathcal{G}_i$ ir netgi a. v. $(\mathcal{G}_i, \mathcal{G}_{i'})^T$ yra vienodai pasiskirstę, todėl įvertiniai $\hat{\alpha}_1, ..., \hat{\alpha}_I$ turi vienodas dispersijas ir kovariacijas. Pažymėkime

$$\boldsymbol{V}(\hat{\alpha}_i) = \theta^2, \quad \mathbf{Cov}\,(\hat{\alpha}_i, \hat{\alpha}_{i'}) = \rho\theta^2.$$

Gauname

$$\boldsymbol{V}\hat{\psi} = \theta^2(\sum_i c_i^2 + \rho\sum_{i\neq i'} c_i c_{i'}) =$$

$$= \theta^2(\sum_i c_i^2 + \rho(\sum_i c_i)^2 - \rho\sum_i c_i^2) = \theta^2(1-\rho)\sum_i c_i^2. \qquad (2.10.24)$$

Kita vertus, pažymėję $\hat{\alpha}_i = \alpha_i + h_i$, $\mathbf{E}(h_i) = 0$, gausime

$$\mathbf{E}\sum_i(\hat{\alpha}_i - \bar{\hat{\alpha}}_\cdot)^2 = \sum_i \alpha_i^2 + \mathbf{E}\{\sum_i h_i^2 - I\bar{h}_\cdot^2\} =$$

$$= \sum_i(\alpha_i - \bar{\alpha}_\cdot)^2 + \theta^2(I-1)(1-\rho),$$

nes

$$\mathbf{E}(I\bar{h}_\cdot^2) = \frac{1}{I}\mathbf{E}(\sum_i h_i)^2 = \frac{1}{I}\mathbf{E}\{\sum_i h_i^2 + \sum_{i\neq i'} h_i h_{i'}\} = \theta^2 + (I-1)\rho\theta^2.$$

Pažymėję, kad $\bar{\hat{\alpha}}_\cdot = \bar{\alpha}_\cdot = 0$, gauname, kad

$$\theta^2(1-\rho) = \frac{1}{I-1}\mathbf{E}\sum_i \hat{\alpha}_i^2 - \sigma_A^2.$$

Įrašę šią išraišką į (2.10.24) ir pasinaudoję (2.10.20), gauname (2.10.24). ▲

Kontrasto ψ įvertinio $\hat{\psi}$ dispersijos įvertiniu imame

$$\hat{\boldsymbol{V}}(\hat{\psi}) = \frac{s^2}{r\mathcal{E}}, \quad s^2 = MS_E.$$

Sudarant pasikliovimo intervalus (2.1.17) reikia imti $\Delta^2 = (I-1)F_\alpha(I-1, n-I-J+1)$.

Atsižvelgiant į randomizacijos principą, atliekant eksperimentą, faktoriaus A lygmenys bloko viduje turėtų būti išdėstomi atsitiktinai.

Subalansuotų nepilnų blokų planų lenteles galima rasti [3].

2.10.4 pavyzdys. Tiriant kineskopo elektros srovės stiprumo priklausomybę nuo keturių katodo kaitinimo siūlelio apdorojimo režimų (R_1, R_2, R_3, R_4), per vieną dieną galima realizuoti tik tris apdorojimo metodus. Todėl eksperimentas atliktas pagal nepilnų subalansuotų blokų schemą, kurioje blokus atitinka dienos. Stebėjimo duomenys pateikti 2.10.6 lentelėje

2.10.6 lentelė. Statistiniai duomenys

Dienos	R_1	R_2	R_3	R_4
1	2	–	20	7
2	–	32	14	3
3	4	13	31	–
4	0	23	–	11

Atlikę skaičiavimus, gauname $MS_A = 293,611$, $MS_B = 2,056$, $MS_{AB} = 72,63$. Kadangi $F_A = MS_A/MS_{AB} = 4,04$ ir P reikšmė $\mathbf{P}\{F_{3,5} > 4,04\} = 0,0834$, tai hipotezė atmetama, jei reikšmingumo lygmuo viršija 0,0834. Tikrindami hipotezę H_B dėl bloko įtakos, gauname $F_B = 0,03$ ir atitinkanti P reikšmė yra $\mathbf{P}\{F_{3,5} > 0,03\} = 0,9928$, tai atmesti hipotezę H_B nėra pagrindo.

2.11. Lotyniškieji kvadratai

2.11.1. Statistinis modelis

Vienfaktorėje blokuotųjų duomenų analizėje, kai nagrinėjama faktoriaus A, turinčio I lygmenų, įtaka tiriamam kintamajam Y, ir yra du trukdantieji faktoriai B ir C, kurių lygmenų skaičiai yra atitinkamai J ir L, tai blokų skaičius yra JL ir matavimų skaičius eksperimente yra IJK (žr. 2.10.3 pastabą). Aptarsime vadinamąjį *lotyniškųjų kvadratų* eksperimentų planą tam atvejui, kai faktorių A, B ir C lygmenų skaičiai vienodi: $I = J = K =: m$. Pagal šį planą eksperimentų skaičių galima sumažinti iki m^2. Planas gali būti naudojamas ir nebūtinai vienfaktorėje blokuotų duomenų analizėje su dviem trukdančiais faktoriais, bet ir dvifaktorėje blokuotųjų duomenų dispersinėje analizėje, kai yra vienas trukdantysis faktorius, bei trifaktorėje dispersinėje analizėje be trukdančiųjų faktorių.

Lotyniškuoju kvadratu vadinama lentelė, turinti m eilučių ir m stulpelių, kuriuose įrašyti skaičiai nuo 1 iki m. Skaičiai išdėstyti taip, kad kiekvienoje eilutėje ir kiekviename stulpelyje kiekvienas skaitmuo parašytas po vieną kartą.

Tai, kad toks skaitmenų išdėstymas yra galimas bet kokiam m, iliustruoja žemiau pateikta lentelė.

2.11.1 lentelė. Standartinis lotynų kvadratas

1	2	3	...	m
2	3	4	...	1
3	4	5	...	2
...	...	...	...	...
m	1	2	...	$m-1$

Tarkime, kad įrašytas skaitmuo rodo faktoriaus A lygmenį, eilutės numeris – faktoriaus B lygmenį, o stulpelio numeris – faktoriaus C lygmenį. Tokiu būdu kiekvieną faktorių B ir C lygmenų rinkinį atitinka tik viena tiriamo faktoriaus A reikšmė (o ne m reikšmių, kaip kryžminėje klasifikacijoje).

Šis planas atsirado taikant dispersinę analizę žemės ūkyje.

2.11.1 pavyzdys. Lyginamas m kviečių veislių (faktorius A) derlingumas (požymis Y). Siekiant pašalinti dirvožemio derlingumo įtaką (šlaituose priklausomai nuo vietos dirvožemio derlingumas skirtingas, nes skiriasi ne tik derlingos žemės sluoksnio gylis, bet ir apšvietimas, drėgmės kiekis), žemės sklypas horizontaliomis ir vertikaliomis linijomis padalijamas į m^2 sklypelių, kurie apsėjami skirtingomis kviečių veislėmis pagal lotyniškojo kvadrato planą. Tariant, kad faktorių sąveikos nėra, eliminuojama trukdančiųjų faktorių B (derlingumo kitimas vertikalia kryptimi) ir C (derlingumo kitimas horizontalia kryptimi) įtaka.

2.11.2 pavyzdys. Testuojant tris sensorines programas (faktorius A) ir lyginant jų efektyvumą Y, parenkami trys skirtingi vertintojai (trukdantysis faktorius B) ir kiekvienas iš jų ryte, vidurdienį ir vakare (trukdantysis faktorius C) įvertina programų efektyvumą pagal lotyniškojo kvadrato planą.

2.11.3 pavyzdys. Tiriama keturių skirtingų hormonų įtaka konkrečiam fermentui karvės kraujuje. Skirtingų karvių kraujo sudėtis skiriasi, be to, ji ilgainiui kinta. Keturi hormonai duodami keturioms karvėms keturiais skirtingais laiko periodais pagal lotyniškojo kvadrato planą.

2.11.4 pavyzdys. Lyginama penkių raketinio kuro paruošimo metodų įtaka degimo greičiui. Mišiniai daromi iš žaliavos, kuri ateina paketais. Paketų sudėtis gali skirtis. Mišinius ruošia skirtingi operatoriai. Atsitiktinai parenkami penki operatoriai ir penki žaliavos paketai ir kiekvienas operatorius paruošia raketinį kurą pagal lotyniškojo kvadrato planą.

Sukeitus lotyniškojo kvadrato bet kurias eilutes arba bet kuriuos stulpelius, vėl gaunamas lotyniškasis kvadratas. Keiskime stulpelius taip, kad pirmoje eilutėje gautume iš eilės surašytus skaitmenis $1, 2, ..., m$. Paskui, keisdami vietomis eilutes (išskyrus pirmąją), pasieksime, kad pirmajame stulpelyje būtų iš eilės surašyti skaitmenys $1, 2, ..., m$. Toks lotyniškasis kvadratas vadinamas standartiniu. Taigi iš kiekvieno standartinio lotyniškojo kvadrato galima gauti $m!(m-1)!$ skirtingų lotyniškųjų kvdratų.

Pagal randomizacijos principą, atliekant eksperimentą, konkretų lotyniškąjį kvadratą reikia pasirinkti atsitiktinai. Tai galima atlikti tokiu būdu. Atsitiktinai parenkame standartinį lotyniškąjį kvadratą. Paskui atsitiktinai parenkame lotyniškąjį kvadratą iš dydžio $m!(m-1)!$ aibės, kuri gali būti gaunama iš šio standartinio keičiant vietomis stulpelius ir eilutes.

Lotyniškųjų kvadratų modelis: a. d. Y_{ijk} aprašomi modeliu

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}, \quad (i,j,k) \in \mathcal{D} \tag{2.11.1}$$

čia indeksai i, j ir k atitinkamai žymi faktorių A, B ir C lygmenų numerius; $\mathcal{D}$ – dydžio m^2 indeksų aibė, kurioje indeksai kinta pagal parinktą lotyniškąjį kvadratą; paklaidos $e_{ijk} \sim N(0, \sigma^2)$ nepriklausomos. Parametrai α_i, β_j ir γ_k tenkina sąlygas

$$\sum_{i=1}^{m} \alpha_i = 0, \quad \sum_{j=1}^{m} \beta_j = 0, \quad \sum_{k=1}^{m} \gamma_k = 0. \tag{2.11.2}$$

2.11.2. Parametrų įvertiniai ir hipotezių tikrinimas

Parametro $\boldsymbol{\theta} = (\mu, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m, \gamma_1, \ldots, \gamma_m)^T$ MK įvertinį randame minimizuodami kvadratinę formą

$$SS(\boldsymbol{\theta}) = \sum_{(i,j,k)\in\mathcal{D}} (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_k)^2.$$

2.11.1 teorema. *Parametrų įvertiniai turi tokį pavidalą:*

$$\hat{\mu} = \bar{Y}_{...} = \frac{1}{m^2} \sum_{(i,j,k)\in\mathcal{D}} Y_{ijk}, \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad i = 1, ..., m,$$

$$\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad j = 1, ..., m, \quad \hat{\gamma}_k = \bar{Y}_{..k} - \bar{Y}_{...}, \quad k = 1, ..., m; \tag{2.11.3}$$

čia

$$\bar{Y}_{i..} = \frac{1}{m} \sum_{(j,k)\in\mathcal{D}_{i..}} Y_{ijk}, \quad \bar{Y}_{.j.} = \frac{1}{m} \sum_{(i,k)\in\mathcal{D}_{.j.}} Y_{ijk}, \quad \bar{Y}_{..k} = \frac{1}{m} \sum_{(i,j)\in\mathcal{D}_{..k}} Y_{ijk},$$

o $\mathcal{D}_{i..}$ *yra aibė tokių porų* (j,k), *kad* $(i,j,k) \in \mathcal{D}$, *o* i *yra fiksuotas; analogiškai apibrėžiamos aibės* $\mathcal{D}_{.j.}$, $\mathcal{D}_{..k}$. *Liekamoji kvadratų suma yra*

$$SS_E = SS(\hat{\boldsymbol{\theta}}) = \sum_{(i,j,k)\in\mathcal{D}} (Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_k)^2$$

$$= \sum_{(i,j,k)\in\mathcal{D}} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{..k} + 2\bar{Y}_{...})^2 \sim \sigma^2 \chi^2_{(m-1)(m-2)}. \tag{2.11.4}$$

Įrodymas. Diferencijuodami kvadratų sumą $SS(\boldsymbol{\theta})$ pagal μ ir prilyginę išvestinę 0, gauname lygtį

$$\sum_{(i,j,k)\in\mathcal{D}} (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_k) = 0.$$

Sumuojant α_i aibėje $\mathcal{D}$ indeksas i kinta nuo 1 iki m. Todėl remdamiesi (2.11.2) susumavę gauname 0. Analogiškai gaunama sumuojant β_j ir γ_k. Taigi parametro μ mažiausiųjų kvadratų įvertinys

$$\hat{\mu} = \bar{Y}_{...} = \frac{1}{m^2} \sum_{(i,j,k)\in\mathcal{D}} Y_{ijk}. \tag{2.11.5}$$

Diferencijuodami pagal α_i ir prilyginę išvestinę 0, gauname lygtis

$$\sum_{(j,k)\in\mathcal{D}_{i..}} (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_k) = 0, \quad i = 1, ..., m.$$

Pagal lotyniškojo kvadrato apibrėžimą aibė $\mathcal{D}_{i..}$ susideda iš tokių porų (j,k), kad kiekvienas indeksas j ir k įgyja reikšmes nuo 1 iki m po vieną kartą. Todėl sumuodami β_j ir γ_k aibėje $\mathcal{D}_{i..}$ gausime nulį. Gauname parametrų α_i įvertinius (2.11.3). Analogiškai gauname kitus (2.11.3) įvertinius.

Liekamoji kvadratinė forma SS_E gaunama įstačius į $SS(\boldsymbol{\theta})$ gautuosius įvertinius (2.11.3). Gauname:

$$SS_E = \sum_{(i,j,k)\in\mathcal{D}} (Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_k)^2 =$$

$$= \sum_{(i,j,k)\in\mathcal{D}} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{..k} + 2\bar{Y}_{...})^2 \sim \sigma^2 \chi^2_{(m-1)(m-2)}.$$

Iš tikrųjų, kadangi nežinomų parametrų (2.11.1) modelyje yra $(3m+1)-3 = 3m-2$ (nes $3m+1$ parametrai $\mu, \alpha_i, \beta_j, \gamma_k$ susieti trimis lygybėmis (2.11.2)), o

stebėjimų skaičius m^2, tai laisvės laipsnių skaičius yra $m^2-(3m-2)=(m-1)(m-2)$. ▲

Pereiname prie faktorių įtakos nebuvimo hipotezių $H_A:\alpha_1=...=\alpha_m=0$, $H_B:\beta_1=...=\beta_m=0$, $H_C:\gamma_1=...=\gamma_m=0$ tikrinimo.

Remiantis 1.3.2 teorema, tikrinant hipotezę H_A reikia rasti sąlyginį kvadratinės formos $SS(\boldsymbol{\theta})$ minimumą SS_{EH_A}, kai visi $\alpha_i=0$, ir skirtumą $SS_A=SS_{EH_A}-SS_E$. Jei hipotezė H_A teisinga, parametrų μ,β_j,γ_k įvertiniai sutampa su (2.11.3). Taigi

$$SS_{EH_A}=\sum_{(i,j,k)\in\mathcal{D}}(Y_{ijk}-\hat{\mu}-\hat{\beta}_j-\hat{\gamma}_k)^2,$$

$$SS_A=SS_{EH_A}-SS_E=m\sum_i(\bar{Y}_{i..}-\bar{Y}_{...})^2. \tag{2.11.6}$$

Analogiškai gauname kvadratų sumas, apibūdinančias faktorių B ir C įtaką

$$SS_B=m\sum_j(\bar{Y}_{.j.}-\bar{Y}_{...})^2,\quad SS_C=m\sum_k(\bar{Y}_{..k}-\bar{Y}_{...})^2. \tag{2.11.7}$$

Kadangi parametrai α_i tenkina lygybę (2.11.2), tai hipotezė H_A gali būti užrašyta šitaip $H_A:\alpha_1=...=\alpha_{m-1}=0$, taigi pagal 1.3.1 pastabą ir pavidalu $H_A:\boldsymbol{H\theta}=\mathbf{0}$; čia $\boldsymbol{H}$ yra $(m-1)\times(3m-2)$ rango $m-1$ matrica. Analogiškai užrašomos hiptezės H_B ir H_C. Pagal 1.3.2 teoremą

$$SS_A/\sigma^2\sim\chi^2(m-1,\lambda_A),\quad SS_B/\sigma^2\sim\chi^2(m-1,\lambda_B),\quad SS_C/\sigma^2\sim\chi^2(m-1,\lambda_C),$$

$$\lambda_A=m\sum_i\alpha_i^2,\quad \lambda_B=m\sum_j\beta_j^2,\quad \lambda_C=m\sum_k\gamma_k^2.$$

Vidutinių kvadratų sumų

$$MS_A=\frac{SS_A}{m-1},\quad MS_B=\frac{SS_B}{m-1},\quad MS_C=\frac{SS_C}{m-1},$$

vidurkiai

$$\mathbf{E}(MS_A)=\sigma^2+m\sigma_A^2,\quad \mathbf{E}(MS_B)=\sigma^2+m\sigma_B^2,\quad \mathbf{E}(MS_C)=\sigma^2+m\sigma_C^2;$$

čia

$$\sigma_A^2=\frac{1}{m-1}\sum_i\alpha_i^2,\quad \sigma_B^2=\frac{1}{m-1}\sum_j\beta_j^2,\quad \sigma_C^2=\frac{1}{m-1}\sum_k\gamma_k^2.$$

Gautus rezultatus surašome į dispersinės analizės lentelę.

2.11.2 lentelė. Dispersinės analizės lentelė

Faktorius	SS	ν	MS	$\mathbf{E}$(MS)
A	SS_A	$m-1$	MS_A	$\sigma^2+m\sigma_A^2$
B	SS_B	$m-1$	MS_B	$\sigma^2+m\sigma_B^2$
C	SS_C	$m-1$	MS_C	$\sigma^2+m\sigma_C^2$
E	SS_E	$(m-1)(m-2)$	MS_E	σ^2

Hipotezės H_A, H_B, H_C yra atmetamos reikšmingumo lygmens α kriterijais, kai statistikos

$$F_A = \frac{MS_A}{MS_E}, \quad F_B = \frac{MS_B}{MS_E}, \quad F_C = \frac{MS_C}{MS_E}$$

atitinkamai viršija Fišerio skirstinio su $m-1$ ir $(m-1)(m-2)$ laisvės laipsnių kritinę reikšmę $F_\alpha(m-1,(m-1)(m-2))$.

Jeigu kuri nors hipotezė atmetama, tai kontrastus, atsakingus už hipotezės atmetimą, galima rasti S metodu. Pavyzdžiui, kontrasto $\psi = \sum_i c_i\alpha_i$, $\sum_i c_i = 0$ įvertinys yra

$$\hat{\psi} = \sum_i c_i\hat{\alpha}_i = \sum_i c_i\bar{Y}_{i..}.$$

Šio įvertinio dispersija ir dispersijos įvertinys

$$\boldsymbol{V}(\hat{\psi}) = \frac{\sigma^2}{m}\sum_i c_i^2, \quad \hat{\boldsymbol{V}}(\hat{\psi}) = \frac{s^2}{m}\sum_i c_i^2, \quad s^2 = MS_E. \tag{2.11.8}$$

Sudarant pasikliovimo intervalus (2.1.17) reikia imti $\Delta^2 = (m-1)F_\alpha(m-1,(m-1)(m-2))$.

Kadangi a. d. $\bar{Y}_{i..}$ turi vienodas dispersijas, tai pritaikomas ir kontrastų palyginimo T metodas.

2.11.5 pavyzdys.[1] Automobilių gamykla tiria keturių tipų padangų kiekvieno iš keturių lengvųjų automobilių tipų dėvėjimosi greitį. Eksperimentas atliktas pagal lotyniškųjų kvadratų planą. Užregistruota padangos nudilimas (mm) po 10 000 mylių ridos. Duomenys pateikti 2.11.3 lentelėje

2.11.3 lentelė. Statistiniai duomenys

2,12 (II)	1,73 (I)	1,65 (IV)	1,89 (III)
1,83 (III)	2,28 (II)	1,67 (I)	2,01 (IV)
1,83 (IV)	2,27 (III)	2,18 (II)	2,03 (I)
1,85 (I)	1,93 (IV)	2,24 (III)	2,52 (II)

Šioje lentelėje stulpeliai atitinka automobilius (faktorius A), eilutės – padangos padėtį ant automobilio (faktorius B); langeliuose nurodytas padangos nusidėvėjimas (tiriamasis kintamasis Y) ir (skliausteliuose) padangos markė (faktorius C).

Atlikę skaičiavimus gauname $MS_A = 0,0376$, $MS_B = 0,0670$, $MS_C = 0,1765$ ir $MS_E = 0,0143$. Tikrindami hipotezes H_A, H_B, H_C gauname statistikų realizacijas $F_A = 2,63$, $F_B = 4,69$, $F_C = 12,36$, kurias atitinka P reikšmės 0,1445, 0,0514, 0,0056. Galima daryti išvadą, kad duomenys prieštarauja padangos markės įtakos nebuvimą prielaidai, neprieštarauja prielaidai, kad automobilis neturi įtakos. Kalbant apie padangos padėties ant automobilio įtaką padangos dėvėjimuisi, tai hipotezė H_B atmetama, jei pasirinktas kriterijaus reikšmingumo lygmuo viršija 0,0514.

2.12. Pratimai

2.1 skyrelis

2.1. Tegu $Y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, ..., I, j = 1, ..., J$; čia e_{ij} yra vienodai pasiskirstę n. a. d. ir $\mathbf{E}(e_{ij}) = 0$, o parametrai α_i tenkina sąlygą $\sum_i d_i\alpha_i = 0$, kai d_i žinomi ir $\sum_i d_i \neq 0$. Raskite parametrų μ ir α_i mažiausiųjų kvadratų įvertinius.

2.2. Tegu $Y_{ij} = \mu_i + e_{ij}, i = 1, ..., I, j = 1, ..., J$; čia e_{ij} yra nepriklausomi ir normalieji $e_{ij} \sim N(0, \sigma^2)$. a) Raskite hipotezės $H : \mu_1 = 2\mu_2 = 3\mu_3$ tikrinimo kriterijų, kai $I = 4$. b) Įrodykite, kad hipotezės $H : \mu_1 = \mu_2$ tikrinimo F kriterijus, kai $I = 2$, yra ekvivalentus Stjudento kriterijui dėl vidurkių lygybės dviejose normaliosiose imtyse.

2.3. Lentelėje pateikti duomenys, apibūdinantys iškvepiamo azoto kiekį Y esant keturioms skirtingoms dietoms (faktoriaus A lygmenys).

A_1	A_2	A_3	A_4
4,079	4,368	4,169	4,928
4,859	5,668	5,709	5,608
3,540	3,752	4,416	4,940
5,047	5,848	5,666	5,291
3,298	3,802	4,123	4,674
4,679	4,844	5,059	5,038
2,870	3,578	4,403	4,905
4,648	5,393	4,496	5,208
3,847	4,374	4,688	4,806

Atlikite duomenų vienfaktorę dispersinę analizę ir kontrastų palyginimą.

2.4. Iš darbininkų, aptarnaujančių didelės įmonės surinkimo konvejerį, buvo atrinkti 4 darbininkai ir kiekvienam iš jų buvo užfiksuotas tam tikros detalės surinkimo laikas.

Darbininkai	Laikas							
1	24,2;	22,2;	24,5;	21,1;	22,0;			
2	19,4;	21,1;	16,2;	21,2;	21,6;	17,8;	19,6;	
3	19,0;	23,1;	23,8;	22,8;				
4	19,9;	15,7;	15,2;	19,8;	18,9;	16,1;	16,2;	18,5

Ar skiriasi darbininkai pagal detalės surinkimo laiką?

2.5. Lentelėje pateikti duomenys, apibūdinantys gumos tempiamąjį atsparumą Ykg/cm^2 priklausomai nuo vulkanizavimo laiko X min.

X_i	Y_{ij}			
20	152	152	147	152
25	158	155	146	169
30	149	159	115	152
40	143	121	116	156
60	126	165	153	157

Patikrinkite hipotezę, kad gumos tempiamasis atsparumas nepriklauso nuo vulkanizavimo laiko.

2.6. Tiriant retųjų elementų pasiskirstymą triaso amžiaus nuogulose netoli Birštono, buvo gauti lentelėje pateikti rezultatai (g/t) trijuose šių nuogulų horizontuose ($\bar{X}$ ir s empiriniai vidurkiai ir vidutinio kvadratinio nuokrypio įvertiniai; stebėjimų skaičius atitinkamai I, II ir III horizontuose yra 136, 77, 111).

Elementai	I		II		III	
	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s
Varis	43	12	47	12	44	22
Švinas	13	4	16	5	22	8
Titanas	3428	701	3531	776	4255	1071
Manganas	940	182	1022	146	828	296
Chromas	74	21	84	22	110	64
Nikelis	44	13	55	16	60	22

Patikrinkite hipotezes, kad elementų koncentracija visuose trijuose horizontuose yra vienoda.

2.2 ir 2.3 skyreliai

2.7. Tegu a. d. $e_{ij}, i = 1, ..., I, j = 1, ..., J$ yra nepriklausomi ir normalieji $e_{ij} \sim N(0, \sigma^2)$. Įrodykite, kad kvadratų sumos

$$\sum_i (\bar{e}_{i.} - \bar{e}_{..})^2, \quad \sum_i \sum_j (e_{ij} - \bar{e}_{i.} - \bar{e}_{.j} + \bar{e}_{..})^2$$

yra nepriklausomos.

2.8. Tegu dvifaktorės dispersinės analizės schemoje faktoriaus A lygmenų skaičius $I = 2$, o faktoriaus B lygmenų skaičius $J \geq 2$; kiekviename langelyje turime po vieną stebėjimą $Y_{ij}, i = 1, 2; j = 1, ..., J$. Įrodykite, kad hipotezės H_A tikrinimo F kriterijus yra ekvivalentus Stjudento kriterijui, grindžiamam skirtumais $Z_j = Y_{1j} - Y_{2j}$, $j = 1, ..., J$. Todėl dispersinės analizės prielaidos gali būti susilpnintos: kriterijus nepakis, jeigu tarsime, kad paklaidų vektoriai $(e_{1j}, e_{2j})^T, j = 1, ..., J$ yra nepriklausomi ir turi dvimatį normalųjį skirstinį su nuliniu vidurkių vektoriumi.

2.9 (**2.8** tęsinys). Lentelėje pateikti miego trukmės padidėjimo duomenys Y_{1j}, naudojant pirmo tipo migdomuosius vaistus, ir Y_{2j} – naudojant antro tipo migdomuosius vaistus; čia j žymi paciento numerį, $j = 1, ..., 10$.

$i; j$	1	2	3	4	5	6	7	8	9	10
1	+0,7	-1,6	-0,2	-1,2	-1,0	+3,4	+3,7	+0,8	0,0	+2,0
2	+1,9	+0,8	+1,1	+0,1	-0,1	+4,4	+5,5	+1,6	+4,6	+3,4

a) Priėmę normalumo prielaidą patikrinkite hipotezę, kad miego trukmės padidėjimo vidurkiai, naudojant pirmo ir antro tipo migdomuosius, nesiskiria; $\alpha = 0,01$.

b) Tarkime, yra žinoma, kad $\boldsymbol{V}(Y_{1j} - Y_{2j}) \leq 1,25$. Kokį skaičių pacientų reikėtų turėti, kad tikrinant vidurkių lygybės hipotezę, ji būtų atmesta su tikimybe, ne mažesne už 0,95, kai miego trukmės padidėjimo vidurkių skirtumas viršija 1 valandą.

2.10. Tarkime, kad **2.3** pratime tiriame iškvepiamo azoto kiekio priklausomybę ne tik nuo dietos (faktorius A), bet ir nuo lyties (faktorius B). Gauti stebėjimų duomenys pateikti lentelėje [1].

	A_1	A_2	A_3	A_4
B_1	4,079	4,368	4,169	4,928
	4,859	5,668	5,709	5,608
	3,540	3,752	4,416	4,940
B_2	2,870	3,578	4,403	4,905
	4,648	5,393	4,496	5,208
	3,847	4,374	4,688	4,806

Atlikite dvifaktorę dispersinę analizę. Patikrinkite pagrindines dispersinės analizės hipotezes.

2.11. Iš kiekvienos 4 pelių (faktorius A) palikuonių buvo atrinkta po 1 peliuką ir jam

buvo taikoma viena iš trijų dietų (faktorius B). Po trijų savaičių išmatuotas svorio priaugis Y [1].

	A_1	A_2	A_3	A_4
B_1	5,2	11,4	4,2	10,7
B_2	7,4	13,0	9,5	8,8
B_3	9,1	13,8	8,8	13,0

Atlikite dvifaktorę dispersinę analizę su vienu stebėjimu langelyje. Remdamiesi Tjukio kriterijumi patikrinkite sąveikos nebuvimo hipotezę.

2.12. Tiriant, kiek aštuonių skirtingų rūšių aliejaus (faktorius A) sugeria spurgos, šešias dienas (faktorius B) buvo gaminamos vienodo didumo spurgų partijos su kiekviena aliejaus rūšimi [14].

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
B_1	164	172	177	178	163	163	150	164
B_2	177	197	184	196	177	193	179	169
B_3	168	167	187	177	144	176	146	155
B_4	156	161	169	181	165	172	141	149
B_5	172	180	179	184	166	176	169	170
B_6	196	190	197	191	178	178	183	167

a) Užpildykite dispersinės analizės lentelę ir patikrinkite pagrindines hipotezes.

b) Jeigu H_A atmetama, tai nurodykite kontrastą, kuris reikšmingai skiriasi nuo 0.

c) Ar naudojant T metodą galima nurodyti aliejaus rūšis, kurias spurgos sugeria skirtingai.

2.13 (**2.12** tęsinys). Tarkime, kad skirtingų aliejaus rūšių sugėrimo vidurkiai tenkina sąlygas $\mu_5 = \mu_7 = \mu_8 = \mu$, $\mu_1 = \mu_2 = \mu_6 = \mu + 12$, $\mu_3 = \mu_4 = \mu + 22$. Kokia tikimybė atmesti hipotezę H_A, jeigu paklaidos dispersija lygi **2.12** pratime surastam įverčiui (kriterijaus reikšmingumo lygmuo $\alpha = 0,05$)?

2.14 (**2.13** tęsinys). Kadangi 5, 7 ir 8 aliejaus rūšys atrodo ekonomiškiausios, tai tolesni eksperimentai bus atliekami tik su šiomis trimis rūšimis. Kiek eksperimentų reikėtų atlikti tikrinant hipotezę $H : \mu_5 = \mu_7 = \mu_8$ kriterijumi, kai reikšmingumo lygmuo $\alpha = 0,05$, kad bet kokį vidurkių skirtumą, viršijantį 10 vienetų, galėtume pastebėti su tikimybe, ne mažesne už 0,8?

2.15. Tiriama konservų dėžutės svorio priklausomybė nuo mėsos tiekėjo (faktorius A) ir nuo dėžutes užpildančio automato užpildymo cilindro (faktorius B). Iš kiekvieno tiekėjo ir kiekvieno cilindro konservų dėžučių partijos atsitiktinai atrenkama po 3 dėžutes. Dėžučių svoriai (sąlyginiais vienetais) pateikti lentelėje [14].

	A_1			A_2			A_3			A_4			A_5		
B_1	1;	1;	2	4;	3;	5	6;	3;	7	3;	1;	3	1;	3;	3
B_2	-1;	3;	-1	-2;	1;	0	3;	1;	5	2;	0;	1	1;	0;	1
B_3	1;	1;	1	2;	0;	1	2;	4;	3	1;	3;	3	3;	3;	3
B_4	-2;	3;	0	-2	0;	1	3;	3;	4	0;	0;	2	0;	1;	1
B_5	1;	1;	-1	2;	1;	5	0;	1;	2	1;	0;	-1	-2;	3;	1
B_6	0;	1;	1	0;	0;	3	3;	3;	4	3;	0;	2	3;	1;	2

Užpildykite dispersinės analizės lentelę ir patikrinkite pagrindines hipotezes.

2.16. Eksperimente hibridinius žiurkiukus maitino hibridinės žiurkių patelės. Lentelėje pateikti žiurkiukų svoriai praėjus 28 dienoms nuo gimimo. Šiame eksperimente faktorius A yra maitinančios žiurkės genotipas, o faktorius B – žiurkiukų vados genotipas.

	A_1	A_2	A_3	A_4		A_1	A_2	A_3	A_4
B_1	61,5	55,0	52,5	42,0	B_3	37,0	56,3	39,6	50,0
	68,2	42,0	61,8	54,0		36,3	69,8	46,0	43,8
	64,0	60,2	49,5	61,0		68,0	67,0	61,3	54,5
	65,0		52,7	48,2				55,3	
	59,7			39,6				55,7	
B_2	60,3	50,8	56,5	51,3	B_4	59,0	59,5	45,2	44,8
	51,7	64,7	59,0	40,5		57,4	52,8	57,0	51,5
	49,3	61,7	47,2			54,0	56,0	61,4	53,0
	48,0	64,0	53,0						42,0
		62,0							54,0

Atlikite dvifaktorę dispersinę analizę, kai stebėjimų skaičiai langeliuose skirtingi. Patikrinkite sąveikos nebuvimo hipotezę. Patikrinkite faktorių A ir B įtakos nebuvimo hipotezes dviem būdais: a) tariant, kad modelis adityvus; b) atsižvelgiant į sąveiką.

2.17. Įrodykite, kad dvifaktorėje dispersinėje analizėje, kai $I = 2$, kvadratų sumas SS_A ir SS_{AB} galima apskaičiuoti šitaip:

$$SS_A = JK(\bar{Y}_{1..} - \bar{Y}_{2..})^2/2,$$
$$SS_{AB} = (K\sum_j(\bar{Y}_{1j.} - \bar{Y}_{2j.})^2 - SS_A)/2,$$

o jei $K = 2$, tai

$$SS_E = \sum_i\sum_j(Y_{ij1} - Y_{ij2})^2/2.$$

2.4 skyrelis

2.18. Tegu $Y_{ij} = \mu + a_i + e_{ij}$ ir a. d. $\{a_i\}$, $\{e_{ij}\}$ nepriklausomi, $a_i \sim N(0, \sigma_A^2)$, $e_{ij} \sim N(0, \sigma^2)$;$i = 1, ..., I, j = 1, ..., J$. Parametrai $\mu, \sigma_A^2, \sigma^2$ nežinomi. Raskite hipotezės $H : \sigma_A^2/\sigma^2 \leq \Delta$, kai alternatyva $\bar{H} : \sigma_A^2/\sigma^2 > \Delta$, tikrinimo TGN kriterijų.

2.19. Atsitiktinai parinkus keturis gaminius po 10 kartų buvo išmatuotas parametras X:

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
6,34	5,95	5,23	4,55	6,85	6,52	5,52	4,73
6,36	6,04	5,27	4,65	6,91	6,60	5,52	4,78
6,41	6,11	5,32	4,68	6,91	6,62	5,53	4,78
6,42	6,31	5,39	4,68	7,02	6,64	5,60	4,84
6,80	6,36	5,40	4,72	7,12	6,71	5,78	4,86

(a) atlikite vienfaktorę dispersinę analizę su vienu atsitiktiniu faktoriumi A (jo lygmenys – gaminių numeriai);
(b) raskite taškinius ir intervalinius ($Q = 0,95$) parametrų μ, σ_A^2, σ^2 įverčius;
(c) raskite parametro σ_A^2/σ^2 pasikliovimo intervalą ($Q = 0,95$).

2.20. Konservų fabriko technologiniame procese kiekvienas pjaustantis abrikosus operatorius buvo stebimas penkis dviejų minučių laikotarpius. Trijose skirtingose linijose buvo pjaustomi trijų skirtingų dydžių vaisiai (didesnis numeris reiškia mažesnį vaisių) ir užfiksuojamas supjaustytų vaisių skaičius Y_{ij}; $i = 1, ..., I$ yra operatoriaus numeris, o $j = 1, ..., 5$ žymi laikotarpio numerį. Analizės rezultatai atskirai kiekvieno dydžio vaisiams pateikti lentelėje.

Dydis	I	$Y_{..}$	MS_A	MS_E
2	9	53,17	59,72	1,144
3	17	52,26	68,20	2,537
4	17	47,32	78,96	4,926

Tardami, kad yra teisingas modelis $Y_{ij} = \mu + a_i + e_{ij}$, kai a. d. $\{a_i\}$ ir $\{e_{ij}\}$ yra nepriklausomi ir normalieji $a_i \sim N(0, \sigma_A^2)$, $e_{ij} \sim N(0, \sigma^2)$, raskite taškinius parametrų $\mu, \sigma_A^2, \sigma^2$ įverčius.

2.21. Tarkime, kad **2.3** pratime faktorius A yra atsitiktinis. Kaip pasikeis dispersinė analizė?

2.5 skyrelis

2.22. Tegu $Y_{ijk} = \mu + a_i + b_j + e_{ijk}$ ir a. d. $\{a_i\}$, $\{b_j\}$, $\{e_{ijk}\}$ nepriklausomi, $a_i \sim N(0, \sigma_A^2)$, $b_j \sim N(0, \sigma_B^2)$, $e_{ijk} \sim N(0, \sigma^2)$;$i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$. Parametrai $\mu, \sigma_A^2, \sigma_B^2, \sigma^2$ nežinomi. Raskite TGN kriterijų

a) hipotezei $H : \tau^2 = \sigma_A^2/(\sigma^2 + K\sigma^2) \leq \Delta$, kai alternatyva $\bar{H} : \tau^2 > \Delta$, tikrinti;

b) hipotezei $H : \sigma_{AB}/\sigma^2 \leq \Delta$, kai alternatyva $\bar{H} : \sigma_{AB}^2/\sigma^2 > \Delta$, tikrinti.

2.23. Atlikus dvifaktorę dispersinę analizę su dviem atsitiktiniais faktoriais A ir B, gauta tokia dispersinės analizės lentelė

Faktorius	ν	SS
A	24	3 243
B	3	46 659
$A \times B$	72	459
E	1100	243

(a) patikrinkite hipotezes H_A, H_B, H_{AB}, kai reikšmingumo lygmuo $P = 0,025$;

(b) raskite dispersijos komponenčių σ_A^2, σ_B^2, σ_{AB}^2, σ^2 įverčius ir jų dispersijų įverčius;

(c) apskaičiuokite kiekvienos dispersijos komponentės apytikslį pasikliovimo intervalą, kai pasikliovimo lygmuo $Q = 0,95$.

2.24. Atlikite dispersinę analizę pagal **2.11** pratimo duomenis tarę, kad abu faktoriai yra atsitiktiniai.

2.25. Atlikite dispersinę analizę pagal **2.15** pratimo duomenis tarę, kad abu faktoriai yra atsitiktiniai.

2.6 skyrelis

2.26. Lentelėje pateikti duomenys, apibūdinantys kuro ištekėjimo greitį iš trijų skirtingų tipų tūtų greitį; matavimus atliko 5 operatoriai, iš kurių kiekvienas atliko po 3 matavimus kiekvienoje tūtoje.

Tūta	1			2			3			4			5		
A	6	6	-15	26	12	5	11	4	4	21	14	7	25	18	25
B	13	6	13	4	4	11	17	10	17	-5	2	-5	15	8	1
C	10	10	-11	-35	0	-14	11	-10	-17	-12	-2	-16	-4	10	24

(a) atlikite mišraus modelio, kuriame faktorius A (tūtos) yra pastovus, o faktorius B (operatorius) – atsitiktinis dispersinę analizę;

(b) atlikite analizę laikydami abu faktorius pastoviais. Kuo paaiškinti skirtingus atsakymus, gautus tikrinant hipotezę, kad rezultatai nepriklauso nuo tūtų tipo.

2.27. Lentelėje pateikta tam tikros medžiagos nepralaidumo vandeniui charakteristika priklausomai nuo trijų tipų staklių (faktorius A), su kuriomis ji buvo pagaminta, per 9 skirtingas dienas (faktorius B) [14].

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
A_1	1,40	1,45	1,91	1,89	1,77	1,66	1,92	1,84	1,54
	1,35	1,57	1,48	1,48	1,73	1,54	1,93	1,79	1,43
	1,62	1,82	1,89	1,39	1,54	1,68	2,13	2,04	1,70
A_2	1,31	1,24	1,51	1,67	1,23	1,40	1,23	1,58	1,64
	1,63	1,18	1,58	1,37	1,40	1,45	1,51	1,63	1,07
	1,41	1,52	1,65	1,11	1,53	1,63	1,44	1.28	1,38
A_3	1,93	1,43	1,38	1,72	1,32	1,63	1,33	1,69	1,70
	1,40	1,86	1,36	1,37	1,34	1,36	1,38	1,80	1,84
	1,62	1,69	1,49	1,43	1,48	1,49	1,29	1,45	1,75

Atlikite duomenų dispersinę analizę, tarę, kad modelis yra mišrusis, kuriame faktorius A yra pastovus, o faktorius B – atsitiktinis.

2.28. Tiriant dujų sunaudojimą per devynias savaites (faktorius A) buvo fiksuojamas dujų sunaudojimas per parą kiekvieną savaitės dieną nuo pirmadienio iki šeštadienio imtinai (faktorius B). Gauti rezultatai (sąlyginiais vienetais) pateikiami lentelėje [14].

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
B_1	5	1	-4	5	-13	-8	-2	-4	-10
B_2	3	6	-10	-2	-7	-2	-4	2	2
B_3	8	4	-14	-3	3	0	5	-11	-12
B_4	8	10	-5	-1	4	-2	4	1	-12
B_5	4	-1	7	-5	5	-3	-7	-3	-6
B_6	3	-9	3	-8	-6	0	-3	8	-1

Atlikite dispersinę analizę tarę, kad faktorius A atsitiktinis, o faktorius B pastovus.

2.29. Atlikite **2.10** pratimo duomenų analizę, tarę, kad faktorius A yra atsitiktinis, o faktorius B – pastovus.

2.30. Atlikite **2.12** pratimo duomenų analizę, tarę, kad faktorius A yra pastovus, o faktorius B – atsitiktinis.

2.7 skyrelis

2.31. Trifaktorės dispersinės analizės skirtingų langelių stebėjimų vidurkiai $\mu_{ijl}, i = 1,2,3;\ j = 1,2,3; l = 1,2$, pateikti lentelėse

C_1	B_1	B_2	B_3	C_2	B_1	B_2	B_3
A_1	5	6	10	A_1	9	7	14
A_2	7	7	1	A_2	9	6	3
A_3	6	5	7	A_3	9	5	10

Įrodykite, kad visų trijų faktorių sąveika $A \times B \times C$ yra lygi 0.

2.32. Nagrinėjame tiesinį modelį $Y_{ijk} = \mu_{ijk} + e_{ijk}, i = 1,...,I; j = 1,...,J; k = 1,...,K$, kuriame $\{e_{ijk}\}$ yra nepriklausomi normalieji $e_{ijk} \sim N(0,\sigma^2)$ a. d. Tarkime, vidurkiai μ_{ijk} tenkina sąlygą

$$\mu_{ijk} = \bar{\mu}_{...} + (\bar{\mu}_{i..} - \bar{\mu}_{...}) + (\bar{\mu}_{ij.} - \bar{\mu}_{i..}) + (\bar{\mu}_{..k} - \bar{\mu}_{...}) =$$

$$= \mu + \alpha_i + \beta_{ij} + \gamma_k.$$

Raskite parametrų $\mu, \alpha_i, \beta_{ij}, \gamma_k$ įvertinius. Sukurkite kriterijų hipotezei $H_A : \alpha_i = 0, i = 1,...,I$.

2.33. Sudarykite trifaktorės dispersinės analizės lentelę, kai visi trys faktoriai yra atsitiktiniai. Sukurkite kriterijus pagrindinėms hipotezėms tikrinti.

2.34. Tarkime, pilnoje dispersinės analizės schemoje,kai vienodas stebėjimų skaičius langelyje, faktorius A turi I lygmenų. Stebėjimų aritmetinius vidurkius, gautus suvidurki-

nus pagal visų kitų faktorių lygmenis, pažymėkime $\bar{Y}_1, ..., \bar{Y}_I$. Tegu šie vidurkiai padalyti į dvi didumo I_1 ir I_2, $I_1 + I_2 = I$ aibes $\bar{Y}_1, ..., \bar{Y}_{I_1}$ ir $\bar{Y}_{I_1+1}, ..., \bar{Y}_I$, o $\bar{Y}^{(1)}$ ir $\bar{Y}^{(2)}$ yra šių aibių vidurkiai. Įrodykite, kad

$$\sum_{i=1}^{I}(\bar{Y}_i - \bar{Y}_\cdot)^2 = \sum_{i=1}^{I_1}(\bar{Y}_i - \bar{Y}^{(1)})^2 + \sum_{i=I_1+1}^{I}(\bar{Y}_i - \bar{Y}^{(2)})^2 + I_1 I_2(\bar{Y}^{(1)} - \bar{Y}^{(2)})^2/I.$$

2.35. Apibendrinkite **2.34** pratimą, kai vidurkiai $\bar{Y}_1, ..., \bar{Y}_I$ padalijami į tris nesikertančias aibes.

2.36. Lentelėje pateikti duomenys, apibūdinantys betoninį kelią priklausomai nuo padengimo storio (faktorius A), nuo pagrindo storio (faktorius B) ir nuo papildomo (apatinio) pagrindo storio (faktorius C). Allikta po du matavimus kiekvieno iš 27 kelio variantų ([7]).

	A_1			A_2			A_3		
	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
C_1	2,8	4,3	5,7	4,1	5,4	6,7	6,0	6,3	7,1
	2,6	4,5	5,3	4,4	5,5	6,9	6,2	6,5	6,9
C_2	4,1	5,7	6,9	5,3	6,5	7,7	6,1	7,2	8,1
	4,4	5,8	7,1	5,1	6,7	7,4	5,8	7,1	8,4
C_3	5,5	7,0	8,1	6,5	7,7	8,8	7,0	8,0	9,1
	5,3	6,8	8,3	6,7	7,5	9,1	7,2	8,3	9,0

Atlikite trifaktorę analizę su trimis pastoviais faktoriais.

2.37. Lentelėje pateikiama izoliacijos kokybės charakteristikos priklausomai nuo keturių faktorių: A – padengimo tipas; B – temperatūra; C – slėgis; D – skirtingos plieninės panelės, kurios buvo naudojamos eksperimente [14].

	A_1	A_1	A_1	A_1	A_2	A_2	A_2	A_2
	B_1	B_1	B_2	B_2	B_1	B_1	B_2	B_2
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
D_1	0,25	0,16	0,30	0,27	0,41	0,10	0,13	0,06
D_2	0,36	0,002	0,18	0,03	0,28	0,04	0,06	0,03
D_3	0,36	0,06	0,44	0,13	0,33	0,03	0,19	0,04
D_4	0,25	0,10	0,34	0,04	0,21	0,01	0,20	0,01

	A_3	A_3	A_3	A_3	A_4	A_4	A_4	A_4
	B_1	B_1	B_2	B_2	B_1	B_1	B_2	B_2
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
D_1	0,44	0,24	0,22	0,18	0,43	0,27	0,26	0,21
D_2	0,65	0,08	0,14	0,36	0,62	0,03	0,51	0,03
D_3	0,42	0,49	0,17	0,25	0,47	0,28	0,21	0,25
D_4	0,47	0,14	0,36	0,19	0,52	0,07	0,32	0,38

Atlikite dispersinę analizę, tarę, kad visų keturių faktorių sąveika ir sąveikos po tris faktorius yra lygios 0.

2.8 - 2.11 skyreliai

2.38. Lentelėje pateikta keturių atspaudų charakteristikos (sąlyginiais vienetais) priklausomai nuo keturių galvučių skirtinguose spausdinimo įrenginiuose [14].

A_1				A_2				A_3			
1	2	3	4	5	6	7	8	9	10	11	12
6	13	1	7	10	2	4	0	0	10	8	7
2	3	10	4	9	1	1	3	0	11	5	2
0	9	0	7	7	1	7	4	5	6	0	5
8	8	6	9	12	10	9	1	5	7	7	4

A_4				A_5			
13	14	15	16	17	18	19	20
11	5	1	0	1	6	3	3
0	10	8	8	4	7	0	7
6	8	9	6	7	0	2	4
4	3	4	5	9	3	2	0

Atlikite dispersinę analizę, tarę, kad faktorius B (galvutės) yra sugrupuotas pagal faktorių A (spausdinimo įrenginiai); galvučių numeriai yra antroje lentelės eilutėje. Faktorių B laikyti atsitiktiniu, o skirtingas kopijas interpretuoti kaip matavimo kartotinumą.

2.39. Lentelėje pateikti duomenys apie kondensatorinio popieriaus poringumą priklausomai nuo partijos (faktorius A) ir nuo atsitiktinai iš partijos atrinkto rulono (faktorius B). Kiekvieną kartą atlikta po tris matavimus [7].

A_1				A_2				A_3			
1	2	3	4	5	6	7	8	9	10	11	12
1,5	1,5	2,7	3,0	1,9	2,3	1,8	1,9	2,5	3,2	1,4	7,8
1,7	1,6	1,9	2,4	1,5	2,4	2,9	3,5	2,9	5,5	1,5	5,2
1,6	1,7	2,0	2,6	2,1	2,4	4,7	2,8	3,3	7,1	3,4	5,0

Atlikite dispersinę analizę, tarę, kad faktorius B (rulonai) yra sugrupuotas pagal faktorių A (partijos); rulonų numeriai yra antroje lentelės eilutėje. Abu faktorius laikyti atsitiktiniais.

2.40. Atliekant eksperimentą 9 jūrų kiaulytės (faktorius B) atsitiktinai buvo suskirstytos į grupes po 3 ir patalpintos į skirtingus narvelius. Kiekvieno narvelio gyvūnai buvo aprūpinami skirtingai NO_2 lygiais (faktorius A); A_1 – kontrolinis; A_2 – dvigubai didesnis už normą; A_3 – trigubai didesnis už normą. Po savaitės buvo atlikta po du kintamojo Y (arterinis PH) matavimus [1].

	B_1		B_2		B_3	
A_1	7,08;	7,02	7,04;	7,07	7,07;	6,98
A_2	7,29;	7,18	7,42;	7,32	7,08;	7,28
A_3	7,74;	7,54	7,53;	7,50	7,51;	7,63

Ištirti kintamojo Y priklausomybę nuo faktoriaus A, tariant, kad faktorius B sugrupuotas pagal faktorių A.

2.41. Keturios pelės turi po 3 peliukus. Pelės atsitiktinai suskirstytos į 2 grupes po dvi. Pirmos grupės peliukams taikoma pirmoji dieta, o antrosios – antroji dieta. Po trijų savaičių užregistruotas peliukų svorio prieaugis Y [1].

Dieta	Pelė	Y			Dieta	Pelė	Y		
Pirmoji	1	11,8	10,5	12,5	Antroji	3	7,4	9,7	8,2
	2	12,3	15,5	11,4		4	7,2	8,6	7,1

Parinkite tinkamą dispersinės analizės schemą ir užpildykite dispersinės analizės lentelę. Patikrinkite pagrindines dispersinės analizės hipotezes.

2.42. Užpildykite trifaktorės dispersinės analizės lentelę, kai faktoriai A ir B dalyvauja eksperimente pagal kryžminės klasifikacijos schemą, o faktoriaus C lygmenys yra sugrupuoti pagal faktorių A ir B lygmenų kombinacijas.

2.43. Penkiolikai pacientų matuotas fermento kiekis iš karto po širdies operacijos (D_0), praėjus vienai dienai (D_1), dviem dienoms (D_2) ir savaitei (D_7) po operacijos.

Pacientas	D_0	D_1	D_2	D_7	Pacientas	D_0	D_1	D_2	D_7
1	108	63	45	42	9	106	65	49	49
2	112	75	56	52	10	110	70	46	47
3	114	75	51	46	11	120	85	60	62
4	129	87	69	69	12	118	78	51	56
5	115	71	52	54	13	110	65	46	47
6	122	80	68	68	14	132	92	73	63
7	105	71	52	54	15	127	90	73	68
8	117	77	54	61					

Patikrinkite, ar fermento lygis kinta po širdies operacijos. Atlikite kontrastų analizę.

2.44. Eksperimento metu buvo tirtas penkių rūšių dažų (faktorius A), naudojamų keliams ženklinti, tinkamumas. Atsitiktinai atrinkus aštuonias vietoves (faktorius B; blokas) ir taikant randomizuotą dažų naudojimo tvarką, kiekvienai vietovei buvo naudojama kiekviena dažų rūšis. Praėjus tam tikram laikui po nudažymo, buvo vertinamas dažų tinkamumas atsižvelgiant į atsparumo ir matomumo charakteristikas (kuo didesnė reikšmė, tuo geresnis įvertinimas) [10]. Atlikite blokuotųjų duomenų dispersinę analizę.

Vieta	Dažai				
	1	2	3	4	5
1	11	13	10	18	15
2	20	28	15	30	18
3	8	10	8	16	12
4	30	35	27	41	28
5	14	16	13	22	16
6	25	27	26	33	25
7	43	46	41	55	42
8	13	14	12	20	13

2.45. Sudarykite dispersinės analizės lentelę, kai eksperimentas atliekamas pagal lotyniškųjų kvadratų schemą ir visi trys faktoriai yra atsitiktiniai.

2.46. Tiriant penkių tipų (A, B, C, D, E) elektrodus eksperimento metu buvo pradeginta po 5 skylutes 5 metalo juostose. Ekaperimentas atliktas pagal lotyniškojo kvadrato schemą, kurioje eilutės atitinka metalo juostas, stulpeliai – skylutės padėtį metalo juostoje, o elektrodo tipas nurodytas skliausteliuose. Registruojama skylutės pradeginimo laikas [7].

3,5 (A)	2,1 (B)	2,5 (C)	3,5 (D)	2,4 (E)
2,6 (E)	3,3 (A)	2,1 (B)	2,5 (C)	2,7 (D)
2,9 (D)	2,6 (E)	3,5 (A)	2,7 (B)	2,9 (C)
2,5 (C)	2,9 (D)	3,0 (E)	3,3 (A)	2,3 (B)
2,1 (B)	2,3 (C)	3,7 (D)	3,2 (E)	3,5 (A)

Atlikite duomenų analizę ir užpildykite dispersinės analizės lentelę. Ar galima tvirtinti, kad elektrodų tipai skiriasi?

2.47. Tiriant tam tikros ankštinės kultūros šešių veislių (A, B, C, D, E, F) derlingumą, eksperimentas atliktas pagal lotyniškojo kvadrato schemą. Lentelėje nurodyti gauti derlingumo rodikliai, o kultūros veislė nurodyta skliausteliuose [14].

220 (B)	98 (F)	149 (D)	92 (A)	282 (E)	169 (C)
74 (A)	238 (E)	153 (B)	228 (C)	48 (F)	188 (D)
118 (D)	279 (C)	118 (F)	272 (E)	176 (B)	65 (A)
295 (E)	222 (B)	54 (A)	104 (D)	213 (C)	163 (F)
187 (C)	90 (D)	242 (E)	96 (F)	66 (A)	122 (B)
90 (F)	124 (A)	195 (C)	109 (B)	79 (D)	211 (E)

Atlikite duomenų dispersinę analizę, taip pat atlikite kultūrų porinius palyginimus naudodami T metodą.

2.48. Eksperimentiškai buvo tirta keturių gumos mišinių įtaka automobilių padangų atsparumui. Gamybos procese vienai padangai galima panaudoti iki trijų mišinių. Padanga (blokas) buvo suskirstyta į tris dalis (bloko dydis $k = 3$) ir skirtingi mišiniai naudoti skirtingose padangos dalyse. Eksperimentui naudotos keturios padangos. Tam tikru būdu buvo išmatuotas padangos nusidėvėjimo lygis [11].

Padanga	Gumos mišinys			
	1	2	3	4
1	238	238	279	–
2	196	213	–	308
3	254	–	334	367
4	–	312	421	412

Ar galima tvirtinti, kad gumos mišinių įtaka automobilių padangų atsparumui skiriasi?

2.49. Reikia palyginti televizorių ryškumo įvertinimą, gautą keturių operatorių (A, B, C, D). Per vieną dieną eksperimente gali dalyvauti tik trys operatoriai. Duomenys pateikti lentelėje [7].

Diena	Operatorius			
	A	B	C	D
1	780	820	800	–
2	950	–	920	940
3	–	880	880	820
4	840	780	–	820

Atlikite duomenų analizę. Ar galima tvirtinti, kad operatorių įvertinimai skiriasi?

2.50. Penkioms skalbimo priemonėms (A, B, C, D, E) palyginti buvo atliktas toks eksperimentas. Skalbimo priemonės buvo lyginamos plaunant specialiai užterštas lėkštes blokuose po tris rezervuarus su skirtingomis skalbimo priemonėmis. Lentelėje pateiktas išplautų vienodu skalbimo priemonės kiekiu lėkščių skaičius [14].

	1	2	3	4	5	6	7	8	9	10
A	27	28	30	31	29	30	–	–	–	–
B	26	26	29	–	–	–	30	21	26	–
C	30	–	–	34	32	–	34	31	–	33
D	–	29	–	33	–	34	31	–	33	31
E	–	–	26	–	24	25	–	23	24	26

Atlikite duomenų analizę. Ar galima tvirtinti, kad skalbimo priemonės skiriasi?

2.13. Atsakymai ir nurodymai

2.1. Tarkime, kad $d_I \neq 0$. Tada $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i d_i / d_I$. Pažymėję $\boldsymbol{Y} = (Y_{11}, ..., Y_{1J}, Y_{21}, ..., Y_{IJ})^T$ ir $\boldsymbol{\beta} = (\mu, \alpha_1, ..., \alpha_{I-1})^T$, gauname tiesinį modelį $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}$; $\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$. **2.2.** Hipotezė atmetama, kai $F = (SS_{EH} - SS_E)2(J-1)/SS_E > F_\alpha(2, 4(J-1))$; čia $SS_E = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$, $SS_{EH} - SS_E = J[(\bar{Y}_{1.} - \tilde{Y})^2 + (\bar{Y}_{2.} - \tilde{Y}/2)^2 + (\bar{Y}_{3.} - \tilde{Y}/3)^2]$, $\tilde{Y} = (6\bar{Y}_{1.} + 3\bar{Y}_{2.} + 2\bar{Y}_{3.})/11$. **2.3.** Statistika (2.1.9) įgijo reikšmę 3,2114; hipotezė atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > pv = \mathbf{P}\{F_{3;32} > 3,2114\} = 0,0359$. **2.4.** Statistika (2.1.9) įgijo reikšmę 9,9285; hipotezė atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > pv = \mathbf{P}\{F_{3;20} > 9,9285\} = 0,00032$. **2.5.** Statistika (2.1.9) įgijo reikšmę 1,3574; $\mathbf{P}\{F_{4;15} > 1,3574\} = 0,2950$; atmesti hipotezę nėra pagrindo. **2.6.** Statistika (2.1.9), kuri esant teisingai hipotezei turi Fišerio skirstinį su 2 ir 321 laisvės laipsniu, įgijo atitinkamai reikšmes: 1,538; 72,507; 31,009; 18,312; 23,904; 27,724. Duomenys neprieštarauja prielaidai, kad vario koncentracija yra vienoda; kitų elementų koncentracijų vienodumo hipotezės atmetamos kriterijais su aukštais reikšmingumo lygmenimis. **2.9.** a) Statistika, kuri esant teisingai

hipotezei turi Fišerio skirstinį su 1 ir 9 laisvės laipsniais, įgijo reikšmę 21,824; hipotezė atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{1,9} > 21,824\} = 0,0012$; b) $n \geq 32$. **2.10.** Gauname statistikų realizacijas $F_A = 2,48$, $F_B = 0,68$, $F_{AB} = 0,03$, kurias atitinka P reikšmės 0,0980, 0,4228, 0,9930. **2.11.** Tjukio kriterijaus statistikos realizacija F_γ įgijo reikšmę 0,7234; $\mathbf{P}\{F_{1;5} > 0,7234\} = 0,4339$; atmesti hipotezę nėra pagrindo. Statistika F_A įgijo reikšmę 9,01, o statistika F_B – reikšmę 4,61. Hipotezė H_A atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{3;6} > 9,01\} = 0,0122$. Hipotezė H_B atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{2;6} > 4,61\} = 0,0613$. **2.12.** a) Tjukio kriterijaus statistikos realizacija F_γ įgijo reikšmę 2,2754; $\mathbf{P}\{F_{1;34} > 2,2754\} = 0,1407$; atmesti hipotezę nėra pagrindo. Statistika F_A įgijo reikšmę 9,11, o statistika F_B – reikšmę 15,31. Hipotezės H_A ir H_B atmetamos kriterijumi su gana aukštu reikšmingumo lygmeniu. **2.13.** ≈ 1. **2.14.** $n \geq 12$. **2.15.** Gauname statistikų realizacijas $F_A = 8,31$, $F_B = 5,94$, $F_{AB} = 1,29$, kurias atitinka P reikšmės 0,000021, 0,0002, 0,2206. **2.16.** Statistika F_{AB} įgijo reikšmę 1,92, ją atitinka P reikšmė 0,0807; a) Gauname statistikų realizacijas $F_A = 4,54$, $F_B = 0,42$, kurias atitinka P reikšmės 0,0066, 0,7362; b) $F_A = 4,40$, $F_B = 0,26$; P reikšmės 0,0085, 0,8546. **2.18.** Hipotezė atmetama α lygmens kriterijumi, kai $F_A = MS_A/MS_E > (1 + J\Delta)F_\alpha(I-1, I(J-1))$. **2.19.** a) $F_A = 161,81$; b) $\hat{\mu} = 5,82$; $(\overline{\mu}; \underline{\mu}) = (4,38; 7,26)$; $\hat{\sigma}^2 = 0,0504$; $(\overline{\hat{\sigma}^2}; \underline{\hat{\sigma}^2}) = (0,033; 0,085)$; $\hat{\sigma}_A^2 = 0,8105$; $(\overline{\hat{\sigma}_A^2}; \underline{\hat{\sigma}_A^2}) = (0,257; 11,33)$; c) $(4,5; 227,3)$. **2.20.** Pagal pirmos eilutės duomenis: $\hat{\mu} = 53,17; \hat{\sigma}^2 = 1,144; \hat{\sigma}_A^2 = 11,715$; pagal antros eilutės duomenis: $\hat{\mu} = 52,26; \hat{\sigma}^2 = 2,537; \hat{\sigma}_A^2 = 13,133$; pagal trečios eilutės duomenis: $\hat{\mu} = 47,32; \hat{\sigma}^2 = 4,926; \hat{\sigma}_A^2 = 14,807$. **2.21.** Kriterijus išliks toks pat, tačiau kriterijaus galia bus išreiškiama centrinio Fišerio skirstinio pasiskirstymo funkcija. **2.22.** a) Hipotezė atmetama α lygmens kriterijumi, kai $F_A = MS_A/MS_{AB} > (1 + JK\Delta)F_\alpha(I-1, (I-1)(J-1))$; b) Hipotezė atmetama α lygmens kriterijumi, kai $F_{AB} = MS_{AB}/MS_E > (1 + K\Delta)F_\alpha((I-1)(J-1), IJ(K-1))$. **2.23.** a) Gauname $F_A = 21,196$, $F_B = 2439,686$, $F_{AB} = 28,846$; visos trys hipotezės atmetamos. b) $\hat{\sigma}^2 = 0,221$; $\hat{\sigma}_{AB}^2 = 0,559$; $\hat{\sigma}_A^2 = 2,926$; $\hat{\sigma}_B^2 = 56,533$; $\hat{\boldsymbol{V}}(\hat{\sigma}^2) = 0,000089$; $\hat{\boldsymbol{V}}(\hat{\sigma}_{AB}^2) = 0,00933$; $\hat{\boldsymbol{V}}(\hat{\sigma}_A^2) = 0,7865$; $\hat{\boldsymbol{V}}(\hat{\sigma}_B^2) = 2132,415$. c) Taikydami normaliąją aproksimaciją gauname parametrų $\sigma_A^2, \sigma_B^2, \sigma_{AB}^2, \sigma^2$ pasikliovimo intervalus: $(1,188; 4,664)$, $(0; 147,042)$, $(0,370; 0,748)$, $(0,203; 0,239)$. **2.24.** Statistika F_A įgijo reikšmę 9,01, o statistika F_B – reikšmę 4,61. Hipotezė H_A atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{3;6} > 9,01\} = 0,0122$. Hipotezė H_B atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{2;6} > 4,61\} = 0,0613$. **2.25.** Gauname statistikų realizacijas $F_A = 6,44$, $F_B = 4,60$, $F_{AB} = 1,29$, kurias atitinka P reikšmės 0,0017, 0,0059, 0,2206. **2.26.** a) Gauname statistikų realizacijas $F_A = 4,04$, $F_B = 2,63$, $F_{AB} = 2,46$, kurias atitinka P reikšmės 0,0613, 0,0537, 0,0534; b) $F_A = 9,92$, $F_B = 2,63$, $F_{AB} = 2,46$; P reikšmės 0,0005, 0,0537, 0,0534. **2.27.** Gauname statistikų realizacijas $F_A = 6,29$, $F_B = 1,05$, $F_{AB} = 2,39$, kurias atitinka P reikšmės 0,0097, 0,4116, 0,0087. **2.28.** Gauname statistikų realizacijas $F_A = 1,94$, $F_B = 0,49$, kurias atitinka P reikšmės 0,0801, 0,7813. **2.29.** $F_A = 2,48$, $F_B = 23,21$, $F_{AB} = 0,03$; P reikšmės 0,0980, 0,0170, 0,9930. **2.30.** Statistika F_A įgijo reikšmę 9,11, o statistika F_B – reikšmę 15,31. Hipotezės H_A ir H_B atmetamos kriterijumi su gana aukštu reikšmingumo lygmeniu. **2.32.** $\hat{\mu} = \bar{Y}_{...}$, $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$, $\hat{\beta}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..}$, $\hat{\gamma}_k = \bar{Y}_{..k} - \bar{Y}_{...}$; hipotezė H_A atmetama α lygmens kriterijumi, kai $(SS_{EH_A} - SS_E)(IJK - IJ - K + 1)/(I-1)SS_E > F_\alpha(I-1, (IJ-1)(K-1))$; $SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.} + \bar{Y}_{..k} - \bar{Y}_{...})^2$; $SS_{EH_A} = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{i..} + \bar{Y}_{..k})^2$. **2.33.** Hipotezė H_{ABC} atmetama, kai $MS_{ABC}/MS_E > F_\alpha((I-1)(J-1)(L-1), IJL(K-1))$; hipotezė H_{AB} atmetama, kai $MS_{AB}/MS_{ABC} > F_\alpha((I-1)(J-1), (I-1)(J-1)(L-1))$: analogiškai hipotezių H_{AC} ir H_{BC} atvejais; neegzistuoja tokie du MS, kad jų vidurkiai sutaptų esant teisingai hipotezei H_A; remiantis 2.7.2 pastaba hipotezė H_A atmetama apytiksliu Fišerio kriterijumi, kai $MS_A/(MS_{AB} - MS_{AC} - MS_{ABC}) > F_\alpha(I-1, \tilde{\nu})$; čia $\tilde{\nu} = (MS_{AB} + MS_{AC} - MS_{ABC})^2/[(MS_{AB})^2/((I-1)(J-1)) + (MS_{AC})^2/((I-1)(L-1)) + (MS_{ABC})^2/((I-1)(J-1)(L-1))]$; analogiškai tikrinamos hipotezės H_B ir H_C. **2.36.** Statistika F_A įgijo reikšmę 480,47, statistika F_B – reikšmę 903,91, statistika F_C – reikšmę 786,43, statistika F_{AB} – reikšmę 17,48, statistika F_{AC} – reikšmę 17,91. Hipotezės H_A, H_B, H_C, H_{AB} ir H_{AC} atmetamos kriterijumi su gana aukštu reikšmingumo lygmeniu. Statistika F_{BC} įgijo reikšmę 2,62, statistika F_{ABC} – reikšmę 4,72. Hipotezė H_{BC} atmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{4;27} > 2,62\} = 0,0571$. Hipotezė H_{ABC} atmetama

kriterijumi, kurio reikšmingumo lygmuo $\alpha > \mathbf{P}\{F_{8;27} > 4,72\} = 0,0011$. **2.37.** Gauname statistikų realizacijas $F_A = 9,12$, $F_B = 5,72$, $F_C = 46,50$, $F_D = 0,54$, kurias atitinka P reikšmės $0,0002$, $0,0227$, $8,85 \times 10^{-8}$, $0,6560$. Gauname statistikų realizacijas $F_{AB} = 1,44$, $F_{AC} = 0,79$, $F_{AD} = 0,34$, kurias atitinka P reikšmės $0,2475$, $0,5070$, $0,9539$. Gauname statistikų realizacijas $F_{BC} = 10,32$, $F_{BD} = 0,82$, $F_{CD} = 1,82$, kurias atitinka P reikšmės $0,0029$, $0,4897$, $0,1627$. **2.38.** $F_A = 0,60$, $F_{B(A)} = 1,76$; P reikšmės $0,6700$, $0,0625$. **2.39.** $F_A = 3,39$, $F_{B(A)} = 4,45$; P reikšmės $0,0798$, $0,0016$. **2.40.** $F_A = 46,78$, $F_{B(A)} = 1,27$; P reikšmės $0,0002$, $0,3590$. **2.41.** Hierarchinės klasifikacijos modelis. Faktoriaus B (pelė) lygmenys yra sugrupuoti pagal faktoriaus A (dieta) lygmenis. Abu faktoriai pastovūs. Gauname statistikų realizacijas $F_A = 28,67$, $F_{B(A)} = 1,08$, kurias atitinka P reikšmės $0,0007$, $0,3838$. **2.43.** Statistikos realizacija 1301,7. Prielaida, kad fermento lygis nekinta po širdies operacijos atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu. **2.44.** Faktorius A (dažų rūšis) yra fiksuotasis. Blokus (faktorius B; atsitiktinis) sudaro penki dažų tinkamumo įvertinimai atlikti toje pačioje vietovėje. Gauname statistikų realizacijas $F_A = 157,92$, $F_B = 30,42$. Hipotezės H_A ir H_B atmetamos kriterijumi su gana aukštu reikšmingumo lygmeniu. **2.46.** Gauname statistikos realizaciją 15,83. Hipotezė, kad elektrodų tipai vienodi, atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu. **2.47.** Gauname statistikos realizaciją 19,63. Hipotezė, kad veislių derlingumas vienodas, atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu. **2.48.** Gumos mišinio įtaka padangų atsparumui yra statistiškai reikšminga (p reikšmė 0,0034). **2.49.** Gauname statistikų realizacijas $F_A = 0,17$ (operatorius), $F_B = 10,61$ (diena), kurias atitinka P reikšmės $0,9131$, $0,0132$. **2.50.** Gauname statistikos realizaciją 27,24. Hipotezė, kad skalbimo priemonės vienodos, atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu.

3 skyrius

Regresinė analizė

Regresinėje analizėje sprendžiami vienų kintamųjų (priklausomų kintamųjų) $Y_1, ..., Y_r$ prognozavimo remiantis kitais kintamaisiais (nepriklausomais ar paaiškinančiaisiais kintamaisiais, kovariantėmis, regresoriais) $X_1, ..., X_m$ uždaviniai. Prognozavimas reikalingas daugelyje praktinių situacijų. Pavyzdžiui, meteorologas prognozuoja ateinančio laikotarpio orus, remdamasis tam tikrų atmosferos požymių matavimais praeityje. Gydytojas prognozuoja ligos eigą, remdamasis tam tikrų simptomų matavimais. Pardavėjas prognozuoja tam tikros markės automobilių kainą atsižvelgdamas į jo gamybos metus, variklio galingumą, nuvažiuotą kelią ir pan. Istoriškai pirmieji regresinės analizės uždaviniai pradėti spręsti pačioje dvidešimtojo amžiaus pradžioje, prognozuojant vaikų ūgį pagal tėvų ūgį bei žemės ūkyje prognozuojant derlių pagal naudojamų trąšų kiekį, dirvožemio tipą ir kitus rodiklius. Paprastai prognozė nebūna visiškai tiksli, nes nagrinėjamo a. d. skirstinys priklauso ne tik nuo stebimų kovariančių, bet ir nuo didelio kitų faktorių, į kuriuos atsižvelgti nėra galimybės, skaičiaus. Todėl prognozės tikslumą galima apibūdinti tik tikimybiškai. Plačiau apie regresinę analizę žr. [2], [10], [16].

3.1. Teoriniai regresijos pagrindai

Šiame skyrelyje aptarsime teorinius prognozės kūrimo ir jos tikslumo įvertinimo klausimus. Tolesniuose skyreliuose aptarsime statistinius regresijos uždavinius, kai prognozė ir jos tikslumas yra vertinami remiantis statistiniais duomenimis.

3.1.1. Optimalioji prognozė

Tarkime, kad norime prognozuoti a. d. Y, remdamiesi a. v. $\mathbf{X} = (X_1, ..., X_m)^T$ su vidurkiu $\mathbf{E}\mathbf{X} = \boldsymbol{\mu} = (\mu_1, ..., \mu_m)^T$ ir baigtine kovariacine matrica

$$\boldsymbol{V}(\mathbf{X}) = \boldsymbol{\Sigma} = [\sigma_{ij}]_{k\times k}, \quad \sigma_{ij} = \mathbf{Cov}\,(X_i,\ X_j), \quad i, j = 1, ..., m.$$

Kiekvieną a. d. $\hat{Y} = h(\mathbf{X})$, čia h yra mačioji funkcija, vadinsime a. d. Y *prognoze*. Skirtumas $Y - h(\mathbf{X})$ vadinamas *prognozės paklaida*, o vidurkis $\mathbf{E}(Y - h(\mathbf{X}))^2$ – *vidutine kvadratine paklaida*.

Visų prognozių, tenkinančių sąlygą $\boldsymbol{V}(h(\mathbf{X})) < \infty$, aibę žymėsime $\mathcal{H}$.

Natūralu ieškoti prognozės, kuri optimaliai prognozuoja a. d. Y.

3.1.1 apibrėžimas. A. d. Y prognozė $\hat{Y} = h(\mathbf{X})$ vadinama *optimaliąja*, jei

$$\mathbf{E}(Y - h(\mathbf{X}))^2 = \min_{\tilde{h}(\boldsymbol{X}) \in \mathcal{H}} \mathbf{E}(Y - \tilde{h}(\mathbf{X}))^2. \tag{3.1.1}$$

3.1.2 apibrėžimas. Sąlyginis vidurkis

$$\mu(\boldsymbol{x}) = \mu(x_1, ..., x_m) = \mathbf{E}(Y|\mathbf{X} = \boldsymbol{x}) \tag{3.1.2}$$

vadinamas a. d. Y *regresija* a. v. $\mathbf{X}$ *atžvilgiu*.

Toliau žymėsime $h = h(\mathbf{X})$, $\mu = \mu(\mathbf{X})$.

3.1.1 teorema. *Regresija $\mu = \mu(\mathbf{X})$ yra optimalioji a.d. Y prognozė.*

Įrodymas. Gauname

$$\mathbf{E}(Y - h)^2 = \mathbf{E}(Y - \mu + \mu - h)^2 = \mathbf{E}(Y - \mu)^2 + \mathbf{E}(\mu - h)^2 \geq \mathbf{E}(Y - \mu)^2,$$

nes

$$\mathbf{E}((Y - \mu)(\mu - h)) = \mathbf{E}\{\mathbf{E}((Y - \mu)(\mu - h)|\mathbf{X})\} = \mathbf{E}\{(\mu - h)\mathbf{E}[(Y - \mu)|\mathbf{X}]\} = 0.$$

▲

Pažymėkime

$$\rho(Y,\ h) = \frac{\mathbf{Cov}\,(Y,\ h)}{\sqrt{\boldsymbol{V}Y\boldsymbol{V}h}} \tag{3.1.3}$$

a. d. Y ir h koreliacijos koeficientą.

3.1.2 teorema. *Koreliacijos koeficientas $\rho(Y,\ \mu)$ neneigiamas ir turi pavidalą*

$$\rho(Y,\ \mu) = \sqrt{\frac{\boldsymbol{V}\mu}{\boldsymbol{V}Y}}. \tag{3.1.4}$$

Optimalioji prognozė μ maksimizuoja koreliacijos koeficientą $\rho(Y, h)$ visų prognozių $h \in \mathcal{H}$ klasėje.

Įrodymas. Su kiekvienu $h \in \mathcal{H}$ turime

$$\mathbf{Cov}\,(Y, h) = \mathbf{E}[(h - \mathbf{E}h)\mathbf{E}((Y - \mathbf{E}Y)|\mathbf{X})] = \mathbf{Cov}\,(\mu, h).$$

Jeigu $h = \mu$, tai $\mathbf{Cov}\,(Y,\ \mu) = \mathbf{Cov}\,(\mu,\ \mu) = \boldsymbol{V}\mu$. Taigi teisinga (3.1.4). Gauname

$$\rho^2(Y,\ h) = \frac{\mathbf{Cov}^{\,2}(Y, h)}{\boldsymbol{V}Y\boldsymbol{V}h}\,\frac{\boldsymbol{V}\mu}{\boldsymbol{V}\mu} = \frac{\mathbf{Cov}^{\,2}(\mu, h)}{\boldsymbol{V}h\boldsymbol{V}\mu}\,\frac{\boldsymbol{V}\mu}{\boldsymbol{V}Y} =$$

$$= \rho^2(\mu, h)\,\rho^2(Y,\ \mu) \le \rho^2(Y,\ \mu).$$

▲

3.1.3 apibrėžimas. Koreliacijos koeficiento $\rho(Y,\mu)$ kvadratas vadinamas *koreliaciniu santykiu* ir žymimas

$$\eta^2_{Y\boldsymbol{X}} = \eta^2_{Y(X_1,\dots,X_m)} = \rho^2(Y,\ \mu) = \frac{\boldsymbol{V}\mu}{\boldsymbol{V}Y}. \tag{3.1.5}$$

Kadangi $Y-\mu$ ir μ nekoreliuoti, tai teisingas dispersijos išskaidymas

$$\boldsymbol{V}Y = \boldsymbol{V}\mu + \boldsymbol{V}(Y-\mu), \tag{3.1.6}$$

todėl

$$1-\eta^2_{Y\boldsymbol{X}} = \frac{\boldsymbol{V}(Y-\mu)}{\boldsymbol{V}Y} \tag{3.1.7}$$

3.1.1 pastaba Iš (3.1.5) ir (3.1.7) išplaukia, kad koreliacinis santykis rodo, kurią dispersijos $\boldsymbol{V}Y$ dalį sudaro optimalios Y prognozės μ dispersija. Prognozės paklaidos $Y-\mu$ dispersija (*liekamoji dispersija*) sudaro $1-\eta^2_{Y\boldsymbol{X}}$ dispersijos $\boldsymbol{V}Y$ dalį.

Jeigu $\eta^2_{Y\boldsymbol{X}}=1$, tai Y ir $\mathbf{X}$ susieti funkcine priklausomybe, nes $\boldsymbol{V}(Y-\mu)=0 \iff Y=\mu(\mathbf{X})$; jeigu $\eta^2_{Y\boldsymbol{X}}=0$, tai prognozuoti Y pagal $\mathbf{X}$ neturi prasmės, nes $\boldsymbol{V}(\mu(\mathbf{X}))=0 \iff \mu(\mathbf{X})=\mathbf{E}Y$, taigi prognozė nepriklauso nuo $\mathbf{X}$.

Taigi koreliacinis santykis apibūdina optimalios prognozės $\mu(\mathbf{X})$ tikslumą. Kuo jis artimesnis vienetui, tuo tikslesnė prognozė.

3.1.2. Tiesinė prognozė

Nagrinėkime a. d. Y tiesines, t. y. $l=l(\mathbf{X})=\alpha+\boldsymbol{\beta}^T\mathbf{X}=\alpha+\beta_1X_1+\dots+\beta_mX_m$ pavidalo prognozes. *Optimalioji tiesinė prognozė* $l^*=\alpha^*+(\beta^*)^T\mathbf{X}$ minimizuoja vidurkį

$$SS(\alpha,\boldsymbol{\beta}) = \mathbf{E}(Y-\alpha-\boldsymbol{\beta}^T\mathbf{X})^2 \tag{3.1.8}$$

parametrų α ir $\boldsymbol{\beta}=(\beta_1,\dots,\beta_m)^T$ atžvilgiu.

Pažymėkime

$$\sigma^2_Y = \boldsymbol{V}Y,\ \ \boldsymbol{\sigma}_{Y\boldsymbol{X}} = \mathbf{Cov}\,(Y,\ \mathbf{X}),\ \ \boldsymbol{\Sigma}=\boldsymbol{V}(\mathbf{X}). \tag{3.1.9}$$

Tarsime, kad kovariacinė matrica $\boldsymbol{\Sigma}$ neišsigimusi. Kiekvienos tiesinės prognozės dispersija yra (žr. 2 priedą (7.3.4))

$$\boldsymbol{V}(l) = \boldsymbol{V}(\boldsymbol{\beta}^T\mathbf{X}) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \tag{3.1.10}$$

3.1.3 teorema. *Optimali tiesinė Y prognozė turi pavidalą $l^*=\alpha^*+\mathbf{X}^T\boldsymbol{\beta}^*$, čia*

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\sigma_{Y\boldsymbol{X}},\ \ \alpha^* = \mathbf{E}Y-(\boldsymbol{\beta}^*)^T\mathbf{E}(\mathbf{X}). \tag{3.1.11}$$

Kvadratinės formos (3.1.8) minimumas

$$SS(\alpha^*, \boldsymbol{\beta}^*) = \sigma_Y^2 - \boldsymbol{\sigma}_{Y\boldsymbol{X}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{Y\boldsymbol{X}}. \tag{3.1.12}$$

A. d. Y ir tiesinės prognozės $l(\mathbf{X})$ koreliacijos koeficientą yra maksimalus, kai $l = l^*$ ir

$$\rho^2(Y,\, l^*) = \frac{\boldsymbol{V}(l^*)}{\sigma_Y^2} = \frac{(\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*}{\sigma_Y^2} = \frac{\boldsymbol{\sigma}_{Y\boldsymbol{X}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{Y\boldsymbol{X}}}{\sigma_Y^2}. \tag{3.1.13}$$

Įrodymas. Gauname

$$SS(\alpha, \boldsymbol{\beta}) = \mathbf{E}(Y - \alpha - \boldsymbol{\beta}^T \mathbf{X})^2 = \mathbf{E}[(Y - \mathbf{E}Y) + (\mathbf{E}Y - \alpha - \beta^T \mathbf{E}\mathbf{X}) - \beta^T (\mathbf{X} - \mathbf{E}\mathbf{X})]^2$$

$$= \boldsymbol{V}(Y) + (\mathbf{E}Y - \alpha - \beta^T \mathbf{E}\mathbf{X})^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \boldsymbol{\sigma}_{Y\boldsymbol{X}}$$

$$\geq \boldsymbol{V}(Y) + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \boldsymbol{\sigma}_{Y\boldsymbol{X}} =: g(\boldsymbol{\beta}).$$

Kadangi $\dot{g}(\boldsymbol{\beta}) = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T \boldsymbol{\sigma}_{Y\boldsymbol{X}}$, matrica $\ddot{g}(\boldsymbol{\beta}) = 2\boldsymbol{\Sigma}$ teigiamai apibrėžta, tai lygties $\dot{g}(\boldsymbol{\beta}) = \mathbf{0}$ sprendinys $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{Y\boldsymbol{X}}$ minimizuoja funkciją $g(\boldsymbol{\beta})$. Jei $\alpha = \alpha^*, \boldsymbol{\beta} = \boldsymbol{\beta}^*$, tai nelygybė virsta lygybe. Taigi $(\alpha^*, \boldsymbol{\beta}^*)$ minimizuoja $SS(\alpha, \boldsymbol{\beta})$. Gauname

$$SS(\alpha^*, \boldsymbol{\beta}^*) = \sigma_Y^2 - (\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^* - 2\boldsymbol{\beta}^{*T} \boldsymbol{\sigma}_{Y\boldsymbol{X}} = \sigma_Y^2 - \boldsymbol{\sigma}_{Y\boldsymbol{X}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{Y\boldsymbol{X}}.$$

Ieškodami koreliacijos koeficiento $\rho(Y, l^*)$, pažymėsime, kad $\boldsymbol{\sigma}_{Y\boldsymbol{X}} = \boldsymbol{\Sigma}\boldsymbol{\beta}^*$,

$$\mathbf{Cov}\,(Y,\, \boldsymbol{\beta}^T \mathbf{X}) = \boldsymbol{\beta}^T \boldsymbol{\sigma}_{Y\boldsymbol{X}} = \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*,$$

$$\mathbf{Cov}\,(Y,\, (\boldsymbol{\beta}^*)^T \mathbf{X}) = (\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^* = \boldsymbol{V}((\boldsymbol{\beta}^*)^T \mathbf{X}).$$

Naudodami Koši nelygybę, turime

$$\rho^2(Y,\, \boldsymbol{\beta}^T \mathbf{X}) = \frac{[\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*]^2}{\sigma_Y^2 \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \leq \frac{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}\, (\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*}{\sigma_Y^2 \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} =$$

$$\frac{[(\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*]^2}{\sigma_Y^2 (\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*} = \rho^2(Y,\, (\boldsymbol{\beta}^*)^T \mathbf{X}).$$

▲

3.1.4 apibrėžimas. Koreliacijos koeficiento kvadratas $\rho^2(Y,\, l^*)$ žymimas

$$\rho_{Y\boldsymbol{X}}^2 = \frac{\boldsymbol{V}(l^*)}{\sigma_Y^2} = \frac{(\boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} \boldsymbol{\beta}^*}{\sigma_Y^2} = \frac{\boldsymbol{\sigma}_{Y\boldsymbol{X}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{Y\boldsymbol{X}}}{\sigma_Y^2}, \tag{3.1.14}$$

o $\rho_{Y\boldsymbol{X}} = \sqrt{\rho_{Y\boldsymbol{X}}^2} \geq 0$ vadinamas *dauginiu koreliacijos koeficientu.*

Kadangi

$$\mathbf{Cov}\,((\boldsymbol{\beta}^*)^T \mathbf{X},\, Y - (\boldsymbol{\beta}^*)^T \mathbf{X}) = 0,$$

tai a. d. Y dispersija tenkina lygybę

$$\boldsymbol{V}Y = \sigma_Y^2 = \boldsymbol{V}(\mathbf{X}^T\boldsymbol{\beta}^*) + \boldsymbol{V}(Y - \mathbf{X}^T\boldsymbol{\beta}^*) = \sigma_Y^2\rho_{Y\boldsymbol{X}}^2 + \sigma_Y^2(1 - \rho_{Y\boldsymbol{X}}^2),$$

todėl

$$1 - \rho_{Y\boldsymbol{X}}^2 = \frac{\boldsymbol{V}(Y - l^*)}{\boldsymbol{V}Y}. \tag{3.1.15}$$

3.1.2 pastaba Iš (3.1.14) ir (3.1.15) išeina, kad dauginio koreliacijos koeficiento kvadratas $\rho_{Y\boldsymbol{X}}^2$ rodo, kurią dispersijos $\boldsymbol{V}Y$ dalį sudaro optimalios tiesinės Y prognozės l^* dispersija. Optimalios tiesinės prognozės paklaidos $Y - l^*$ dispersija sudaro $1 - \rho_{Y\boldsymbol{X}}^2$ dispersijos $\boldsymbol{V}Y$ dalį.

Jeigu $\rho_{Y\boldsymbol{X}}^2 = 1$, tai Y ir $\mathbf{X}$ susieti tiesine priklausomybe; jeigu $\rho_{Y\boldsymbol{X}}^2 = 0$, tai tiesinis Y prognozavimas naudojantis $\mathbf{X}$ neturi prasmės.

3.1.3 pastaba. Iš (3.1.14) išplaukia, kad geriausioji tiesinė prognozė l^* ir jos tikslumo matas (3.1.14) išreiškiami a. v. $(Y,\ X_1, ..., X_k)^T$ pirmųjų ir antrųjų momentų terminais.

3.1.4 pastaba. Jeigu $k = 1$, tai $\rho_{Y(X_1)}^2$ yra a. d. Y ir X_1 įprastinio koreliacijos koeficiento kvadratas.

3.1.5 pastaba. Kadangi optimali tiesinė prognozė minimizuoja liekamąją dispersiją siauresnėje klasėje negu optimali nebūtinai tiesinė prognozė, tai galioja nelygybės

$$0 \le \rho_{Y\boldsymbol{X}}^2 \le \eta_{Y\boldsymbol{X}}^2 \le 1. \tag{3.1.16}$$

Galime pateikti pavyzdžius, kai $\eta_{Y\boldsymbol{X}}^2 = 1$, tuo tarpu $\rho_{Y\boldsymbol{X}}^2 = 0$ (žr., pvz., 3.4 pratimą). Tai reiškia, kad Y ir $\mathbf{X}$ gali būti susieti funkcine priklausomybe, tačiau tiesinis prognozavimas pagal $\mathbf{X}$ neturi prasmės.

Skirtumas $\eta_{Y\boldsymbol{X}}^2 - \rho_{Y\boldsymbol{X}}^2$ rodo, kiek galima pagerinti prognozavimą vietoje tiesinių $\mathbf{X}$ funkcijų pasirenkant funkcijas iš platesnės klasės.

3.1.6 pastaba. Jei a.v. $(Y, \mathbf{X}^T)^T$ turi daugiamatį normalųjį skirstinį, tai tiesinė prognozė negali būti pagerinta, t. y. $\eta_{Y\boldsymbol{X}}^2 = \rho_{Y\boldsymbol{X}}^2$. Tai yra todėl, kad normaliojo vektoriaus bet kurios koordinatės sąlyginis vidurkis kitų koordinačių atžvilgiu (regresija) yra tiesinė tų koordinačių funkcija (žr. 2 priedą (7.4.7)).

3.1.3. Papildomos priklausomybės apibūdinimas

Dažnai svarbu ištirti, kiek pagerėja Y prognozės tikslumas padidinus a. v. $\mathbf{X}$ dimensiją.

Suskaidykime a. v. $\mathbf{X}$ į dvi komponentes $\mathbf{X}^{(1)} = (X_1, ..., X_k)^T$ ir $\mathbf{X}^{(2)} = (X_{k+1}, ..., X_m)^T$. Prognozuojant a. d. Y pagal $\mathbf{X}^{(1)}$ paklaidos dispersija yra $\sigma_Y^2(1 - \eta_{Y\boldsymbol{X}^{(1)}}^2)$, o prognozuojant pagal $\mathbf{X}$ paklaidos dispersija yra $\sigma_Y^2(1 - \eta_{Y\boldsymbol{X}})$. Papildomą priklausomybę galime apibūdinti *paklaidos dispersijos santykiniu sumažėjimu* pridėjus prie $\mathbf{X}^{(1)}$ a. v. $\mathbf{X}^{(2)}$.

3.1.5 apibrėžimas. Prognozės paklaidos dispersijos santykinis sumažėjimas žymimas

$$\eta^2_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}} = \frac{\eta^2_{Y\boldsymbol{X}} - \eta^2_{Y\boldsymbol{X}^{(1)}}}{1-\eta^2_{Y\boldsymbol{X}^{(1)}}}. \tag{3.1.17}$$

o dydis $\eta_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)})} = \sqrt{\eta^2_{Y\boldsymbol{X}\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}}} \geq 0$ vadinamas *daliniu koreliaciniu santykiu.*

Iš vieneto atėmę abi (3.1.17) lygybės puses ir padauginę gautos lygybės abi puses iš $1-\eta^2_{Y\boldsymbol{X}^{(1)}}$, gauname

$$1-\eta^2_{Y\boldsymbol{X}} = (1-\eta^2_{Y\boldsymbol{X}^{(1)}})(1-\eta^2_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)})}), \tag{3.1.18}$$

iš čia gaunama lygybė

$$1-\eta^2_{Y(X_1,\ldots,X_m)} = \prod_{j=2}^{m}(1-\eta^2_{YX_j|(X_1,\ldots,X_{j-1})})(1-\eta^2_{YX_1}). \tag{3.1.19}$$

3.1.7 pastaba. Galimas ir kitas dalinio koreliacinio santykio interpretavimas. Prognozuojant Y pagal $\mathbf{X}^{(1)} = (X_1,\ldots,X_k)^T$, lieka paklaida $Y-\mu_1(\mathbf{X}^{(1)})$. Dalinis koreliacinis santykis $\eta^2_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}}$ parodo, kokią šios atsitiktinės paklaidos sklaidos dalį paaiškina naujų kintamųjų $X_{k+1},\ldots,X_m$ pridėjimas:

$$\eta^2_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}} = \eta^2_{Y-\mu_1(\boldsymbol{X}^{(1)}),\boldsymbol{X}} = \frac{\boldsymbol{V}(\mathbf{E}(Y-\mu_1(\mathbf{X}^{(1)})|\mathbf{X}))}{\boldsymbol{V}(Y-\mu_1(\mathbf{X}^{(1)}))}, \tag{3.1.20}$$

taigi dalinis koreliacinis santykis yra tiesiog koreliacinis santykis prognozuojant paklaidą $Y-\mu_1(\mathbf{X}^{(1)})$ pagal visą a. v. $\mathbf{X}$, kartu tai yra Y ir $\mathbf{X}^{(2)}$ priklausomybės matas, eliminavus $\mathbf{X}^{(1)}$ įtaką.

Įrodymas. Pažymėkime $Z = Y-\mu_1(\mathbf{X}^{(1)})$. Turime

$$\mathbf{E}(Z|\mathbf{X}) = \mu(\mathbf{X}) - \mu_1(\mathbf{X}^{(1)}), \quad \mu(\mathbf{X}) = \mathbf{E}(Y|\mathbf{X}).$$

Pastebėkime, kad

$$\mu_1(\mathbf{X}^{(1)}) = \mathbf{E}(Y \mid \mathbf{X}^{(1)}) = \mathbf{E}(\mathbf{E}(Y \mid \mathbf{X}) \mid \mathbf{X}^{(1)}) = \mathbf{E}(\mu(\mathbf{X}) \mid \mathbf{X}^{(1)}),$$

todėl

$$\mathbf{Cov}\,(\mu,\mu_1) = \mathbf{E}(\mathbf{E}(\mu(\mathbf{X})-\mathbf{E}Y)(\mu_1(\mathbf{X}^{(1)})-\mathbf{E}Y) \mid \mathbf{X}^{(1)})) = \boldsymbol{V}(\mu_1(\mathbf{X}^{(1)})).$$

Taigi

$$\boldsymbol{V}(\mathbf{E}(Z|\mathbf{X})) = \sigma^2_Y\eta^2_{Y\boldsymbol{X}} - \sigma^2_Y\eta^2_{Y\boldsymbol{X}^{(1)}},$$

$$\boldsymbol{V}(Z) = \boldsymbol{V}(Y-\mu_1) = \boldsymbol{V}(Y) + \boldsymbol{V}(\mu_1) - 2\mathbf{Cov}\,(Y,\mu_1) = \sigma^2_Y - \sigma^2_Y\eta^2_{Y\boldsymbol{X}^{(1)}}.$$

Iš paskutinių dviejų lygybių ir 3.1.5 apibrėžimo išplaukia (3.1.18).

▲

3.1.6 apibrėžimas. Tiesinės prognozės paklaidos dispersijos santykinis sumažėjimas žymimas

$$\rho^2_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}} = \frac{\rho^2_{Y\boldsymbol{X}} - \rho^2_{Y\boldsymbol{X}^{(1)}}}{1-\rho^2_{Y\boldsymbol{X}^{(1)}}}, \tag{3.1.21}$$

o dydis $\rho_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}} = \sqrt{\rho^2_{Y\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(1)}}} \geq 0$ vadinamas *daliniu koreliacijos koeficientu.*

Analogiškai daliniam koreliaciniam santykiui (3.1.7 pastaba) tai yra tiesinės Y ir $\mathbf{X}^{(2)}$ priklausomybės matas, eliminavus $\mathbf{X}^{(1)}$ įtaką. Jis gali žymiai skirtis nuo dauginio Y ir $\mathbf{X}^{(2)}$ koreliacijos koeficientą, kurį skaičiuojant $\mathbf{X}^{(1)}$ įtaka neeliminuota.

Analogiškai koreliaciniam santykiui dauginis koreliacijos koeficientas išreiškiamas daliniais koeficientais:

$$1-\rho^2_{Y(X_1,\ldots,X_m)} = \prod_{j=2}^{m}(1-\rho^2_{YX_j|(X_1,\ldots,X_{j-1})})(1-\rho^2_{YX_1}). \tag{3.1.22}$$

Kuo daugiau prognozuojančių a. d., tuo didesnis dauginis koreliacijos koeficientas.

3.2. Tiesinė vieno kintamojo regresija

3.2.1. Statistinis modelis

Tarkime, kad turime prognozuoti a. d. Y remdamiesi vienmate kovariante X. Fiksavus kovariantės reikšmes $x_1,\ldots,x_n$ gauta imtis $Y_1,\ldots,Y_n$. Darome prielaidą, kad dydžiai x_i yra neatsitiktiniai arba yra nepriklausomų vienodai pasiskirsčiusių a. d. $X_1,\ldots,X_n$ realizacijos. Pastaruoju atveju analizė yra sąlyginė, naudojamasi tik šiomis realizacijomis, bet ne a. d. X_i skirstiniais. Tiesiniame regresijos modelyje daroma prielaida, kad a. d. Y sąlyginis vidurkis žinant kovariantę, yra tiesinė funkcija, o sąlyginė dispersija nepriklauso nuo kovariantės.

Nagrinėsime tiesinį modelį:

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \tag{3.2.1}$$

čia $e_1,\ldots,e_n$ yra nepriklausomi vienodai pasiskirstę a. d., $\mathbf{E}(e_i)=0$, $\boldsymbol{V}(e_i)=\sigma^2$.

Patogu modelį modifikuoti tokiu pavidalu:

$$Y_i = \alpha + \beta(x_i - \bar{x}) + e_i, \tag{3.2.2}$$

čia

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \beta = \beta_1, \quad \alpha = \beta_0 + \beta_1\bar{x},$$

nes, kaip matysime, tada parametrų α ir β MK įvertiniai nekoreliuoti.

Iš (3.2.1) gauname, kad vidurkiai

$$\mu(x_i) = \mathbf{E}(Y|x_i) = \mathbf{E}Y_i = \alpha + \beta(x_i - \bar{x}) = \beta_0 + \beta_1 x_i, \quad i = 1, ..., n,$$

yra tiesinės x_i funkcijos.

Jei x_i yra a. d. X realizacijos, tai vidurkiai $\mu(x_i)$ yra Y regresijos $\mu(X) = \mathbf{E}(Y \mid \mathbf{X})$ (kuri yra optimalus Y prediktorius) realizacijos. Modelyje tariama, kad regresija tiesinė, taigi, atsižvelgiant į 3.1.6 pastabą, jis ypač rekomenduotinas, kai (Y, X) skirstinys yra normalusis.

3.2.2. Parametrų įvertiniai

Pažymėję $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{\beta} = (\alpha, \beta)^T$, $\boldsymbol{e} = (e_1, \ldots, e_n)^T$, gauname tiesinį modelį

$$Y = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}$$

su plano matrica $\boldsymbol{A}$, turinčia n eilučių ir 2 stulpelius:

$$\boldsymbol{A}^T = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 - \bar{x} & \cdots & x_n - \bar{x} \end{pmatrix}, \quad \boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix}. \quad (3.2.3)$$

Taškinį parametro $\boldsymbol{\beta}$ įvertinį $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\beta})$ gauname mažiausiųjų kvadratų metodu, minimizuodami kvadratinę formą

$$SS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \alpha - \beta(x_i - \bar{x}))^2.$$

Šis įvertinys turi pavidalą (žr. (1.2.3) formulę):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T Y,$$

o kvadratų sumos minimumas

$$SS_E = \min_{\boldsymbol{\beta}} SS(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2. \quad (3.2.4)$$

Pažymėkime

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

3.2.1 teorema. *Nežinomų parametrų α ir β MK įvertiniai turi pavidalą*

$$\hat{\alpha} = \bar{Y}, \quad \hat{\beta} = s_{xy}/s_x^2.$$

Jie nepaslinktieji, o jų antrieji momentai yra

$$\boldsymbol{V}\hat{\alpha} = \frac{\sigma^2}{n}, \quad \boldsymbol{V}\hat{\beta} = \frac{\sigma^2}{(n-1)s_x^2}, \quad \mathbf{Cov}\,(\hat{\alpha}, \hat{\beta}) = 0. \quad (3.2.5)$$

Nepriklausantis nuo $\hat{\boldsymbol{\beta}}$ dispersijos σ^2 įvertinys yra

$$s^2 = MS_E = \frac{SS_E}{n-2}. \tag{3.2.6}$$

Nepaslinktieji įvertinių $\hat{\alpha}$ ir $\hat{\beta}$ dispersijų įvertiniai yra

$$s_\alpha^2 = \hat{\boldsymbol{V}}\hat{\alpha} = \frac{s^2}{n}, \quad s_\beta^2 = \hat{\boldsymbol{V}}\hat{\beta} = \frac{s^2}{(n-1)s_x^2}.$$

Įrodymas. Pirmasis teiginys išplaukia iš lygybių

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} 1/n & 0 \\ 0 & 1/\sum_{i=1}^n (x_i-\bar{x})^2 \end{pmatrix} \begin{pmatrix} \sum_i Y_i \\ \sum_i Y_i(x_i-\bar{x}) \end{pmatrix}$$
$$= \begin{pmatrix} \bar{Y} \\ s_{xy}/s_x^2 \end{pmatrix}.$$

Pasinaudoję tuo, kad

$$\mathbf{Cov}\,(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i(x_i-\bar{x})) = \sigma^2 \sum_{i=1}^n (x_i-\bar{x}) = 0, \quad \boldsymbol{V}(\sum_{i=1}^n (x_i-\bar{x})) = \sigma^2(n-1)s_x^2$$

gauname (3.2.5) formules.

Nepaslinktojo dispersijos σ^2 įvertinio pavidalas ir jo nepriklausomumas nuo įvertinio $\hat{\boldsymbol{\beta}}$ gaunamas iš 1.2.1 teoremos. Paskutinis teoremos teiginys akivaizdus.▲

3.2.1 pastaba. Regresijos tiesės $y = \mu(x) = \alpha + \beta(x-\bar{x})$ krypties koeficiento β įvertinys $\hat{\beta}$ yra tuo tikslesnis, kuo didesnę reikšmę įgyja suma s_x^2. Jeigu visi matavimai atlikti viename taške $\bar{x}$, tai įvertinti kampinio koeficiento β negalima (per vieną tašką galima išvesti be galo daug skirtingų tiesių). Taigi, norint patikimiau įvertinti regresijos tiesę, reikia taip planuoti eksperimentą, kad X įgytų reikšmių kuo platesnėje srityje.

3.2.2 pastaba. Kiekvienos tiesinės funkcijos $c_1\alpha + c_2\beta$ nepaslinktasis įvertinys yra $c_1\hat{\alpha} + c_2\hat{\beta}$. Pavyzdžiui, vidurkio $\mu(x) = \alpha + \beta(x-\bar{x})$ nepaslinktasis įvertinys yra

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}(x-\bar{x}), \quad \boldsymbol{V}\hat{\mu}(x) = \sigma^2 b^2(x), \quad b^2(x) = \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right). \tag{3.2.7}$$

3.2.3 pastaba. Jei e_i yra normalieji $N(0,\sigma^2)$ vienodai pasiskirstę nepriklausomi a. d., tai pagal 1.2.1 teoremą gautieji įvertiniai pasižymi tokiomis savybėmis

$$\frac{s^2(n-2)}{\sigma^2} \sim \chi^2(n-2), \quad \frac{\hat{\alpha}-\alpha}{s_\alpha} \sim S(n-2), \quad \frac{\hat{\beta}-\beta}{s_\beta} \sim S(n-2), \tag{3.2.8}$$

$$\frac{\hat{\mu}(x)-\mu(x)}{s\,b(x)} \sim S(n-2). \tag{3.2.9}$$

Taigi parametrų α, β, $\sigma^2, \mu(x)$ pasikliovimo lygmens $Q = 1 - 2\alpha$ pasikliovimo intervalai yra

$$\hat{\alpha} \pm s_\alpha \, t_\alpha(n-2), \quad \hat{\beta} \pm s_\beta \, t_\alpha(n-2), \quad \left(\frac{SS_E}{\chi^2_\alpha(n-2)}, \frac{SS_E}{\chi^2_{1-\alpha}(n-2)} \right), \quad (3.2.10)$$

$$\hat{\mu}(x) \pm s_{\mu(x)} t_\alpha(n-2), \quad s^2_{\mu(x)} = s^2 b^2(x), \quad (3.2.11)$$

čia $t_\alpha(n-2)$ ir $\chi^2_\alpha(n-2)$ žymi Stjudento ir chi kvadrato skirstinių su $n-2$ laisvės laipsniais α kritines reikšmes. Regresijos tiesės $\mu(x)$ reikšmės taške x pasikliovimo intervalas yra trumpiausias, kai $x = \bar{x}$; šis intervalas tuo ilgesnis, kuo didesnis atstumas $|x - \bar{x}|$.

Remiantis (1.3.29) ir (1.3.32) galima sudaryti pasikliovimo intervalų rinkinį regresijos tiesei $\mu(x)$ taškuose $x_1, x_2, ..., x_k$, kad jie uždengtų visas $\mu(x_i)$ reikšmes su tikimybe ne mažesne už $Q = 1 - 2\alpha$. Tuo tikslu formulėje (3.2.11) reikia įrašyti paeiliui argumento reikšmes $x_1, x_2, ..., x_k$ ir kritinę reikšmę $t_\alpha(n-2)$ reikia pakeisti į $t_{\alpha/k}(n-2)$, jei naudojama Bonferonio nelygybė, ir pakeisti į $(2F_\alpha(2, n-2))^{1/2}$, jei naudojamas S metodas. Pastarųjų intervalų rinkinį, kai $x \in \mathbf{R}$, galima traktuoti kaip regresijos tiesės pasikliovimo zoną, kai pasikliovimo lygmuo $Q = 1 - 2\alpha$.

3.2.3. Parametro β lygybės nuliui hipotezės tikrinimas

Nagrinėkime hipotezę $H : \beta = 0$, kai alternatyva yra $\bar{H} : \beta \neq 0$. Jei ši hipotezė teisinga, kintamojo Y skirstinys nepriklauso nuo kovariantės x reikšmių, taigi prognozė neturi prasmės. Iš (3.2.8) išplaukia, kad esant teisingai hipotezei

$$T = \hat{\beta}/s_\beta \sim S(n-2),$$

taigi hipotezė H atmetama reikšmingumo lygmens α kriterijumi, kai

$$| \, T \, | > t_{\alpha/2}(n-2), \quad (3.2.12)$$

arba P-reikšmių terminais, kai

$$pv = 2 \min(\mathbf{P}\{T < t\}, \mathbf{P}\{T > t\}) < \alpha,$$

čia t yra statistikos T realizacija.

3.2.4. Tolesnio matavimo prognozė

Sakykime, kad pagal aprašytus stebėjimo rezultatus Y_i gautas regresijos tiesės įvertinys $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}(x - \bar{x})$. Reikia nurodyti kintamojo Y tolesnio nepriklausomo matavimo Y_{n+1} prognozės intervalą, jeigu žinoma, kad bus matuojama, kai kovariantės reikšmė x.

3.2.1 apibrėžimas. Toks atsitiktinis intervalas (U_1, U_2), kai

$$\mathbf{P}\{U_1 < Y_{n+1} < U_2\} = Q,$$

vadinamas atsitiktinio dydžio Y_{n+1} lygmens Q *prognozės intervalu*.

Atsitiktiniai dydžiai Y_{n+1} ir $\hat{\mu}(x)$ yra nepriklausomi ir pagal (3.2.5)

$$Y_{n+1} \sim N(\mu(x), \sigma^2), \quad \hat{\mu}(x) \sim N(\mu(x), \sigma^2 b^2(x)),$$

čia

$$b^2(x) = \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}.$$

Taigi

$$Y_{n+1} - \hat{\mu}(x) \sim N(0, \sigma^2(1+b^2(x)))$$

ir pagal 3.2.1 teoremą šis a. d. nepriklauso nuo dispersijos σ^2 įvertinio s^2. Todėl

$$\frac{Y_{n+1} - \hat{\mu}(x)}{s\sqrt{1+b^2(x)}} \sim S(n-2). \tag{3.2.13}$$

Taigi a. d. Y_{n+1} lygmens $Q = 1-2\alpha$ prognozės intervalas yra

$$\hat{\mu}(x) \pm s\sqrt{1+b^2(x)}\, t_\alpha(n-2). \tag{3.2.14}$$

Jis platesnis už vidurkio $\mu(x)$ pasikliovimo intervalą $\hat{\mu}(x) \pm sb(x)\, t_\alpha(n-2)$.

3.2.4 pastaba. Prognozės intervalas yra simetriškas atžvilgiu $\hat{\mu}(x)$ ir tuo platesnis, kuo x daugiau skiriasi nuo kovariančių reikšmių, kuriomis buvo vertinama regresijos tiesė, vidurkio $\bar{x}$. Todėl prognozavimas, kai skirtumo $|x-\bar{x}|$ reikšmės didelės, mažai prasmingas. Reikia atsižvelgti ir į tai, kad regresijos tiesinis pavidalas dažnai yra aproksimacija, naudojama kovariantės reikšmių $x_1, \ldots, x_n$ siauroje srityje, taigi ši aproksimacija gali būti bloga, kai x nepatenka į minėtą sritį.

3.2.5 pastaba. Jeigu reikia sudaryti pasikliovimo intervalų rinkinį, kuris uždengtų tolesnius nepriklausomus matavimus $Y_{n+1}, ..., Y_{n+k}$, gautus taškuose $x^{(1)}, ..., x^{(k)}$, tai pakanka formulėje (3.2.14) įrašyti paeiliui $x^{(1)}, ..., x^{(k)}$, o kritinę reikšmę $t_\alpha(n-2)$ pakeisti į $t_{\alpha/k}(n-2)$(naudojama Bonferonio nelygybė), arba pakeisti $(kF_\alpha(k, n-2))^{1/2}$ (naudojamas S metodas).

3.2.6 pastaba. Kartais reikia spręsti atvirkščią prognozavimo uždavinį, kai kintamojo Y reikšmė Y_{n+1} yra žinoma, o reikia rasti kintamojo reikšmės x, kurią atitinka Y_{n+1}, pasikliovimo intervalą. Tokį intervalą galima gauti taip pat pasinaudojus (3.2.12) sąryšiu, tačiau šiuo atveju x yra ne tik skaitiklyje, bet ir vardiklyje, taigi turime spręsti antrojo laipsnio nelygybę

$$(Y_{n+1} - \hat{\alpha} - \hat{\beta}(x-\bar{x}))^2 \le s^2(1+b^2(x))t_\alpha^2(n-2).$$

Gauname intervalo rėžius

$$\bar{x} + (B \mp \sqrt{B^2 - AC})/A; \tag{3.2.15}$$

čia

$$A = \hat{\beta}^2 - \frac{s^2 t_\alpha^2(n-2)}{(n-1)s_x^2}, \quad B = \hat{\beta}(Y_{n+1} - \hat{\alpha}),$$

$$C = (Y_{n+1} - \hat{\alpha})^2 - \frac{n+1}{n} s^2 t_\alpha^2(n-2).$$

3.2.5. Regresijos tiesiškumo hipotezės tikrinimas

Norint patikrinti vidurkio $\mathbf{E}(Y \mid x)$ tiesiškumo (regresijos tiesiškumo) hipotezę, reikalingi pasikartojantys matavimai kai fiksuotos kovariantės reikšmės. Tarkime, kad turima n_i a. d. Y stebėjimų, kai kovariantės x reikšmė yra x_i $(i = 1, \ldots, k)$. Taigi turimi duomenys $Y_{ij},\ j = 1, ..., n_i\ i = 1, ..., k$. Pažymėkime $n = \sum_{i=1}^{k} n_i$.

Tegu Y regresija X atžvilgiu yra bet kokio pavidalo. Pažymėkime $\mu_i = \mathbf{E}(Y \mid x_i)$ ir $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)^T$. Nagrinėjame modelį

$$Y_{ij} = \mu_i + e_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., k.$$

Jei teisinga tikrinamoji hipotezė, tai vidurkiai $\mu_1, ..., \mu_k$ yra dviejų parametrų α ir β tiesinės funkcijos, t. y. teisinga hipotezė

$$H : \mu_i = \alpha + \beta(x_i - \bar{x}), \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \quad i = 1, ..., k.$$

Taigi formaliai tikrinsime, ar duomenys neprieštarauja hipotezei H. Pavyzdyje 1.2.3 (ir jo tęsinyje) kritinė sritis turėjo pavidalą (imame $m = 1$):

$$F = \frac{(SS_{EH} - SS_E)(n-k)}{SS_E(k-2)} > F_\alpha(k-2,\ n-k); \tag{3.2.16}$$

čia $F_\alpha(k-2,\ n-k)$ – Fišerio skirstinio α kritinė reikšmė, o

$$SS_E = \min_{\boldsymbol{\mu}} \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \mu_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2, \quad \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij},$$

$$SS_{EH} = \min_{\boldsymbol{\mu}:\mu_i = \alpha + \beta(x_i - \bar{x})} SS(\boldsymbol{\mu}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2,$$

$$SS_{EH} - SS_E = \sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2,$$

čia $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)^T$.

3.2.7 pastaba. Analogiškai galime rasti kriterijų hipotezei, kad vidurkiai $\mu_1, ..., \mu_k$ priklauso, pavyzdžiui, parabolei, ar kitai kreivei, tiesiškai priklausančiai nuo mažesnio už k parametrų skaičiaus, tikrinti. Pavyzdžiui, hipotezė $\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,\ i = 1, ..., k$, reiškia, kad visi vidurkiai $\mu_1, ..., \mu_k$ yra trijų parametrų $\beta_0, \beta_1, \beta_2$ funkcijos. Tuo atveju statistikos skaitiklyje SS_{EH} išraiškoje vietoje $\hat{\alpha} + \hat{\beta}(x_i - \bar{x})$ įrašome $\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$; čia $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ yra parametrų β_i mažiausiųjų kvadratų įvertiniai modelyje $Y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$. Tokio tipo modeliai nagrinėjami 3.3 skyrelyje. Kriterijus turi (3.2.16) pavidalą, tiktai $k-2$ pakis į $k-3$ statistikos vardiklyje ir (3.2.16) nelygybės dešinėje pusėje. Tai vėl gaunama iš 1.2.3 pavyzdžio rezultatų.

3.2.8 pastaba. Jeigu kiekviename taške x_i turime tik po vieną stebėjimą Y_i (visi $n_i = 1$), tai kriterijaus (3.2.16) pritaikyti negalime, nes $SS_E = 0$. Šiuo atveju tam tikros informacijos apie regresijos pavidalą suteikia likutiniai skirtumai

$$\hat{e}_i = Y_i - \hat{Y}_i, \quad \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})), \quad i = 1, ..., n.$$

Pažymėję

$$\hat{\boldsymbol{Y}} = (\hat{Y}_1, ..., \hat{Y}_n)^T, \quad \hat{\boldsymbol{e}} = (\hat{e}_1, ..., \hat{e}_n)^T,$$

gauname

$$\hat{\boldsymbol{e}} = \boldsymbol{DY}, \quad \boldsymbol{D} = \boldsymbol{I}_n - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T = [d_{ij}]_{n\times n}, \quad \mathbf{E}(\hat{\boldsymbol{e}}) = \boldsymbol{0}, \quad \boldsymbol{V}(\hat{\boldsymbol{e}}) = \sigma^2\boldsymbol{D}.$$

Taigi

$$\hat{e}_i \sim N(0, \sigma^2 d_{ii}).$$

Kadangi

$$\hat{\sigma}^2 = MS_E = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} \xrightarrow{P} \sigma^2, \quad n \to \infty,$$

tai

$$\tilde{e}_i = \frac{\hat{e}_i}{\sqrt{MS_E d_{ii}}} \xrightarrow{d} Z \sim N(0,\, 1).$$

Atsitiktiniai dydžiai $\tilde{e}_i$ vadinami *standartizuotomis liekanomis.*

Kai regresijos tiesiškumo prielaida teisinga, $\tilde{e}_i$ skirstiniai mažai skiriasi nuo standartinio normaliojo. Aišku, net kai n yra didelis, vektoriaus $(\tilde{e}_1, ..., \tilde{e}_n)^T$ negalima traktuoti kaip paprastosios imties, nes a. d. $\tilde{e}_i$ yra priklausomi. Daugelyje statistinių programų paketų naudojami neformalūs diagnostiniai grafiniai metodai modelio adekvatumui tikrinti, remiantis standartizuotomis liekanomis. Šiame vadovėlyje neformalių metodų neaptariame, informacijos apie juos galima rasti taikomojo pobūdžio knygose [4] I dalis; [5].

3.2.6. Atsitiktinių kovariančių atvejis

Atsižvelgiant į eksperimento planą, priklausomo kintamojo Y ir nepriklausomo kintamojo X sąryšį galima aprašyti trimis būdais:

1. Kovariantė x nėra a. d., o tam tikras determinuotas kintamasis (pvz., kalendorinis laikas), nuo kurio gali priklausyti a. d. Y skirstinys.

2. Vektorius $(Y, X)^T$ yra atsitiktinis, tačiau eksperimentas planuojamas taip, kad a. d. Y nepriklausomi stebėjimai gaunami kai yra iš anksto parinktos fiksuotos a. d. X reikšmės $x_1, ..., x_m$. Pavyzdžiui, tarkime, Y reiškia žmogaus ūgį, o X – svorį. Eksperimentas atliekamas taip, kad galima atsitiktinai atrinkti n_i individų, kurių svoris vienodas ir lygus x_i, ir pamatuoti jų ūgį Y_{ij}, $j = 1, ..., n_i$; $i = 1, ..., k$.

3. Stebint a. v. $(Y, X)^T$ gauta imtis $(Y_i,\ X_i)^T$, $i = 1, ..., n$.

Regresinio modelio nežinomų parametrų įvertinių savybes ir statistinius kriterijus nagrinėjome, tarę, kad kovariantės X reikšmės yra fiksuotos. Kadangi gautųjų įvertinių skirstiniai (žr. (3.2.8), (3.2.9), (3.2.12)) nepriklauso nuo fiksuotųjų X reikšmių $x_1, ..., x_m$, tai gautos išvados yra teisingos visais trimis atvejais.

3.2.7. Regresijos ir koreliacijos koeficientų sąryšis

Tarkime, kad turime imtį

$$(Y_1, X_1)^T, \ldots, (Y_n, X_n)^T,$$

gautą stebint a. v. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Kai yra normalusis skirstinys, sąlyginiai vidurkiai $\mathbf{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$ yra tiesinės X_i funkcijos, $\boldsymbol{V}(Y_i|X_i) = \sigma^2$ nepriklauso nuo X_i, todėl kai kovariantės X reikšmės fiksuotos, yra teisingas tiesinės regresijos modelis (3.2.2).

Pažymėkime vidurkius $\mu_X = \mathbf{E}X$, $\mu_Y = \mathbf{E}Y$ ir antruosius momentus $\sigma_X^2 = \boldsymbol{V}X$, $\sigma_Y^2 = \boldsymbol{V}Y$, $\rho\sigma_X\sigma_Y = \mathbf{Cov}\,(X, Y)$; čia ρ yra a. d. Y ir X koreliacijos koeficientas.

3.2.2 teorema. *Koreliacijos koeficientas tenkina lygybes:*

$$\beta = \rho\frac{\sigma_Y}{\sigma_X}, \quad \sigma^2 = \sigma_Y^2(1 - \rho^2). \tag{3.2.17}$$

Įrodymas. Naudosimės dvimačio normaliojo skirstinio savybėmis (žr 2 priedą (7.4.7)):

$$\mu(x) = \mathbf{E}(Y \mid X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \quad \boldsymbol{V}(Y \mid X = x) = \sigma_Y^2(1 - \rho^2).$$

Gauname

$$\mathbf{E}(Y \mid X = x_i) = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(\bar{x} - \mu_X) + \rho\frac{\sigma_Y}{\sigma_X}(x_i - \bar{x}),$$

$$\boldsymbol{V}(Y \mid X = x_i) = \sigma_Y^2(1 - \rho^2).$$

Bet pagal tiesinės regresijos modelį (3.2.2)

$$\mathbf{E}(Y \mid X = x_i) = \alpha + \beta(x_i - \bar{x}), \quad \boldsymbol{V}(Y \mid X = x_i) = \sigma^2.$$

Taigi gauname (3.2.17) lygybes.▲

Jeigu Y ir X skirstiniai neišsigimę, hipotezė $H : \rho = 0$ yra ekvivalenti hipotezei $\beta = 0$ (arba hipotezei, kad pirmiau įvestas tiesinės prognozės tikslumo matas – dauginis koreliacijos koeficientas $\rho_{Y.X}^2 = \rho^2 = 0$).

Parametrų $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \mathbf{Cov}\,(X, Y), \rho$ įvertiniai:

$$\hat{\mu}_X = \bar{Y}, \quad \hat{\mu}_Y = \bar{X}, \quad s_X^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2, \quad s_Y^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2,$$

$$s_{XY} = \frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y}), \quad r = \frac{s_{XY}}{s_X s_Y}. \tag{3.2.18}$$

Naudodamiesi empirinio koreliacijos koeficiento r išraiška, 3.2.1 teoremoje pateiktus parametrų α ir β įvertinius galime užrašyti kitu pavidalu:

$$\hat{\alpha} = \bar{Y}, \quad \hat{\beta} = r\frac{s_Y}{s_X}.$$

3.2.9 pastaba. Kai skirstinys dvimatis normalusis, hipotezės $H : \rho = 0$ tikrinimo kriterijus remiansi sąryšiu (žr. 1 dalies 5.7.3 skyrelį)

$$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim S(n-2).$$

Šiame skyrelyje hipotezei $\beta = 0$ tikrinti naudojomės tuo, kad

$$\hat{\beta}/s_\beta \sim S(n-2).$$

Jei paskutinėje statistikoje x_i pakeisime į X_i (o tai galime padaryti, nes Stjudento skirstinys nepriklauso nuo kovariančių X_i fiksuotųjų reikšmių x_i), tai pirmiau pateiktos statistikos sutampa. Iš tikrųjų, kadangi pagal (3.2.4)

$$SS_E = \sum_i (Y_i - \bar{Y} - r\frac{s_Y}{s_X}(X_i - \bar{X}))^2 = (n-1)s_Y^2(1-r^2),$$

$$s^2 = \frac{SS_E}{n-2} = \frac{n-1}{n-2}s_Y^2(1-r^2), \quad s_\beta^2 = \frac{s^2}{(n-1)s_X^2},$$

tai

$$\frac{\hat{\beta}}{s_\beta} = \frac{r\, s_Y/s_X}{(s^2/((n-1)s_X^2))^{1/2}} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}.$$

Taigi turime identiškus kriterijus.

3.2.10 pastaba. Jeigu nesame tikri, kad $(Y, X)^T$ skirstinys yra dvimatis normalusis, tai, norint parinkti geriau prognozuojantį netiesinį modelį, reikėtų įvertinti koreliacinį santykį η_{YX}^2. Nežinant regresijos kreivės pavidalo, tiesiogiai įvertinti koreliacinį santykį pagal imtį $(Y_i,\ X_i),\ i = 1, ..., n$ negalima. Tam tikslui reikėtų kovariantės X reikšmes sugrupuoti (t. y. suapvalinti tam tikru tikslumu).

Tarkime, kad turimos suapvalintų stebėjimų poros $(X_i,\ Y_{ij})$, $j = 1, ..., n_i$, $i = 1, ..., k$, $n = \sum_i n_i$. Koreliacinio santykio įvertinys yra

$$\hat{\eta} = \frac{s_\mu^2}{s_Y^2}, \quad s_\mu^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2, \quad \bar{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}. \tag{3.2.19}$$

3.2.1 pavyzdys. 3.2.1 lentelėje pateikti duomenys, gauti matuojant $n = 50$ mokinių svorį Y (kg) ir ūgį X (cm). Tiksliau kalbant, lentelėje pateikti sugrupuoti duomenys; ūgio grupavimo intervalo ilgis 5 cm; svorio – 3 kg; X_i ir Y_j – grupavimo intervalų centrai. Pagal šiuos duomenis įvertinsime tiesinės regresijos lygties koeficientą, koreliacijos koeficientą ir koreliacinį santykį prognozuodami a. d. Y pagal X.

3.2.1 lentelė. Statistiniai duomenys

X_iY_i	24	27	30	33	36	Σ
120	1	3	-	-	-	4
125	-	2	6	1	-	9
130	-	1	5	5	-	11
135	-	1	6	7	2	16
140	-	-	1	4	2	7
145	-	-	-	1	1	2
150	-	-	-	-	1	1
Σ	1	7	18	18	6	50

Randame pirmųjų ir antrųjų momentų įverčius

$$\hat{\mu}_X = 132,3;\ \hat{\mu}_Y = 31,26;\ s_X^2 = 47,153;\ s_Y^2 = 8,115;\ s_{XY} = 13,573.$$

Pasinaudoję šiais įverčiais, randame koreliacijos koeficiento ρ, dauginio koreliacijos koeficiento ρ_{YX}^2 ir koreliacinio santykio η_{YX}^2 įverčius:

$$\hat{\rho} = r = 0,694;\ \hat{\rho}_{YX}^2 = r^2 = 0,482;\ \hat{\eta}_{YX}^2 = 0,518.$$

Iš įverčių $\hat{\rho}_{YX}^2$ ir $\hat{\eta}_{YX}^2$ artumo galima daryti išvadą, kad tiesinės regresijos modelis šiame uždavinyje yra priimtinas. Analogišką išvadą gauname taikydami kriterijų (3.2.6). Regresijos tiesės įvertis

$$\hat{Y}(x) = \hat{\alpha} + \hat{\beta}(x - \bar{X}) = \bar{Y} + r\frac{s_Y}{s_X}(x - \bar{X}) =$$
$$= 31,26 + 0,288(x - 132,3).$$

Koreliacijos koeficiento pasikliovimo intervalas, surastas pagal 1 dalies 4.7.3 skyrelį, yra $(0,515,\ 0,815)$; pasikliovimo lygmuo $Q = 0,95$.

3.2.8. Netiesinė regresija

Sprendžiant praktinius uždavinis retai pasitaiko, kad regresijos kreivė būtų tiksliai tiesinė. Nepaisant to, šis modelis gana dažnai taikomas, jei argumento kitimo sritis nėra didelė ir regresijos kreivė nedaug skiriasi nuo tiesės šioje srityje.

Jeigu regresijos kreivė yra akivaizdžiai netiesinė, tai ją kartais galima pertvarkyti į kelių kintamųjų tiesinę regresiją, įtraukint papildomas kovariantes. Pavyzdžiui, iš netiesinės regresijos modelio, kuriame $\mu(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 e^x$, gaunamas tiesinės kelių kintamųjų regresijos modelis $\mu(x_1,\ x_2, x_3) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$; čia $x = x_1,\ x_2 = x^2,\ x_3 = e^x$.

Kartais netiesinį modelį galima pertvarkyti į tiesinį, jei egzistuoja tolydžios ir monotoninės funkcijos $U = g(Y)$ ir $V = h(X)$, tokios, kad a. d. U regresija V atžvilgiu yra tiesinė:

$$\mathbf{E}(U|V = v) = \alpha + \beta v.$$

Pažymėkime $U = g(Y)$ ir $V = h(X)$. Jeigu modelyje $U_i = \alpha + \beta v_i + e_i$ a. d. e_i tenkina kitas šio skyrelio pradžioje suformuluotas tiesinės regresijos modelio sąlygas, tai, atlikus analizę šiame modelyje, išvadas apie kovariantės x įtaką a. d. Y skirstiniui galima daryti grįžus prie pradinių kintamųjų Y ir X atvirkštinėmis funkcijomis. Pavyzdžiui, suradus MK įvertinius $\hat{\alpha}$ ir $\hat{\beta}$ tiesiniame modelyje $U_i = \alpha + \beta v_i + e_i$, regresijos $\mu(x) = \mathbf{E}(Y|X = x)$ įvertiniu galima naudoti $\hat{\mu}(x) = g^{-1}(\hat{\alpha} + \hat{\beta}h(x))$.

Pateiksime keletą dažniausiai naudojamų funkcijų, transformuojančių netiesinį modelį į tiesinį:

1. Jei $y = \alpha x^\beta$, tai $u = \ln y,\ v = \ln x$ ir $u = \ln\alpha + \beta v$.
2. Jei $y = \alpha e^{\beta x}$, tai $u = \ln y,\ x = v$ ir $u = \ln\alpha + \beta v$.
3. Jei $y = \frac{x}{\alpha x + \beta}$, tai $u = \frac{1}{y},\ v = \frac{1}{x}$ ir $u = \alpha - \beta v$.
4. Jei $y = \alpha + \beta\ln x$, tai $u = y,\ v = \ln x$ ir $u = \alpha + \beta v$.
5. Jei $y = e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})$, tai $u = \ln\frac{y}{1+y},\ v = x$ ir $u = \alpha + \beta v$.

Pagaliau, jeigu regresija $\mu(x) = \mathbf{E}(Y|X = x)$ nurodytais būdais nepertvarkoma į tiesinę, bet užrašoma žinoma funkcija $\mu(x, \boldsymbol{\beta})$ nuo kovariantės x ir nuo

baigtiniamačio parametro $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$, tai parametrą $\boldsymbol{\beta}$ galima vertinti mažiausiųjų kvadratų metodu minimizuojant kvadratinę formą

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \mu(x_i, \boldsymbol{\beta}))^2.$$

Šiuo atveju, minimizuojant kvadratinę formą $SS(\boldsymbol{\beta})$, nepakanka išspręsti tiesinių lygčių sistemą, tenka ekstremumo ieškoti artutiniais skaitiniais metodais. Bendru atveju apie gautų įvertinių optimalumo savybes mažai ką galime pasakyti.

3.2.2 pavyzdys. 3.2.2 lentelės duomenys apibūdina tam tikro cheminės reakcijos produkto išeigą Y (padidėjimas gramais kiekvienam gramui pradinės medžiagos), atsižvelgiant į reakcijos laiką X (valandos).

3.2.2 lentelė. Statistiniai duomenys

i	X_i	Y_{ij}	n_i	$\bar{Y}_{i\cdot}$
1	1	1,11 1,12 1,07	3	1,100
2	2	1,34 1,38 1,39	3	1,370
3	3	1,47 1,44 1,38 1,38 1,41	5	1,416
4	7	1,51 1,48 1,53 1,52 1,55	5	1,618
5	28	1,62 1,62 1,60 1,55 1,57	5	1,592

Šiame pavyzdyje tiesinės regresijos modelis stebėjimams aprašyti netinka. Regresijos tiesiškumo hipotezė atmetama aukštu reikšmingumo lygmeniu, nes statistika F iš (3.2.6) įgyja reikšmę 89,47 (kai hipotezė teisinga, statistika turi Fišerio skirstinį su 3 ir 16 laisvės laipsnių).

Parinkime kitą regresijos kreivės pavidalą. Regresijos kreivė turėtų tenkinti tokias sąlygas: jeigu $X = 0$, tai funkcijos reikšmė lygi 0; mažoms argumento reikšmėms, kol daug nesureagavusios medžiagos, funkcijos pokyčiai turėtų būti didesni negu didelėms; argumentui didėjant funkcija turėtų artėti prie tam tikros maksimaliai galimos reikšmės (asimptotos), kai visa medžiaga sureaguoja. Matavimo rezultatų kitimas patvirtina šias prielaidas. Suformuluotas sąlygas tenkina, pavyzdžiui, funkcija

$$f(x) = \gamma e^{-\beta \frac{1}{x}}, \ \ \gamma > 0, \ \beta > 0.$$

Tada, nagrinėjant naujus kintamuosius $U = \ln Y$ ir $V = 1/X$, a. d. U regresija V atžvilgiu turėtų būti artima tiesinei

$$\mu(v) = \mathbf{E}(U|V = v) = \alpha - \beta(v - \bar{V});$$

$$\alpha = \ln\gamma - \beta\bar{V}, \ \bar{V} = \frac{1}{n}\sum_{i=1}^{5} n_i V_i = \frac{1}{n}\sum_{i=1}^{5} \frac{n_i}{X_i}.$$

Stebėjimo duomenys perėjus prie naujų kintamųjų U ir V, pateikti 3.2.3 lentelėje.

3.2.3 lentelė. Transformuoti duomenys

i	V_i	U_{ij}	n_i	$\bar{U}_{i\cdot}$
1	1,000	0,1044 0,1133 0,0677	3	0,0951
2	0,500	0,2927 0,3221 0,3293	3	1,370
3	0,333	0,3853 0,3646 0,3221 0,3221 0,3436	5	0,3475
4	0,143	0,4121 0,3920 0,4253 0,4187 0,4383	5	0,4173
5	0,036	0,4824 0,4824 0,4700 0,4383 0,4511	5	0,4648

Pagal šios lentelės duomenis regresijos tiesiškumo hipotezę atmesti nėra pagrindo (statistika (3.2.16) įgijo reikšmę 1,258).

Gauname regresijos tiesės įvertį

$$\hat{\mu}(v) = 0,3513 - 0,3737(v - 0,3362).$$

Grįžę prie kintamųjų X ir Y gauname

$$\hat{f}(x) = \hat{\gamma}e^{-\hat{\beta}\frac{1}{x}} = 1,6113e^{-0,3737\frac{1}{x}}.$$

3.3. Tiesinė keleto kintamųjų regresija

3.3.1. Statistinis modelis

Tarkime, kad norime prognozuoti a. d. Y remdamiesi m kovariantėmis $X_1, \ldots, X_m$. Žymėkime $\mathbf{X} = (X_0,\ X_1, ..., X_m)^T$ kovariančių vektorių papildytą koordinate $X_0 \equiv 1$. Fiksavus kovariantės $\mathbf{X}$ reikšmes $\mathbf{X}^{(i)} = \boldsymbol{x}^{(i)} = (x_{0i}, x_{1i}, ..., x_{mi})^T$, gauti nepriklausomi a. d. Y stebėjimai $Y_i,\ \ i = 1, ..., n$.

Kaip ir vienos kovariantės atveju darome prielaidą, kad dydžiai $\boldsymbol{x}^{(i)}$ yra neatsitiktiniai arba yra nepriklausomų vienodai pasiskirsčiusių a. v. $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ realizacijos. Pastaruoju atveju analizė yra sąlyginė, naudojamasi tik šiomis realizacijomis, bet ne a. v. $\mathbf{X}^{(i)}$ skirstiniais.

Tiesinis kelių kintamųjų regresijos modelis:

$$Y_i = \mu(\boldsymbol{x}^{(i)}) + e_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + e_i, \quad i = 1, ..., n; \tag{3.3.1}$$

čia e_i vienodai pasiskirstę nekoreliuoti a. d. su nuliniais vidurkiais ir vienodomis dispersijomis $\boldsymbol{V}e_i = \sigma^2$. Tariama, kad sąlyginis vidurkis $\mu(\boldsymbol{x}^{(i)}) = \mathbf{E}(Y_i|\boldsymbol{x}^{(i)}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}$ yra tiesinė $m+1$ kovariantės funkcija su nežinomais koeficientais. Kita vertus, šis vidurkis yra tiesinė nežinomų parametrų funkcija su žinomais koeficientais.

Pažymėkime

$$\boldsymbol{Y} = (Y_1, ..., Y_n)^T, \quad \boldsymbol{\beta} = (\beta_0,\ \beta_1, ..., \beta_m)^T, \quad \boldsymbol{e} = (e_i, ..., e_n)^T,$$

$$\boldsymbol{A} = \begin{pmatrix} 1\ x_{1i}\ ...\ x_{1m} \\ 1\ x_{21}\ ...\ x_{2m} \\ ...\ ...\ ...\ ... \\ 1\ x_{n1}\ ...\ x_{nm} \end{pmatrix}.$$

Tada stebėjimus (3.3.1) galima užrašyti (1.1.2) matriciniu pavidalu

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.3.2}$$

$$\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\beta}, \quad \boldsymbol{V}(\boldsymbol{Y}) = \boldsymbol{V}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}. \tag{3.3.3}$$

Taigi turime atskirą tiesinio modelio (1.1.2) atvejį su konkrečiais koeficientais prie nežinomų parametrų.

3.3.2. Koeficientų β interpretacija

Tarkime, kad j-oji kovariantė yra tolydi. Imkime du kovariančių vektorius $\boldsymbol{x}^{(1)}$ ir $\boldsymbol{x}^{(2)}$, kurių visos koordinatės, išskyrus j-ąją, yra vienodos, o $x_j^{(2)} = x_j^{(1)} + 1$. Iš (3.3.1) formulės išplaukia, kad

$$\mu(\boldsymbol{x}^{(2)}) - \mu(\boldsymbol{x}^{(1)}) = \mu(x_1, ..., x_j^{(2)}, ..., x_m) - \mu(x_1, ..., x_j^{(1)}, ..., x_m) = \beta_j. \quad (3.3.4)$$

Taigi *parametras β_j lygus priklausomo kintamojo Y vidurkio pokyčiui, kai j-oji kovariantė padidėja vienetu, kitoms kovariantėms nepakitus.*

Jei j-oji kovariantė nominali, norint, kad modelio parametrai turėtų prasmę, kovariantę reikia koduoti.

Tarkime, kad j-oji kovariantė x_j nominali, pavyzdžiui, ligos stadija, lytis, rasė, ir pan., ir įgyja k skirtingų reikšmių (sakykime, reikšmes 1, 2,...,k). Tada vietoje $\beta_j x_j$ tiesinės regresijos modelyje imamas narys $\beta_{j1} z_{j1} + \beta_{j2} z_{j2} + \cdots + \beta_{j,k-1} z_{j,k-1}$, kur

$$z_{jl} = \begin{cases} 1, & \text{jei } x_j = l+1 \quad (l = 1, \ldots, k-1); \\ 0, & \text{jei } x_j = 1. \end{cases}$$

Taigi modelis (3.3.1) modifikuojamas tokiu būdu:

$$\mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \sum_{i=1}^{k-1} \beta_{ji} z_{ji} + ... + \beta_m x_m.$$

Tuo atveju j-ąją kovariantę atitinkantis narys įgyja tokias reikšmes:

$$\sum_{i=1}^{k-1} \beta_{ji} z_{ji} = \begin{cases} \beta_{j,l-1}, & \text{jei } x_j = l, \quad l = 2, \ldots, k; \\ 0, & \text{jei } x_j = 1. \end{cases}$$

Jei, pavyzdžiui, x_j yra sėklos rūšis, įgyjanti 3 reikšmes, tai modelyje (3.3.1) imti narį $\beta_j x_j$ su x_j, įgyjančiu reikšmes $1, 2$ ir 3, būtų neteisinga, nes toks modelis reikštų, kad mes iš karto darome prielaidą, jog pereinantėjimas nuo 1-osios prie 2-sios rūšies gaunamas toks pat vidutinis priklausomo kintamojo, pavyzdžiui, derlingumo pokytis, kaip ir pereinant nuo 2-osios prie 3-sios rūšies. Bet regresinės analizės tikslas yra būtent nustatyti, kaip vidutinis derlingumas priklauso nuo sėklos rūšies. Šiuo atveju vietoje nario $\beta_j x_j$ imamas narys $\beta_{j1} z_{j1} + \beta_{j2} z_{j2}$, čia z_{j1} įgyja reikšmę 1 antrajai javų rūšiai, o z_{j2} įgyja reikšmę 1 trečiajai javų rūšiai.

Panagrinėkime koeficientų ir modelio interpretaciją po kodavimo. Imkime du kovariančių vektorius $\boldsymbol{x}^{(1)}$ ir $\boldsymbol{x}^{(2)}$, kuriems visos kovariantės, išskyrus j-ąją nominalią kovariantę, yra vienodos, j-josios kovariantės reikšmė pirmajam vektoriui yra vienetas, o antrajam $- l + 1$. Jeigu nominaliosios kovariantės reikšmė lygi 1, tai ją atitinka vektorius $(z_{j1}^{(1)}, ..., z_{j,k-1}^{(1)})^T = (0, ..., 0)^T$, o jeigu nominaliosios kovariantės reikšmė yra $l + 1$, tai ją atitinka vektorius $(z_{j1}^{(2)}, ..., z_{j,k-1}^{(2)})^T = (0, ..., 0, 1, 0, ..., 0)^T$; čia 1 yra l-oje pozicijoje. Gauname

$$\mu(x^{(2)}) - \mu(x^{(1)}) = \beta_{jl}, \quad l = 1, ..., k-1.$$

Taigi *parametras β_{jl} parodo kintamojo Y vidurkio pokytį, kai j-osios kovariantės reikšmė pakinta nuo pirmosios iki $l+1$-osios, kitoms kovariantėms nepakitus.* Pavyzdžiui, jei x_j yra sėklos rūšis, įgyjanti 3 reikšmes, tai β_{j1} parodo derlingumo vidurkio pokytį, pereinant nuo pirmosios prie antrosios rūšies, o β_{j2} rodo derlingumo vidurkio pokytį, pereinant nuo pirmosios prie trečiosios rūšies.

3.3.3. Modelis, kai yra kovariančių sąveika

Jei tas pats kovariantės x_j reikšmės pokytis sukelia skirtingą vidutinės priklausomo kintamojo reikšmės pokytį esant įvairioms kitoms kovariančių reikšmėms, turime x_j ir šių kovariančių sąveiką. Tada (3.3.1) modelis modifikuojamas. Pavyzdžiui, kai yra dvi tolydžiosios kovariantės, naudojamas modelis

$$\mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

o kai yra trys kovariantės:

$$\mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3.$$

Kai $m = 2$, turime vidurkio pokytį

$$\mu(x_1 + 1, x_2) - \mu(x_1, x_2) = \beta_1 + \beta_3 x_2. \tag{3.3.5}$$

Taigi *priklausomo kintamojo Y vidurkio pokytis, padidėjus pirmajai kovariantei vienetu, priklauso nuo antrosios kovariantės reikšmės x_2 ir lygus $\beta_1 + \beta_3 x_2$.* Jei, pavyzdžiui, Y yra automobilio kaina (litais), x_1 amžius (metais), x_2 galingumas (cm^3), ir yra amžiaus ir galingumo sąveika, tai akivaizdu, kad kaina mažėja kasmet, bet kainos mažėjimas skirtingas automobilių su skirtingu galingumu. x_2 (cm^3) galingumo automobiliui metinis kainos sumažėjimas yra $\beta_1 + \beta_3 x_2$ (litų).

Jei x_1 yra tolydi kovariantė, o x_2 yra nominali kovariantė, įgyjanti tris reikšmes, tai naudojamas modelis

$$\mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_{21} z_{21} + \beta_{22} z_{22} + \beta_{121} x_1 z_{121} + \beta_{122} x_1 z_{22}.$$

Tada

$$\mu(x_1 + 1, z_{21}, z_{22}) - \mu(x_1, z_{21}, z_{22}) = \beta_1 + \beta_{121} z_{21} + \beta_{122} z_{22}.$$

Taigi *priklausomo kintamojo Y vidurkio pokytis, padidėjus pirmajai kovariantei vienetu, priklauso nuo antrosios kovariantės reikšmės x_2 ir lygus β_1, $\beta_1 + \beta_{121}$ ir $\beta_1 + \beta_{122}$, kai yra atitinkamai nulinės, pirmosios ir trečiosios nominalios kovariantės reikšmės.*

3.3.4. Parametrų įvertiniai

Tarkime, kad matrica $\boldsymbol{A}^T\boldsymbol{A}$ neišsigimusi. Tada parametro $\boldsymbol{\beta}$ mažiausiųjų kvadratų įvertinys (1.2.3) yra

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}, \tag{3.3.6}$$

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}. \tag{3.3.7}$$

Analogiškai, jeigu $\theta = \boldsymbol{L}^T\boldsymbol{\beta} = L_0\beta_0 + L_1\beta_1 + \cdots + L_m\beta_m$ yra tiesinė regresinių parametrų funkcija, tai pagal (1.2.3) teoremą įvertinys

$$\hat{\theta} = \boldsymbol{L}^T\hat{\boldsymbol{\beta}}, \tag{3.3.8}$$

yra minimalios dispersijos įvertinys visų nepaslinktųjų tiesinių parametro θ įvertinių klasėje, ir

$$\mathbf{E}(\hat{\theta}) = \boldsymbol{\theta}, \quad \boldsymbol{V}(\hat{\theta}) = \sigma^2 b^2, \quad b^2 = \boldsymbol{L}^T\boldsymbol{C}\boldsymbol{L}, \quad \boldsymbol{C} = (\boldsymbol{A}^T\boldsymbol{A})^{-1} = [c_{ij}]_{(m+1)\times(m+1)}. \tag{3.3.9}$$

Imdami $\boldsymbol{L} = \boldsymbol{x}$ gauname regresijos $\mu(\boldsymbol{x}) = \mathbf{E}(Y|\boldsymbol{x})$ įvertinį:

$$\hat{\mu}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}^T\boldsymbol{x} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_m x_m.$$

Pagal 1.2.2 teoremą nepaslinktasis dispersijos σ^2 įvertinys yra

$$\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-m-1}, \quad \mathbf{E}s^2 = \sigma^2, \tag{3.3.10}$$

$$SS_E = (\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}^{(i)})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{3.3.11}$$

$$\hat{Y}_i = \hat{\mu}(\boldsymbol{x}^{(i)}) = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im}. \tag{3.3.12}$$

Atsitiktiniai dydžiai $\hat{Y}_i$ ir Y_i vadinami atitinkamai *prognozuojamomis* ir *stebėtomis* priklausomo kintamojo Y reikšmėmis, o a. d. $\hat{e}_i = Y_i - \hat{Y}_i$ – *liekamosiomis arba prognozės paklaidomis.*

Prognozuojamų ir stebėtų reikšmių sklaidą apibūdina kvadratų sumos:
$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ – *liekamųjų paklaidų kvadratų suma*,
$SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ – *pilnoji kvadratų suma*,
$SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ – *regresijos kvadratų suma.*

Pilnoji kvadratų suma SS_T apibūdina stebėjimų Y_i sklaidą apie jų aritmetinį vidurkį $\bar{Y}$; regresijos kvadratų suma SS_R – regresijos modeliu prognozuojamų reikšmių $\hat{Y}_i$ sklaidą apie $\bar{Y}$; liekamųjų kvadratų suma SS_E – atstumą tarp stebėtų ir prognozuojamų reikšmių. Kaip įsitikinsime, $SS_E = SS_T - SS_R$, taigi ši kvadratų suma dar parodo, kuri Y_i sklaidos dalis lieka nepaaiškinta regresiniu modeliu.

3.3.1 teorema. *Teisingos lygybės*

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i, \quad SS_T = SS_E + SS_R. \tag{3.3.13}$$

Įrodymas. Pažymėkime $\hat{\boldsymbol{e}} = (\hat{e}_1, \ldots, \hat{e}_n)^T$, $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$. Kadangi

$$\boldsymbol{A}^T\hat{\boldsymbol{e}} = \boldsymbol{A}^T(\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{A}^T(\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} - \boldsymbol{Y}) = 0, \tag{3.3.14}$$

o pirmoji matricos $\boldsymbol{A}^T$ eilutė yra $\mathbf{1}_n = (1, ..., 1)^T$, tai $\sum_{i=1}^n \hat{e}_i = 0$, iš čia išplaukia pirmoji (3.3.13) lygybė. Gauname

$$\boldsymbol{Y}^T\boldsymbol{Y} = (\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}})^T(\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}) = \hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}} + 2\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} = \hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}, \tag{3.3.15}$$

nes iš lygybės (3.3.14) išplaukia $\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{e}} = \hat{\beta}^T\boldsymbol{A}^T\hat{\boldsymbol{e}} = 0$. Lygybę (3.3.15) užrašykime taip:

$$\sum_{j=1}^n Y_j^2 = \sum_{j=1}^n \hat{Y}_j^2 + \sum_{j=1}^k (Y_j - \hat{Y}_j)^2.$$

Tada

$$\sum_{j=1}^n Y_j^2 - n\bar{Y}^2 = \sum_{j=1}^n \hat{Y}_j^2 - n\bar{Y}^2 + \sum_{j=1}^n (Y_j - \bar{Y}_j)^2,$$

iš kur išplaukia, kad

$$\sum_{j=1}^n (Y_j - \bar{Y})^2 = \sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2 + \sum_{j=1}^n (Y_j - \hat{Y}_j)^2.$$

Teorema įrodyta. ▲

3.3.5. Koeficientų $\boldsymbol{\beta}$ ir jų tiesinių darinių pasikliovimo intervalai

Priimkime papildomą prielaidą, kad nepriklausomi a. d. e_i yra normalieji $N(0, \sigma^2)$. Įvertinių savybės normaliuoju atveju pateiktos 1.3 poskyryje:

$$\hat{\boldsymbol{\beta}} \sim N_{m+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}), \quad \frac{s^2(n-m-1)}{\sigma^2} \sim \chi^2(n-m-1),$$

be to, atsitiktiniai dydžiai $\hat{\boldsymbol{\beta}}$ ir s^2 yra nepriklausomi.

Pagal (3.3.9) parametro $\theta = \boldsymbol{L}^T\boldsymbol{\beta}$ įvertinys $\hat{\theta} = \boldsymbol{L}^T\hat{\boldsymbol{\beta}}$ turi savybę

$$\frac{\hat{\theta} - \theta}{s\, b(\boldsymbol{L})} \sim S(n-m-1), \quad b^2(\boldsymbol{L}) = \boldsymbol{L}^T\boldsymbol{C}\boldsymbol{L},$$

čia $\boldsymbol{C} = (\boldsymbol{A}^T\boldsymbol{A})^{-1} = [c_{ij}]_{(m+1)\times(m+1)}$. Imdami $\theta = \beta_i$ ir $\theta = \mu(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ gauname

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim S(n-m-1), \quad \frac{\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})}{sb(\boldsymbol{x})} \sim S(n-m-1),\ b^2(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x}. \tag{3.3.16}$$

Parametro θ pasikliovimo intervalas, kai pasikliovimo lygmuo $Q = 1 - \alpha$, yra

$$\hat{\theta} \pm b\, s\, t_{\alpha/2}(n - m - 1). \tag{3.3.17}$$

Parametro β_i pasikliovimo intervalas yra

$$\hat{\beta}_i \pm \sqrt{c_{ii}} s\, t_{\alpha/2}(n - m - 1), \tag{3.3.18}$$

o parametro $\mu(\boldsymbol{x})$

$$\hat{\mu}(\boldsymbol{x}) \pm b(\boldsymbol{x})\, s\, t_{\alpha/2}(n - m - 1). \tag{3.3.19}$$

Jeigu reikia rasti pasikliovimo intervalų rinkinį regresijos kreivės reikšmėms taškuose $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(k)}$, formulėje (3.3.19) įrašoma paeiliui $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(k)}$ ir $t_{\alpha/2}(n - m - 1)$ pakeičiama į $t_{\alpha/(2k)}(n - m - 1)$ (remiamasi Bonferonio nelygybe), arba $((m + 1)F_\alpha(m + 1, n - m - 1))^{1/2}$ (naudojamas S metodas).

3.3.6. Naujo stebėjimo reikšmės prognozė

Analogiškai, kaip ir esant vienai kovariantei, sudarysime tolesnio nepriklausomo stebėjimo Y_{n+1}, atliekamo, kai kovariančių vektorius $\boldsymbol{x} = (1, x_1, ..., x_m)^T$, prognozės intervalą.

Atsitiktiniai dydžiai Y_{n+1} ir $\hat{\mu}(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ yra nepriklausomi ir

$$Y_{n+1} \sim N(\mu(\boldsymbol{x}), \sigma^2), \quad \hat{\mu}(\boldsymbol{x}) \sim N(\mu(\boldsymbol{x}), \sigma^2 b^2(\boldsymbol{x})),$$

taigi

$$Y_{n+1} - \hat{\mu}(\boldsymbol{x}) \sim N(0, \sigma^2(1 + b^2(\boldsymbol{x}))$$

ir

$$t = \frac{Y_{n+1} - \hat{\mu}(\boldsymbol{x})}{s\sqrt{(1 + b^2(\boldsymbol{x}))}} \sim S(n - m - 1).$$

Reikšmingumo lygmens $Q = 1 - \alpha$ prognozės intervalas reikšmei Y_{n+1} yra

$$\hat{\mu}(\boldsymbol{x}) \pm s\sqrt{(1 + b^2(\boldsymbol{x})}\, t_{\alpha/2}(n - m - 1).$$

Jis platesnis už pasikliovimo intervalą vidurkiui $\mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. Jeigu tolesnis stebėjimas Y_{n+1} yra žinomas, o reikia nurodyti kovariantę $\boldsymbol{x}$, kuriai esant jis buvo gautas, tai kovariantės $\boldsymbol{x}$ pasiklovimo sritis $\boldsymbol{B} \in \mathbf{R}^{m+1}$, kai pasikliovimo lygmeniu $Q = 1 - \alpha$, yra

$$\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{Y}, Y_{n+1}) = \{\boldsymbol{x} : |t| < t_{\alpha/2}(n - m - 1)\}.$$

3.3.7. Hipotezių apie regresijos parametrų reikšmes tikrinimas

Nagrinėkime hipotezę

$$H_{j_1...j_k} : \beta_{j_1} = \cdots = \beta_{j_k} = 0, \tag{3.3.20}$$

čia $1 \le j_1 \le \cdots \le j_k \le m$, k fiksuotas skaičius, $k = 1, ..., m$. Jei ši hipotezė teisinga, tai kovariantės $x_{j_1}, ..., x_{j_k}$ nėra reikšmingos priklausomo kintamojo prognozei ir jas galima išmesti iš modelio.

Atveju $k = 1$, $j_1 = j$ turime hipotezę

$$H_j : \beta_j = 0, \tag{3.3.21}$$

kuri reiškia, kad kovariantė x_j nėra reikšminga priklausomo kintamojo prognozei. Kai $k = m$, turime hipotezę

$$H_{1...m} : \beta_1 = ... = \beta_m = 0. \tag{3.3.22}$$

Ši hipotezė reiškia, kad tiesinės regresijos apskritai nėra. Žinant kovariančių reikšmes negaunama jokios papildomos informacijos apie Y reikšmes.

Hipotezėms tikrinti naudosimės 1.3.2 poskyrio 1.3.1 pavyzdžio rezultatais, iš kurių išplaukia, kad tiesiniame modelyje: $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ reikšmingumo lygmens α kritinė sritis hipotezei $H_{j_1...j_k}$ tikrinti turi pavidalą

$$F_{j_1,,,j_k} = \frac{SS_E^{(m-k)} - SS_E}{k\, s^2} > F_\alpha(k, n-m-1).$$

Čia $SS_E^{(m-k)}$ yra SS_E analogas modeliui be kovariančių $x_{j_1}, ..., x_{j_k}$:

$$SS_E^{(m-k)} = \sum_{i=1}^{n} (Y_i - \tilde{Y}_i)^2, \quad \tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_{s_1} x_{s_1} + ... + \tilde{\beta}_{s_{m-k}} x_{s_{m-k}};$$

$s_1, ..., s_{m-k}$ papildo $j_1, ..., j_k$ iki $1, 2, ..., m$, o $\tilde{\beta}_0, \tilde{\beta}_{s_1}, ..., \tilde{\beta}_{s_{m-k}}$ yra regresijos parametrų įvertiniai modelyje be kovariančių $x_{j_1}, ..., x_{j_k}$.

Hipotezės $H_j : \beta_j = 0$ atveju kriterijaus statistika yra

$$F_j = \frac{SS_E^{(m-1)} - SS_E}{s^2}.$$

Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai

$$F_j > F_\alpha(1, n-m-1).$$

Remiantis (3.3.16) šią hipotezę galima tikrinti ir naudojant Stjudento kriterijų. Hipotezė atmetama α lygmens kriterijumi, kai

$$|t| = \frac{|\hat{\beta}_j|}{s\sqrt{c_{jj}}} > t_{\alpha/2}(n-m-1).$$

Abu šie kriterijai yra ekvivalentūs.

Reikia pažymėti, kad hipotezės $H_{1...m}$ atveju

$$SS_E^{(0)} = \min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = SS_T, \quad SS_E^{(0)} - SS_E = SS_R,$$

čia SS_T yra pilnoji kvadratų suma, SS_R – regresijos kvadratų suma.

Taigi kriterijaus statistika hipotezei $H_{1\dots m}$ yra

$$F_{1\dots m} = \frac{SS_R/m}{SS_E/(n-m-1)} = \frac{MS_R}{MS_E} \sim F_{m,n-m-1}.$$

Regresijos nebuvimo hipotezė $H_{1\dots m}$ apie regresijos nebuvimą yra atmetama reikšmingumo lygmens α kriterijumi, kai

$$F_{1\dots m} > F_\alpha(m, n-m-1).$$

3.3.8. Determinacijos koeficientas

Priminsime, kad pilnoji kvadratų suma $SS_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$ apibūdina stebėjimų Y_i sklaidą apie jų aritmetinį vidurkį $\bar{Y}$; regresijos kvadratų suma $SS_R = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ – regresijos modeliu prognozuojamų reikšmių $\hat{Y}_i$ sklaidą apie $\bar{Y}$; liekamųjų paklaidų kvadratų suma $SS_E = \sum_{i=1}^n (Y_i - \hat{Y})^2$ parodo, kuri Y_i sklaidos dalis lieka nepaaiškinta regresiniu modeliu. Naudodami šias sumas apibrėšime prognozavimo (tiesiniu regresiniu modeliu) kokybės matą.

3.3.1 apibrėžimas. Atsitiktinis dydis

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T} \tag{3.3.23}$$

vadinamas *determinacijos koeficientu*.

Determinacijos koeficientas R^2 įgyja reikšmes iš intervalo $[0, 1]$. Jis parodo santykinę Y_i sklaidos dalį, paaiškinamą regresiniu modeliu, taigi apibūdina prognozavimo kokybę.

Jei prognozavimas idealus, t. y. $\hat{Y}_i = Y_i$, tai $SS_E = 0$ ir $R^2 = 1$. Jei nėra regresijos, t. y. su visais $\boldsymbol{x}^{(i)}$ vidurkio $\mu(\boldsymbol{x}^{(i)})$ prognozė nepriklauso nuo $\boldsymbol{x}_i$, tai $\hat{Y}_i = \bar{Y}$, taigi $SS_E = SS_T$ ir $R^2 = 0$. Tokiu būdu, R^2 charakterizuoja prognozavimo kokybę. Skirtingai nuo sumų SS_E ir SS_R, jo reikšmė nepriklauso nuo matavimo vienetų parinkimo.

3.3.2 apibrėžimas. Atsitiktinis dydis $R_{Y\boldsymbol{x}} = \sqrt{R^2}$ vadinamas *empiriniu dauginiu koreliacijos koeficientu*.

3.3.2 teorema. *Empirinis dauginis koreliacijos koeficientas lygus empiriniam stebėtų reikšmių Y_i ir prognozuojamų reikšmių $\hat{Y}_i$ koreliacijos koeficientui:*

$$R_{Y\boldsymbol{x}} = r_{Y\hat{Y}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

čia $\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^n \hat{Y}_i$.

Įrodymas. Remdamiesi (3.3.13) ir (3.3.14) gauname:

$$\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{e}} = 0, \quad \bar{\hat{Y}} = \bar{Y}.$$

Taigi

$$\sum_{i=1}^{n}(\hat{Y}_i-\bar{\hat{Y}})e_i=\sum_{i=1}^{n}\hat{Y}_ie_i=\hat{\boldsymbol{Y}}^T\boldsymbol{e}=0,$$

$$\sum_{i=1}^{n}(\hat{Y}_i-\bar{\hat{Y}})(Y_i-\bar{Y})=\sum_{i=1}^{n}(Y_i-\bar{\hat{Y}})(\hat{e}_i+\hat{Y}_i-\bar{\hat{Y}})=\sum_{i=1}^{n}(\hat{Y}_i-\bar{\hat{Y}})^2$$

ir

$$r_{Y\hat{Y}}=\sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=R_{Y\boldsymbol{x}}.$$

Teorema įrodyta.▲

3.3.1 pastaba. Jeigu imtis

$$(Y_1,X_1^{(1)},\dots,X_m^{(1)}),\dots,(Y_n,X_1^{(n)},\dots,X_m^{(n)}),$$

gauta stebint normalųjį a. v. $(Y_1,X_1,\dots,X_m)$, tai su bet kuriomis fiksuotomis kovariančių reikšmėmis $\boldsymbol{x}^{(i)}=(1,\,x_1^{(i)},\dots,x_m^{(i)})^T$ teisingas tiesinės regresijos modelis, todėl pagal 3.1.4 apibrėžimą *dauginio koreliacijos koeficiento kvadratas* $\rho^2_{Y\boldsymbol{X}}$ yra Y ir optimalios prognozės $\boldsymbol{\beta}^*\mathbf{X}$ koreliacijos koeficiento kvadratas, o pagal 3.3.2 teoremą *empirinio dauginio koreliacijos koeficiento kvadratas.* Taigi determinacijos koeficientą $R^2_{Y\boldsymbol{X}}=r^2_{Y\hat{Y}}$ galima interpretuoti kaip tiesinės regresijos tikslumo mato $\rho^2_{Y\boldsymbol{X}}$ įvertinį.

3.3.9. Empiriniai dalinės koreliacijos koeficientai

Nagrinėkime statistiką

$$R^2_{Y(k+1\dots m)|(1\dots k)}=\frac{SS_R^{(k)}-SS_R^{(m)}}{SS_R^{(k)}}=\frac{SS_E^{(m)}-SS_E^{(k)}}{SS_T-SS_E^{(k)}}.$$

Kvadratų sumos $SS_E^{(m)}$ ir $SS_E^{(k)}$ parodo Y_i reikšmių sklaidos dalis, paaiškintas regresiniu modeliu atitinkamai su m ir k kovariančių, taigi $SS_E^{(m)}-SS_E^{(k)}$ parodo Y_i reikšmių sklaidos dalį, paaiškintą pridedant prie $x_1,...,x_k$ papildomas kovariantes $x_{k+1}...,x_m$. Skirtumas $SS_R^{(k)}=SS_T-SS_E^{(k)}$ parodo likutinę Y_i reikšmių sklaidos dalį, nepaaiškintą modeliu su k kovariančių.

Taigi statistika $R^2_{Y(k+1\dots m)(1\dots k)}$ yra *dalis liekamosios sklaidos* modelyje su m kovariančių, *paaiškinta įtraukiant naujas kovariantes* x_{k+1}, ..., x_m.

3.3.3 apibrėžimas. Statistika $R_{Y(k+1\dots m)|(1\dots k)}=\sqrt{R^2_{Y(k+1\dots m)|(1\dots k)}}$ vadinama empiriniu dalinės koreliacijos koeficientu tarp Y ir $x_{k+1},...,x_m$ dalinės koreliacijos koeficientu.

3.3.3 teorema. *Empirinis dalinės koreliacijos koeficientas* $R_{Y(k+1\dots m)|(1\dots k)}$ *yra empirinių dauginių koreliacijos koeficientų* $R_{Y(1\dots k)}$ *ir* $R_{Y(1\dots m)}$ *funkcija:*

$$R^2_{Y(k+1\dots m)|(1\dots k)}=\frac{R^2_{Y(1\dots m)}-R^2_{Y(1\dots k)}}{1-R^2_{Y(1\dots k)}}.$$

Įrodymas. Teoremos rezultatas gaunamas iš sąryšių

$$SS_R^{(m)} = SS_T(1 - R^2_{Y(1\ldots m)}), \quad SS_R^{(k)} = SS_T(1 - R^2_{Y(1\ldots k)}).$$

▲

Atskiru atveju, empirinio Y ir x_m dalinės koreliacijos koeficiento kvadratas $R^2_{Ym(1\ldots k)}$ yra *dalis liekamosios sklaidos modelyje su* $(m-1)$ *kovariantėmis* $x_1, ..., x_{m-1}$*, paaiškinta įtraukus papildomą* m*-ją kovariantę* x_m. Turime

$$R^2_{Ym(1\ldots m-1)} = \frac{R^2_{Y(1\ldots m)} - R^2_{Y(1\ldots m-1)}}{1 - R^2_{Y(1\ldots m-1)}}. \tag{3.3.24}$$

3.3.4 teorema. *Teisinga lygybė*

$$1 - R^2_{Y(1\ldots m)} = \prod_{j=2}^{m} \left(1 - R^2_{Yj(1\ldots j-1)}\right) \left(1 - R^2_{Y(1)}\right)$$

Įrodymas. Iš vieneto atėmę abi (3.3.24) lygybės puses ir padauginę iš $1 - R^2_{Ym(1\ldots m)}$, gauname sąryšį

$$1 - R^2_{Y(1\ldots m)} = \left(1 - R^2_{Y(1\ldots m-1)}\right) \left(1 - R^2_{Ym(1\ldots m-1)}\right),$$

iš kurio remdamiesi indukcija gauname teoremos rezultatą. ▲

Teorema parodo, kokia dalimi sumažinama dar nepaaiškinta liekamosios sklaidos dalis, laipsniškai įtraukiant įvedant papildomas kovariantes.

3.3.2 pastaba. Jeigu imtis

$$(Y_1, X_1^{(1)}, \ldots, X_m^{(1)}), \ldots, (Y_n, X_1^{(n)}, \ldots, X_m^{(n)})$$

gauta stebint normalųjį a. v. $(Y_1, X_1, \ldots, X_m)$, tai su bet kuriomis fiksuotomis kovariančių reikšmėmis $\boldsymbol{x}^{(i)} = (1, x_1^{(i)}, ..., x_m^{(i)})^T$ teisingas tiesinės regresijos modelis, todėl *empirinio dalinio koreliacijos koeficiento kvadratą* $R^2_{Y(k+1,\ldots m)|(1\ldots k)}$ galima interpretuoti kaip *dalinės koreliacijos koeficiento kvadrato* $\rho^2_{Y(X_{k+1}\ldots X_m)|(X_1\ldots X_k)}$ įvertinį, jei visose formulėse x_i keičiame į X_i.

3.3.10. Regresijos tiesinio pavidalo hipotezės tikrinimas

Tarkime, kad $\boldsymbol{x}^{(i)} = (1,\ x_{1i}, ..., x_{mi})^T,\ i = 1, ..., k$, yra skirtingos kovariantės reikšmės ir, kai yra reikšmė $\boldsymbol{x}^{(i)}$, yra gauta n_i nepriklausomų a. d. Y stebėjimų $Y_{ij},\ j = 1, ..., n_i,\ n = n_1 + \cdots + n_k$.

Sakykime Y regresija $\boldsymbol{x}$ atžvilgiu yra bet kokio pavidalo ir $\mu_i = \mathbf{E}(Y|\boldsymbol{x}^{(i)})$, $i = 1, ..., k$. Taigi nagrinėjame tiesinį modelį

$$Y_{ij} = \mu_i + e_{ij}, \quad i = 1, ..., k$$

Jeigu teisinga prielaida apie regresijos tiesinį pavidalą, tai vidurkių vektorius $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)^T$ yra parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ tiesinė funkcija, t. y. teisinga hipotezė

$$H : \mu_i = \boldsymbol{\beta}^T \boldsymbol{x}^{(i)}, \quad i = 1, ..., k.$$

Formaliai tikrinsime, ar duomenys neprieštarauja hipotezei H. Pavyzdyje 1.3.3 (ir jo tęsinyje) kritinė sritis turi pavidalą

$$F = \frac{(SS_{EH} - SS_E)/(k - m - 1)}{SS_E/(n - k)}; \tag{3.3.25}$$

čia $F_\alpha(k - m - 1, n - k)$ – Fišerio skirstinio α kritinė reikšmė, o

$$SS_E = \min_{\boldsymbol{\mu}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2, \quad \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

Sąlyginis kvadratinės formos minimumas, kai hipotezė teisinga, yra

$$SS_{EH} = \min_{\boldsymbol{\mu}: \mu_i = \boldsymbol{\beta}^T \boldsymbol{x}^{(i)}} SS(\boldsymbol{\mu}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^{(i)})^2.$$

Skirtumas

$$(SS_{EH} - SS_E) = \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}^{(i)})^2.$$

3.3.3 pastaba. Jeigu su kiekvienu $\boldsymbol{x}^{(i)}$ turime tik vieną stebėjimą Y_i, tai kriterijaus (3.3.25) pritaikyti negalima, nes $SS_E = 0$. Kaip ir turint vieną regresorių, tam tikros informacijos apie regresijos pavidalą suteikia likutiniai skirtumai

$$e_i = Y_i - \hat{Y}_i, \quad \hat{Y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}^{(i)}, \quad i = 1, ..., n.$$

Pažymėję

$$\hat{\boldsymbol{Y}} = (\hat{Y}_1, ..., \hat{Y}_n)^T, \quad \hat{\boldsymbol{e}} = (\hat{e}_1, ..., \hat{e}_n)^T = \boldsymbol{Y} - \hat{Y},$$

gauname

$$\hat{\boldsymbol{e}} = \boldsymbol{D}\boldsymbol{Y}, \quad \boldsymbol{D} = \boldsymbol{I}_n - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T = [d_{ij}]_{n\times n}, \quad \mathbf{E}(\hat{\boldsymbol{e}}) = \mathbf{0}, \quad \boldsymbol{V}(\hat{\boldsymbol{e}}) = \sigma^2 \boldsymbol{D}.$$

Taigi

$$\hat{e}_i \sim N(0, \sigma^2 d_{ii}).$$

Kadangi

$$\hat{\sigma}^2 = MS_E = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n - m - 1} \xrightarrow{P} \sigma^2, \quad n \to \infty,$$

tai

$$\tilde{e}_i = \frac{\hat{e}_i}{\sqrt{MS_E d_{ii}}} \xrightarrow{d} Z \sim N(0, 1).$$

Kai regresijos tiesiškumo prielaida teisinga, $\tilde{e}_i$ skirstiniai mažai skiriasi nuo standartinio normaliojo skirstinio. Aišku, net kai n yra didelis, vektoriaus $(\tilde{e}_1, ..., \tilde{e}_n)^T$ negalime traktuoti kaip paprastąją imtį, nes a. d. $\tilde{e}_i$ yra priklausomi. Daugelyje statistinių programų paketų naudojami neformalūs diagnostiniai grafiniai metodai modelio adekvatumui tikrinti, remiantis standartizuotomis liekanomis. Šiame vadovėlyje neformalūs metodai neaptariami, informacijos apie juos galima rasti taikomojo pobūdžio knygose [4],I dalis; [5].

3.3.11. Pažingsninė regresija

Dažnai, taikydamas regresinę analizę, eksperimentuotojas neturi pakankamai informacijos, kurie iš kintamųjų $X_1, ..., X_m$ labiau tinka a. d. Y prognozuoti. Be to, galutinai parinktas modelis yra tuo vertingesnis, kuo mažesniu skaičiumi kovariančių jis nusakomas. Suprantama, stengiamasi, kad kintamųjų $X_1, ..., X_m$ skaičiaus sumažinimas iš esmės nepablogintų prognozės tikslumo. Matematinėje statistikoje naudojama daug įvairių metodų, leidžiančių sumažinti vektoriaus $(X_1, ..., X_m)^T$ dimensiją iš esmės nepabloginant prognozės tikslumo. Pateiksime vieną iš paprastesnių ir palyginti dažnai naudojamų metodų.

Sakykime, vektorius $(Y, X_1, ..., X_m)^T$ pasiskirstęs pagal $(m+1)$-matį normalųjį skirstinį. Stebint šį a. v. gauta didumo n imtis $(Y_i, X_{1i}, ..., X_{mi})^T$, $i = 1, ..., n$. Kadangi tiesinio prognozavimo tikslumo matas yra (3.1.7) dauginio koreliacijos koeficiento kvadratas, tai būtų natūralu parinkti tokį kovariančių rinkinį, kad dauginio koreliacijos koeficiento įvertinys būtų kuo didesnis. Tačiau tokio metodo negalima realizuoti praktiškai, jeigu kovariančių skaičius gana didelis. Pavyzdžiui, jeigu $m = 10$, tai galima sudaryti $2^{10} - 1 = 1023$ skirtingų tiesinės regresijos lygčių (neatsižvelgiant į galimas kovariančių sąveikas). Vienas iš paprasčiausių ir dažniausiai naudojamų metodų kovariančių rinkiniui parinkti yra pažingsninė regresija.

N u l i n i s ž i n g s n i s. Parenkame du reikšmingumo lygmenis P ir P'; randame a. d. Y ir X_i koreliacijos koeficientų įverčius $r(Y, X_i) = r_{YX_i} = r_{0i}$, $i = 1, ..., m$.

P i r m a s i s ž i n g s n i s. Išrenkame maksimalų įvertį

$$\max_{1 \le i \le m} |r_{0i}| = |r_{0i_1}|.$$

Kintamasis X_{i_1} įtraukiamas į kintamųjų Y prognozuoti sąrašą, jeigu teisinga nelygybė

$$F_{0i_1} = r^2_{0i_1}(n-2)/(1 - r^2_{0i_1}) > F_P(1, n-2).$$

Priešingu atveju daroma išvada, kad Y prognozuoti pagal $X_1, ..., X_m$ negalima.

A n t r a s i s ž i n g s n i s. Randame dalinių koreliacijos koeficientų įverčius

$$r_{YX_i(X_{i_1})} = r_{0i(i_1)}$$

ir iš jų išrenkame maksimalų:

$$\max_{i \ne i_1} |r_{0i(i_1)}| = |r_{0i_2(i_1)}|.$$

Kintamasis X_{i_2} įtraukiamas į sąrašą, jeigu teisinga nelygybė

$$r^2_{0i_2(i_1)}(n-3)/(1-r^2_{0i_2(i_1)}) > F_P(1,\ n-3). \tag{3.3.26}$$

Priešingu atveju sąraše paliekamas tik kintamasis X_{i_1}.

Jeigu kintamasis X_{i_2} į sąrašą įtrauktas, tikrinama, ar nereikia išbraukti iš sąrašo X_{i_1}, kai į jį įrašytas X_{i_2}. Kintamasis X_{i_1} išbraukiamas, jeigu teisinga nelygybė

$$r^2_{0i_1(i_2)}(n-3)/(1-r^2_{0i_1(i_2)}) < F_{P'}(1,\ n-3). \tag{3.3.27}$$

Taigi po antrojo žingsnio a. d. Y prognozuoti gali likti tik kintamasis X_{i_1} (nelygybė (3.3.26) neteisinga); lieka tik kintamasis X_{i_2} (nelygybė (3.3.26) ir (3.3.27) teisingos); vektorius $(X_{i_1},\ X_{i_2})^T$ (nelygybė (3.3.26) teisinga, o nelygybė (3.3.27) neteisinga). Pirmu atveju procesas užbaigiamas; antru ir trečiu atvejais pereinama prie trečiojo žingsnio.

T r e č i a s i s, k e t v i r t a s i s, ... ž i n g s n i a i. Rekurentiškai kartojamas antrasis žingsnis. Tiksliau, remiantis dalinių koreliacijos koeficientų įverčiais, pagal taisykles, analogiškas (3.3.26), daroma išvada, ar reikia papildyti kintamųjų sąrašą dar vienu kintamuoju. Įrašius naują kintamąjį, tikrinama, ar nereikia kurio nors kintamojo išbraukti iš sąrašo. Procesas užbaigiamas, kai nelygybės, analogiškos (3.3.26), (3.3.27), netenkinamos.

3.4. Pratimai

3.1 skyrelis

3.1. Vektorius $(X,\ Y)^T$ pasiskirstęs pagal dvimatį normalųjį skirstinį su nuliniais vidurkiais, vienetinėmis dispersijomis ir koreliacijos koeficientu ρ. Raskite regresijos kreives:
(a) a. d. X atžvilgiu Y;
(b) a. d. X atžvilgiu Y^2;
(c) a. d. X^2 atžvilgiu Y^2;
(d) a. d. X^2 atžvilgiu Y.
Apskaičiuokite koreliacinius santykius ir palyginkite juos su koreliacijos koeficientų kvadratais.

3.2. Atsitiktinis vektorius $(X,\ Y)^T$ turi tolygųjį skirstinį viduje elipsės

$$ax^2 + 2hxy + by^2 = c, \quad h \neq 0, \quad h^2 < a\,b; \quad a,\ b > 0.$$

Įrodykite, kad a. d. X ir Y regresija kito kintamojo atžvilgiu yra tiesinė.

3.3. Atsitiktinis vektorius $(X,\ Y)^T$ tolygiai pasiskirstęs lygiagretainyje, apribotame tiesėmis $x = 3(y+1)$, $x = 3(y-1)$, $x = y+1$, $x = y-1$. Įrodykite, kad atsitiktinio dydžio Y regresija X atžvilgiu yra tiesinė, o X regresija Y atžvilgiu yra kreivė, susidedanti iš trijų atkarpų.

3.4. Atsitiktinis vektorius $(X,Y)^T$ tolygiai pasiskirstęs ant pusapskritimio $y = +\sqrt{1-x^2}$, $0 \le x \le 1$. Įrodykite, kad $\eta^2_{YX} = 1$, o $\rho^2_{YX} = 0$.

3.5. Atsitiktinis vektorius $(X,Y)^T$ tolygiai pasiskirstęs skritulyje $(x-a)^2 + (y-b)^2 \le r^2$. Įrodykite, kad $\eta^2_{YX} = \rho^2_{YX} = 0$.

3.6. Tegu $\boldsymbol{\rho} = [\rho_{ij}]_{k\times k}$ yra koreliacijos koeficientų matrica. Įrodykite, kad dauginis koreliacijos koeficientas ir daliniai koreliacijos koeficientai susieti lygybėmis:

$$1-\rho^2_{1.(2\ldots k)} = (1-\rho^2_{12})(1-\rho^2_{13.2})\ldots(1-\rho^2_{1k.2\ldots k-1});$$

$$\rho_{12.3} = \frac{(\rho_{12}-\rho_{13}\rho_{23})}{(1-\rho^2_{13})^{1/2}(1-\rho^2_{23})^{1/2}};$$

$$\rho_{12} = \frac{(\rho_{12.3}-\rho_{13.2}\rho_{23.1})}{(1-\rho^2_{13.2})^{1/2}(1-\rho^2_{23.1})^{1/2}};$$

$$\rho_{12.(3\ldots k-1)} = \frac{(\rho_{12.(3\ldots k)}-\rho_{1k.(2\ldots k-1)}\rho_{2k.(13\ldots k-1)})}{(1-\rho^2_{1k.(2\ldots k-1)})^{1/2}(1-\rho^2_{2k.(13\ldots k-1)})^{1/2}};$$

$$\rho_{12.(3\ldots k)} = \frac{(\rho_{12.(3\ldots k-1)}-\rho_{1k.(3\ldots k-1)}\rho_{2k.(3\ldots k-1)})}{(1-\rho^2_{1k.(2\ldots k-1)})^{1/2}(1-\rho^2_{2k.(3\ldots k-1)})^{1/2}};$$

3.2 skyrelis

3.7. Pagal didumo n imtį $Y_i = \alpha+\beta(X_i-\bar{X})+e_i$, $i = 1,\ldots,n$, kai a. d. $\{e_i\}$ nepriklausomi ir $e_i \sim N(0,\sigma^2)$ gauti įvertiniai $\hat{\alpha}$ ir $\hat{\beta}$. Sukurkite TGN kriterijų hipotezei $H : \theta = a\alpha+b\beta = \theta_0$, kai alternatyva $\bar{H} : \theta \neq \theta_0$, tikrinti; čia a, b, θ_0 – žinomi.

3.8. Pagal dvi nepriklausomas didumo n_1 ir n_2 imtis įvertinti tiesinės vieno kintamojo regresijos koeficientai $\hat{\alpha}_1$, $\hat{\beta}_1$ ir $\hat{\alpha}_2$, $\hat{\beta}_2$. Tarkime, kad visi stebėjimai yra nepriklausomi ir normalieji, turi vienodas dispersijas σ^2. Raskite kriterijus

(a) hipotezei $H : \beta_1 = \beta_2$ tikrinti;
(b) hipotezei $H : \alpha_1 = \alpha_2$ tikrinti;
(c) hipotezei $H : \alpha_1 = \alpha_2$, $\beta_1 = \beta_2$ tikrinti;
(d) hipotezei, kad regresijos tiesės susikerta taške $x = c$, tikrinti;

3.9 (**3.8** tęsinys). Apibendrinkite **3.8** pratimą tuo atveju, kai stebėjimo rinkinių skaičius didesnis už du.

3.10. Raskite parametrų α, β ir σ^2 įverčius naudodami šiuos stebėjimus:

$$0,15 = \alpha - 3\beta + e_1,$$
$$2,07 = \alpha - \beta + e_2,$$
$$4,31 = \alpha + \beta + e_3,$$
$$6,49 = \alpha + 3\beta + e_4,$$

čia $e_1,\ldots,e_4$ – nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai $e_i \sim N(0,\sigma^2)$, $i = 1,\ldots,4$.

3.11. Taškas tolygiai juda tiese. Laiko momentais $t = 0,\ 1,\ 2,\ 3,\ 4$ buvo užfiksuotos tokios jo koordinačių reikšmės: $S_t = 12,98;\ 13,05;\ 13,35;\ 13,97;\ 14,22$. Tegu visų matavimų paklaidos nepriklausomos ir turi normalųjį skirstinį $N(0,\ \sigma^2)$. Raskite greičio v ir dispersijos σ^2 taškinius ir intervalinius įverčius, kai pasikliovimo lygmuo $Q = 0,95$.

3.12. Matuojant gauti stebėjimai, kurie yra tiesinės parametro α funkcijos:

$$x_1 = -0,42 = e_1,$$
$$x_2 = 0,30 = \alpha + e_1 + e_2,$$
$$x_3 = 1,30 = 2\alpha + e_1 + e_2 + e_3,$$
$$x_4 = 2,56 = 3\alpha + e_1 + e_2 + e_3 + e_4,$$
$$x_5 = 3,26 = 4\alpha + e_1 + e_2 + e_3 + e_4 + e_5,$$

čia $e_1,\ldots,e_5$ – nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai $e_i \sim N(0,\ \sigma^2)$, $i = 1,\ldots,5$. Raskite parametro α NMD įvertį. Sudarykite parametrų α ir σ^2 pasikliovimo intervalus, kai pasikliovimo lygmuo $Q = 0,95$.

3.13. Atsitiktinis vektorius $\mathbf{X} = (X_1, ..., X_5)^T$ turi normalųjį skirstinį su parametrais $\mathbf{E}X_i = (i-1)\alpha,\ i = 1, ..., 5;\ \boldsymbol{V}X_i = \mathbf{Cov}\,(X_i,\ X_j) = i,\ j > i,\ i = 1, ..., 5$. Raskite parametro α pasikliovimo intervalą, kai pasikliovimo lygmuo $Q = 0,95$, pagal atsitiktinio vektoriaus $\mathbf{X}$ realizaciją $\boldsymbol{x} = (-0,42;\ 0,30;\ 1,30;\ 2,56;\ 3,26)^T$.

3.14. Išmatuotas grunto atsparumas poslinkiui Y, veikiant įvairiems slėgiams $P\ kg/cm^2$ ir gauti šitokie rezultatai:

P_i	Y								
	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}	Y_{i6}	Y_{i7}	Y_{i8}	Y_{i9}
1	0,962	0,612	0,612	0,888	1,038	1,188	0,812	0,988	0,875
2	0,962	0,688	0,662	1,112	1,118	1,325	0,988	1,025	1,075
3	1,125	0,800	0,700	0,900	1,475	1,400	1,150	1,175	1,300
4	1,412	0,850	0,650	1,075	1,312	1,425	1,250	1,175	1,162
5	1,125	0,900	0,675	1,225	1,225	1,650	1,025	1,075	1,500
6	1,138	0,950	0,625	1,500	1,375	1,588	1,025	1,150	1,100

a) patikrinkite Y regresijos P atžvilgiu tiesiškumo hipotezę;
b) laikydami Y regresiją atžvilgiu P tiesine, įvertinkite jos koeficientus;
c) sudarykite regresijos koeficientų pasikliovimo intervalus ($Q = 0,95$).

3.15. Sąlyginis Y skirstinys, kai X reikšmės fiksuotos, yra normalusis su dispersija $\sigma^2 = 0,1$. Tarkime, kad taškuose $x = 0,0;\ 0,1; ...; 1,0$ turimos vienodo didumo n nepriklausomos imtys. Koks turi būti n, kad regresijos tiesiškumo hipotezė būtų atmesta su tikimybe, ne mažesne už 0,95, jeigu $\mathbf{E}(Y|X = x) = x^2$ (kriterijaus reikšmingumo lygmuo $P = 0,05$).

3.16. Tarkime, kad pagal nepriklausomus stebėjimus $Y_i = \alpha + \beta(x_i - \bar{x}) + e_i,\ i = 1, ..., n$, kai a. d. $\{e_i\}$ nepriklausomi ir normalūs $e_i \sim N(0, \sigma^2)$, įvertinta regresijos tiesė $\hat{\alpha} + \hat{\beta}(x - \bar{x})$. Tegu taške x yra gauta k nepriklausomų Y_{n+1} matavimų $Y_{n+1}^{(1)}, ..., Y_{n+1}^{(k)}$. Sukonstruokite stebėjimo Y_{n+1} prognozės intervalą ir argumento x pasikliovimo intervalą.

3.17. Pasiūlykite, kaip pertvarkyti lygtį

$$y = \frac{\alpha\beta}{\alpha \sin^2 x + \beta \cos^2 x}$$

į tiesinį pavidalą.

3.18. Kalibruojamas pieno rūgšties koncetracijos kraujuje matavimo prietaisas. Tuo tikslu gauti $n = 20$ pavyzdžių matavimai Y, kai koncentacija X tiksliai žinoma [1]

X	1	1	1	1	3	3	3	3	3	5
Y	1,1	0,7	1,8	0,4	3,0	4,4	4,9	4,4	4,5	7,3
X	5	5	10	10	10	10	15	15	15	15
Y	8,2	6,2	12,0	13,1	12,6	13,2	18,7	19,7	17,4	17,1

Tardami, kad paklaidos normaliosios, įvertinkite tiesinės regresijos parametrus. Patikrinkite regresijos tiesiškumo hipotezę. Raskite tolesnių nepriklausomų Y matavimų prognozės intervalus, kai pasikliovimo lygmuo $Q = 0,95$, jei žinoma, kad matavimai bus atliekami taškuose $X = 12$, $X = 18$.

3.19. Pamatavus $n = 108$ pacientų arterinį PH (kintamasis Y) ir veninį PH (kintamasis X) gauti vidurkių ir kovariacijų matricos elementų įverčiai: $\bar{X} = 7,373$, $\bar{Y} = 7,413$, $s_X^2 = 0,1253, s_Y^2 = 0,1184, s_{XY} = 0,1101$ [1]. Priėmę prielaidą dėl a.v. $(X, Y)^T$ normalumo, įvertinkite a. d. Y regresijos tiesę atžvilgiu a. d. X. Patikrinkite hipotezę $H : \beta = 0$. Kokią dalį a. d. Y sklaidos paaiškina regresijos lygtis?

3.20. Kintamasis Y reiškia tam tikro cheminio proceso savybę priklausomai nuo laiko t [1]:

t	0,0	1,0	1,5	2,0	2,5	3,0	3,5	4,0
Y	0,000	0,025	0,035	0,045	0,055	0,065	0,075	0,082
t	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0
Y	0,088	0,094	0,100	0,105	0,110	0,115	0,120	0,125

a) Įvertinkite tiesinės regresijos parametrus. Panagrinėkite prognozės paklaidas ir aptarkite regresijos tiesinio pavidalo priimtinumą. b) Tarkime, kad iš teorinių samprotavimų gauta, jog regresijos kreivė turėtų būti tokio pavidalo: $Y = \alpha(1-e^{-\beta t})$. Įvertinkite netiesinio modelio parametrus ir raskite jų apytikslius pasikliovimo intervalus.

3.21. Farmakinetikoje vertinant vaisto koncentracijos Y priklausomybę nuo laiko t dažnai naudojamasmas dvikomponentis matematinis modelis $Y = \alpha_1 e^{-\alpha_2 t} + \beta_1 e^{-\beta_2 t}$. Pagal pateikiamus duomenis [1]

t	0,10	0,25	0,50	1,00	1,50	2,00	4,00	8,00	12,00	24,00	48,00
Y	18,7	16,9	14,5	11,1	8,9	7,5	5,2	3,6	2,6	1,0	0,2

įvertinkite regresijos parametrus. Nagrinėdami prognozės paklaidas aptarkite prognozės tikslumą.

3.3 skyrelis

3.22. Lentelėje pateiktas 78 žuvų ilgis Y (milimetrais) ir amžius X (metais) [16].

X	Y										
1	67	62									
2	109	83	91	88	123	100	109				
3	137	131	122	122	118	115	131	143	142	122	140
3	150	140	150	150	140	150	130	130			
4	138	135	146	146	145	145	144	140	150	152	157
4	155	153	154	158	162	161	162	165	171	162	169
4	167	171	150	170	140	140	150	150	150	160	150
4	150	150	150	140	160	170	160	160	170		
5	170	188	150	150	160	160	180				
6	170										

a) patikrinkite Y regresijos X atžvilgiu tiesiškumo hipotezę;
b) parinkite polinominės regresijos modelį ir aptarkite prognozės tikslumą.

3.23. Tarkime, kad įvertinome regresiją $\mathbf{E}(Y) = \beta_0 + \beta_1 x$, o tikroji regresijos lygtis yra $\mathbf{E}(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ pavidalo. Koks bus įvertinių $\hat{\beta}_0$ ir $\hat{\beta}_1$ poslinkis, kai jie įvertinti pagal nepriklausomus stebėjimus taškuose $x = -3, -2, -1, 0, +1, +2, +3$?

3.24. Tegu $Y_1 = \theta_1 + \theta_2 + e_1$, $Y_2 = 2\theta_2 + e_2$, $Y_3 = -\theta_1 + \theta_2 + e_3$; čia $\{e_i\}$ nepriklausomi ir $e_i \sim N(0, \sigma^2)$. Sudarykite kriterijų hipotezei $H : \theta_1 = 2\theta_2$ tikrinti.

3.25. Tegu Y_1, Y_2, Y_3, Y_4 yra keturkampio kampų $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ matavimai. Tarkime, kad matavimo paklaidos yra normalieji nepriklausomi a. d., turintys nulinius vidurkius ir vienodas dispersijas. Patikrinkite hipotezę, kad keturkampis yra lygiagretainis, t. y. $H : \varphi_1 = \varphi_3, \varphi_2 = \varphi_4$.

3.26. Įrodykite, kad pilno rango modelį

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi} + e_1$$

įvedus naujus parametrus galima užrašyti šitaip

$$Y_i = \gamma_0 + \gamma_1 Z_{1i} + ... + \gamma_m Z_{mi} + e_i,$$

kad plano matrica turėtų ortogonalius stulpelius ir $\gamma_r = ... = \gamma_m = 0$ tada ir tik tada, kai $\beta_r = ... = \beta_m = 0$, $r = 1, ..., m$.

3.27. Tarkime, $\mathbf{E}(Y_t) = \beta_0 + \beta_1 \cos(2\pi k_1 t/n) + \beta_2 \sin(2\pi k_2 t/n)$; čia $t = 1, ..., n$, o k_1 ir k_2 – žinomos teigiamos konstantos. Raskite parametrų $\beta_0, \beta_1, \beta_2$ mažiausiųjų kvadratų įvertinius.

3.28. Tegu $Y_i = \beta_0 + \beta_1(X_{1i} - \bar{X}_{1.}) + \beta_2(X_{2i} - \bar{X}_{2.})$, $i = 1, ..., n$. Įrodykite, kad parametrų $\beta_0, \beta_1, \beta_2$ įvertinius galima rasti atlikus tokią dvižingsnę procedūrą: 1) parenkame vieno kintamojo regresiją $Y_i = \beta_0 + \beta_1(X_{1i} - \bar{X}_{1.})$; 2) parenkame liekanų $Y_i - \hat{Y}_i$, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(X_{1i} - \bar{X}_{1.})$ regresiją $X_{2i} - \bar{X}_{2i}, i = 1, ..., n$, atžvilgiu.

3.29. Nagrinėjamas benzino oktaninis skaičius Y priklausomai nuo priedų A ir B koncentracijų X_1 ir X_2. Pagal pateikiamus duomenis [1]

X_1	2	2	2	2	3	3	3	3
X_2	2	3	4	5	2	3	4	5
Y	96,3	95,7	99,9	99,4	95,1	97,8	99,3	104,9
X_1	4	4	4	4	5	5	5	5
X_2	2	3	4	5	2	3	4	5
Y	96,2	100,1	103,2	104,3	97,8	102,2	104,7	108,8

a) Įvertinkite regresijos parametrus tardami, kad a. d. Y regresija X_1 ir X_2 atžvilgiu yra tiesinė. b) Tarę, kad paklaidos normaliosios, patikrinkite regresijos koeficientų lygybės 0 hipotezes. c) Raskite a. d. Y tolesnio nepriklausomo stebėjimo, kuris atliktas taške $X_1 = 4,5; X_2 = 3,5$, prognozės intervalą, kai pasikliovimo lygmuo $Q = 0,95$.

3.30. Dviem skirtingais būdais buvo matuojamas 141 paciento arterinis kraujo spaudimas. Pirmu būdu matuojant sistolinį X_1, diastolinį X_2 ir vidutinį X_3 kraujo spaudimą buvo naudojamas kateteris. Antru būdu buvo matuojamas sistolinis X_4 ir diastolinis X_5 kraujo spaudimas naudojant kompresinę manžetę. Atlikus pradinę analizę gauti šie rezultatai.

i	$\bar{X}_i$	s_i	r_{ij}				
1	112,2	28,6	1,000	0,839	0,927	0,871	0,753
2	59,4	17,1		1,000	0,967	0,778	0,828
3	76,8	21,0			1,000	0,845	0,852
4	107,0	28,9				1,000	0,837
5	66,8	19,3					1,000

Šioje lentelėje $\bar{X}_i$ – aritmetiniai vidurkiai, o s_i – vidutinių kvadratinių nuokrypių įvertiniai, $i = 1, ..., 5$; r_{ij} - koreliacijos koeficientų įvertiniai.

Naudodamidami pažingsninės regresijos metodą, parinkite vektoriaus $(X_1, X_2, X_3)^T$ koordinates kintamiesiems X_4 ir X_5 prognozuoti ($P' = 0,01$, $P = 0,005$). Raskite regresijos koeficientus ir suvestinius koreliacijos koeficientus.

3.31. Reikia parinkti tokio pavidalo regresijos modelį:

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 \varphi(X_i), \quad i = 1, 2, 3;$$

čia $\varphi(x)$ – antrojo laipsnio polinomas. Parinkite $\varphi(x)$ taip, kad plano matrica turėtų ortogonalius stulpelius, kai $X_1 = -1, X_2 = 0, X_3 = +1$.

3.32. Pagal pateikiamos lentelės duomenis parinkite trečiojo laipsnio polinomą.

Y (indeksas)	9,8	11,0	13,2	15,1	16,0
X (metai)	1950	1951	1952	1953	1954

Tardami, kad paklaidos nepriklausomos ir normaliosios su vienodomis dispersijomis σ^2, patikrinkite antrojo laipsnio polinomo adekvatum hipotezę. Išspręskite uždavinį naudodami ortogonaliuosius polinomus, t. y. vietoje x^i parinkdami i-ojo laipsnio polinomą taip, kad plano matricos stulpeliai būtų ortogonalūs.

3.5. Atsakymai ir nurodymai

3.1. a) $\mathbf{E}(X|Y=y)=\rho y$; $\eta^2_{X|Y}=\rho^2_{X|Y}=\rho^2$; b) $\mathbf{E}(X|Y^2=y^2)=0$; $\eta^2_{X|Y}=\rho^2_{X|Y}=0$; c) $\mathbf{E}(X^2|Y^2=y^2)=1-\rho^2+\rho^2y^2$; d) $\mathbf{E}(X^2|Y=y)=1-\rho^2+\rho^2y^2$; $\eta^2_{X|Y}=\rho^2_{X|Y}=\rho^4$. **3.3.** $\mathbf{E}(Y|X=x)=2x/3$, kai $-3\le x\le 3$; $\mathbf{E}(X|Y=y)=2y+1$, kai $-2\le y\le -1$; $\mathbf{E}(X|Y=y)=y$, kai $-1\le y\le 1$; $\mathbf{E}(X|Y=y)=2y-1$, kai $1\le x\le 2$. **3.7.** Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai $|\hat\theta-\theta_0|/(sc)>t_{\alpha/2}(n-2)$; $\hat\theta=a\hat\alpha+b\hat\beta$, $s^2=\sum_i(Y_i-\hat\alpha-\hat\beta(x_i-\bar x))^2/(n-2)$; $c^2=a^2/n+b^2/\sum_i(x_i+\bar x)^2$. **3.8.** Tegu pirmosios imties elementai yra (Y_{1i},X_{1i}), $i=1,...,n_1$, o antrosios – (Y_{2i},X_{2i}), $i=1,...,n_2$. Vertinamos regresijos tiesės pavidalo: $\alpha_1+\beta_1(X_{1i}-\bar X_{1.})$ ir $\alpha_2+\beta_2(X_{2i}-\bar X_{2.})$. a) Hipotezė atmetama α lygmens kriterijumi, kai $|\hat\beta_1-\hat\beta_2|/(sc)>t_{\alpha/2}(n_1+n_2-4)$; $s^2=[\sum_i(Y_{1i}-\hat\alpha_1-\hat\beta_1(X_{1i}-\bar X_{1.}))^2+\sum_j(Y_{2i}-\hat\alpha_2-\hat\beta_2(X_{2i}-\bar X_{2.}))^2]/(n_1+n_2-4)$, $c^2=1/\sum(X_{1i}-\bar X_{1.})^2+1/\sum(X_{2i}-\bar X_{2.})^2$. b) Hipotezė atmetama α lygmens kriterijumi, kai $|\hat\alpha_1-\hat\alpha_2|/(sb)>t_{\alpha/2}(n_1+n_2-4)$; $b^2=1/n_1+1/n_2$. c) Hipotezė atmetama α lygmens kriterijumi, kai $(SS_{EH}-SS_E)(n_1+n_2-4)/(2SS_E)>F_\alpha(2,n_1+n_2-4)$; $SS_E=s^2(n_1+n_2-4)$; $SS_{EH}=[\sum_i(Y_{1i}-\hat\alpha-\hat\beta(X_{1i}-\bar X_{..}))^2+\sum_i(Y_{2i}-\hat\alpha-\hat\beta(X_{2i}-\bar X_{..}))^2]$, čia $\bar X_{..}=(n_1\bar X_{1.}+n_2\bar X_{2.})/(n_1+n_2)$, o $\hat\alpha$ ir $\hat\beta$ yra regresijos tiesės $\alpha+\beta(x-\bar X_{..})$ parametrų įvertiniai, gauti sujungus visus n_1+n_2 stebėjimus. d) perkelkime koordinatų pradžią: $Z_{1i}=X_{1i}-c$, $i=1,...,n_1$ ir $Z_{2j}=X_{2j}-c$, $j=1,...,n_2$. Kintamųjų (Y_{1i},Z_{1i}) ir (Y_{2j},Z_{2j}) terminais tikrinamoji hipotezė ekvivalenti p. b) hipotezei. **3.10.** $\hat\alpha=(Y_1+Y_2+Y_3+Y_4)/4=3,255$; $\hat\beta=(-3Y_1-Y_2+Y_3+3Y_4)/20=1,063$, $s^2=SS_E/2=0,01206$. **3.11.** $\hat v=0,34$, $\hat\sigma^2=0,0259$; $(\underline v;\overline v)=(0,178;0,502)$; $(\underline{\sigma^2};\overline{\sigma^2})=(0,0083;0,3602)$. **3.12.** $\hat\alpha=0,92$; $s^2=0,0967$; $(\underline\alpha;\overline\alpha)=(0,4883;1,3517)$; $(\underline{\sigma^2};\overline{\sigma^2})=(0,0347;0,7985)$. *Nurodymas.* A. d. $Z_1=X_1, Z_j=X_j-X_{j-1}, j=2,3,4,5$ yra nepriklausomi su vienetinėmis dispersijomis ir vidurkiais $\mathbf{E}Z_1=0, \mathbf{E}Z_2=...=\mathbf{E}Z_5=\alpha$. **3.13.** $\hat\alpha=0,92$; $(\underline\alpha;\overline\alpha)=(-0,060;1,900)$. *Nurodymas.* A. d. $Z_1=X_1, Z_j=X_j-X_{j-1}, j=2,3,4,5$ yra nepriklausomi; jų dispersijos σ^2, o vidurkiai $\mathbf{E}Z_1=0, \mathbf{E}Z_2=...=\mathbf{E}Z_5=\alpha$. **3.14.** a) statistika (3.2.16) įgijo reikšmę 0,3985; kadangi P reikšmė yra $\mathbf{P}\{F_{4,48}>0,3985\}=0,8087$, tai atmesti tiesiškumo hipotezę nėra pagrindo; b) $\hat\beta_0=0,8873$, $\hat\beta_1=0,0540$; c) $(\underline{\beta_0};\overline{\beta_0})=(0,7322;1,0425)$; $(\underline{\beta_1};\overline{\beta_1})=(0,0141;0,0938)$. **3.15.** $n>32$. **3.16.** Formulėje (3.2.13) reikia $b^2(x)$ apibrėžti tokiu būdu: $b^2(x)=k/n+(x-\bar x)^2/((n-1)s_x^2)$. **3.17.** Pažymėkime $u=1/y, v=\sin^2 x$. Tada $u=\gamma+\delta v$; $\gamma=1/\alpha, \delta=(\alpha-\beta)/(\alpha\beta)$. **3.18.** $\hat\alpha=8,535$, $\hat\beta=1,2066$, $\hat\sigma^2=0,8115$; statistika (3.2.16) įgijo reikšmę 1,9864; kadangi P reikšmė yra $\mathbf{P}\{F_{3,15}>1,9864\}=0,1594$, tai atmesti tiesiškumo hipotezę nėra pagrindo; tolesnio matavimo prognozės intervalas yra (12,9420; 16,9179), kai $x=12$ ir (20,0178; 24,3213), kai $x=18$. **3.19.** $\hat\mu=0,9343+0,8787x$; hipotezė $H:\beta=0$ atmetama: a. d., kai teisinga hipotezė, turintis Stjudento skirstinį su 106 laisvės laipsniais įgijo reikšmę 21,7608; paaiškina 0,817 sklaidos dalį. **3.20. a)** $\hat\beta_0=0,0144$, $\hat\beta_1=0,0149$, $\hat\sigma^2=0,00004$. Determinacijos koeficientas 0,9697. Sklaidos diagramoje prognozės paklaidos išsidėsčiusios neatsitiktinai, todėl regresijos tiesinis pavidalas nepriimtinas. **b)** $\hat\alpha=0,1758$, $\hat\beta=0,1531$. Esant pasikliovimo lygmeniu 0,95, gauname $(\underline\alpha;\overline\alpha)=(0,1690;0,1827)$, $(\underline\beta;\overline\beta)=(0,1437;0,1626)$. **3.21.** $\hat\alpha_1=13,2255$, $\hat\alpha_2=1,0158$, $\hat\beta_1=6,7978$, $\hat\beta_2=0,0796$. Esant pasikliovimo lygmeniu 0,95, gauname $(\underline{\alpha_1};\overline{\alpha_1})=(13,0724;13,3787)$, $(\underline{\alpha_2};\overline{\alpha_2})=(0,9915;1,0402)$, $(\underline{\beta_1};\overline{\beta_1})=(6,6336;6,9620)$, $(\underline{\beta_2};\overline{\beta_2})=(0,0767;0,0825)$. **3.22. a)** Statistika (3.2.16) įgijo reikšmę 6,29; tiesiškumo hipotezė atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu; **b)** Statistika (3.3.25) įgijo reikšmę 0,29; kadangi P reikšmė yra 0,8295, hipotezę apie antro laipsnio polinominės regresijos modelio tinkamumą atmesti nėra pagrindo. $\hat\mu=13,62+54,05x-4,72x^2$. Determinacijos koeficientas yra 0,8011. **3.23.** $\mathbf{E}\hat\beta_0-\beta_0=4\beta_2$, $\mathbf{E}\hat\beta_1-\beta_1=7\beta_3$. **3.24.** Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai $(SS_{EH}-SS_E)/SS_E>F_\alpha(1\,1)$ čia $SS_E=(Y_1-Y_2+Y_3)^2/3$; $SS_{EH}=(Y_1-3\hat\theta)^2+(Y_2-2\hat\theta)^2+(Y_3+\hat\theta)^2$; $\hat\theta=(3Y_1+2Y_2-Y_3)/14$. **3.25.** Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai $F=[(Y_1-Y_3)^2/2+(Y_2-Y_4)^2]/(\sum_i Y_i-2\pi)^2>F_\alpha(2,1)$. **3.27.** Pažymėkime $\boldsymbol{Y}=(Y_1,...,Y_n)^T$ ir $\boldsymbol\beta=(\beta_0,\beta_1,\beta_2)^T$. Tada $\hat{\boldsymbol\beta}=(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$; čia matrica $\boldsymbol{A}^T$ turi 3 eilutes ir n stulpelių; pirmoji eilutė yra $(1,1,...,1)$; antroji $(\cos(2\pi k_1/n),\cos(4\pi k_1/n),...,\cos(2n\pi k_1/n))$;

trečioji $(\sin(2\pi k_1/n), \sin(4\pi k_1/n), ..., \sin(2n\pi k_1/n))$. **3.29.** **a)** $\hat{\beta}_0 = 84,55$, $\hat{\beta}_1 = 1,83$, $\hat{\beta}_2 = 2,68$; **b)** Gauname $F_1 = 5,89$ ir $F_2 = 8,62$; regresijos koeficientų lygybės nuliui hipotezės atmetamos kriterijais su gana aukštu reikšmingumo lygmeniu; **c)** $(99,02; 105,36)$. **3.30.** Pažingsninės regresijos metodu gauname, kad prognozuojant X_4 reikia naudotis visais kintamaisiais $(X_1, X_2, X_3)^T$; regresijos įvertis $\hat{X}_4 = 8,29 + 0,60X_1 - 0,14X_2 + 0,52X_3$; $r_{X_4\,(X_1,\,X_2,\,X_3)} = 0,8770$. **3.31.** $\psi(x) = 2 - 3x^2$. **3.32.** Atlikime keitimą $Z_i = X_i - 1952$. Tada Z_i įgyja reikšmes -2, -1, 0, 1, 2. Nagrinėkime regresijas pavidalo $Y_i = \beta_0\varphi_0(Z_i) + \beta_1\varphi_1(Z_i) + ... + \beta_m\varphi_m(Z_i)$, kai $\varphi_j(z)$ yra ortogonalūs polinomai: $\varphi_0(z) \equiv 1; \varphi_1(z) = z; \varphi_2(z) = z^2 - 2; \varphi_3(z) = 5z^3 - 17z$. Parametrų įvertiniai $\hat{\beta}_j = \sum_i Y_i\varphi_j(Z_i)/\sum \varphi_j^2(Z_i)$, $\boldsymbol{V}\hat{\beta}_j = \sigma^2/\sum \varphi_j^2(Z_i)$, $j = 0,\ 1,\ 2,\ 3$. Liekamosios kvadratų sumos $SS_E^{(m)} = \sum_i Y_i^2 - \hat{\beta}_0^2 \sum \varphi_0^2(Z_i) - ... - \hat{\beta}_m^2 \sum \varphi_m^2(Z_i)$, $m = 0,\ 1,\ 2,\ 3$. Pagal turimus stebinius gauname $\hat{\beta}_0 = 13,02$, $\hat{\beta}_1 = 1,65$, $\hat{\beta}_2 = -0,0643$, $\hat{\beta}_3 = -0,0333$. Liekamosios kvadratų sumos $SS_E^{(0)} = 27,688$, $SS_E^{(1)} = 0,463$, $SS_E^{(2)} = 0,405$, $SS_E^{(3)} = 0,005$. Matome, kad imant trečiojo laipsnio polinomą liekamoji kvadratų suma įgyja mažą reikšmę. Hipotezė $H : \beta_3 = 0$ atmetama: statistika $F = \hat{\beta}_3^2 \sum_i \varphi_3^2(Z_i)/s^2$, kuri esant teisingai hipotezei turi Fišerio skirstinį su 1 ir 2 laisvės laipsniais, įgijo reikšmę 155,5556, $pv = \mathbf{P}\{F_{1,2} > 155,5556\} = 0,0064$.

4 skyrius

Kovariacinė analizė

4.1. Kovariacinės analizės modeliai

Dispersinė analizė naudojama, kai norima ištirti įvairių kokybinių faktorių (nominaliųjų kovariančių) įtaką priklausomam kintamajam. Pagrindinės hipotezės, tikrinamos dispersine analize - faktorių įtakos ir jų sąveikos nebuvimo hipotezės. Regresinėje analizėje daugiau pabrėžiamas priklausomo kintamojo prognozavimas pagal kovariančių (tiek tolydžiųjų, tiek ir nominaliųjų) reikšmes.

Kai kintamojo Y skirstinys priklauso ne tiktai nuo kokybinių faktorių, bet ir nuo kiekybinių faktorių (tolydžiųjų kovariančių), bet norima ištirti tiktai kokybinių faktorių įtaką, būtina naudoti regresinį modelį tiek su nominaliomis, tiek ir su tolydžiomis kovariantėmis, nes, naudojant dispersinės analizės metodus, tolydžiųjų kovariančių egzistavimas gali iškreipti analizės rezultatus. Pavyzdžiui, tiriant gyvūnų svorio prieaugio priklausomybę nuo maitinimo raciono, prieaugis gali priklausyti nuo tokių kovariančių, kaip antai pradinis gyvūno svoris, jo amžius ir kt. Tiriant padangų dilimo greičio priklausomybę nuo jų markės, protektoriaus dilimo greitis gali priklausyti nuo padangų apkrovos, kelio dangos, meteorologinių sąlygų, vairuotojo stažo ir kt.

Jeigu, be tiriamų kokybinių faktorių ir priklausomo kintamojo Y matavimų, galima gauti ir „trukdančių“ kiekybinių kovariančių matavimus, tai, atliekant statistinę analizę, pastarųjų įtaką reikia eliminuoti. Tokia duomenų analizė vadinama *kovariacine analize*. Iš esmės tai (kaip ir dispersinė analizė) tam tikras regresinės analizės atvejis, turintis savo specifiką. Detaliau apie kovariacinę analizę žr.[12], [14].

Tarkime, kad tiriama vieno faktoriaus A įtaka kintamajam Y ir, kai faktoriaus A lygmuo yra A_i, gauti kintamojo Y ir kovariantės X matavimai

$$(Y_{i1}, x_{i1}), \dots (Y_{iJ_i}, x_{iJ_i}), \quad i = 1, ..., I,$$

tada naudojamas *vienfaktorės vieno kintamojo kovariacinės analizės modelis:*

$$Y_{ij} = \mu + \alpha_i + \gamma x_{ij} + e_{ij}, \quad \sum_i \alpha_i = 0, \tag{4.1.1}$$

čia $e_{ij} \sim N(0, \sigma^2)$ yra vienodai pasiskirstę nepriklausomi normalieji atsitiktiniai dydžiai.

Jeigu kovariantė yra daugiamatė: $\mathbf{X} = (x^{(1)}, ..., x^{(k)})^T$, tai gaunami matavimai

$$(Y_{i1}, x_{i1}^{(1)}, \dots x_{i1}^{(k)}), \dots (Y_{iJ_i}, x_{iJ_i}^{(1)}, \dots x_{iJ_i}^{(k)}), \quad i = 1, ..., I,$$

ir naudojamas *vienfaktorės keleto kintamųjų kovariacinės analizės modelis:*

$$Y_{ij} = \mu + \alpha_i + \gamma_1 x_{ij}^{(1)} + ... + \gamma_k x_{ij}^{(k)} + e_{ij}, \quad \sum_i \alpha_i = 0. \tag{4.1.2}$$

Jeigu dvifaktorėje dispersinėje analizėje su vienu stebėjimu langelyje tarsime, kad faktorių sąveikos nėra ir kartu su Y stebima kovariantė x, tai gautume *adityvųjį dvifaktorės kovariacinės analizės modelį:*

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma x_{ij} + e_{ij}, \quad \sum_i \alpha_i = \sum_j \beta_j = 0. \tag{4.1.3}$$

4.1.1 pastaba. Kaip ir regresinėje analizėje, kovariančių $x^{(1)}, ..., x^{(k)}$ reikšmės dažniausiai būna iš anksto parinktos ir fiksuotos. Arba tai gali būta atsitiktinio vektoriaus $(X^{(1)}, ..., X^{(k)})^T$ realizacijos. Pastaruoju atveju analizė sąlyginė, tariant, kad kovariacijų vektoriaus realizacijos fiksuotos.

4.1.2 pastaba. Jeigu esant fiksuotoms nagrinėjamų faktorių lygmenų kombinacijoms stebimos nepriklausomos atsitiktinio vektoriaus $(Y, X^{(1)}, ..., X^{(k)})^T$ realizacijos, tai gautiems duomenims analizuoti galima taikyti daugiamatę dispersinę analizę, kuri aptariama IV knygos dalyje.

Bendru atveju kovariacinėje analizėje stebėjimų vektoriaus $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ vidurkis užrašomas dviejų dėmenų suma

$$\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A\beta} + \boldsymbol{C\gamma}. \tag{4.1.4}$$

Pirmasis dėmuo apibūdina kažkokį dispersinės analizės modelį, antrasis aprašo tiesinę Y regresiją kintamųjų $x^{(1)}, ..., x^{(k)}$ atžvilgiu. Matrica $\boldsymbol{A}$ yra dispersinės analizės plano matrica, kurios elementai yra 1 arba 0. Matricos $\boldsymbol{C}$ i-asis stulpelis susideda iš kovariantės $\boldsymbol{x}^{(i)}$ matavimų.

Tarę, kad matrca $\boldsymbol{C}$ fiksuota, ir naudodamiesi (4.1.4) gauname *bendrąjį kovariacinės analizės modelį:*

$$\boldsymbol{Y} = \boldsymbol{B\delta} + \boldsymbol{e}; \tag{4.1.5}$$

čia $\boldsymbol{Y}$ yra priklausomo kintamojo n matavimų vektorius, $\boldsymbol{B} = (\boldsymbol{A}\vdots\boldsymbol{C})$, $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T = (\beta_1, ..., \beta_m, \gamma_1, ..., \gamma_k)^T$, $\boldsymbol{A}$ – $n \times m$ matrica, $m = Rang(\boldsymbol{A})$, $\boldsymbol{C}$ – $n \times k$ matrica, $k = Rang(\boldsymbol{C})$; $\boldsymbol{e}$ yra n vienodai pasiskirsčiusių, nepriklausomų normaliųjų $N(0, \sigma^2)$ atsitiktinių dydžių vektorius.

Šis modelis yra tiesinis, todėl jam analizuoti pritaikoma bendra 1 skyriaus metodika.

Atsitiktinio vektoriaus $\boldsymbol{Y}$ koordinatės turi pavidalą

$$Y_i = a_{i1}\beta_1 + ... + a_{im}\beta_m + \gamma_1 x_{1i} + ... + \gamma_k x_{ki} + e_i, \quad i = 1, ..., n; \tag{4.1.6}$$

čia a_{ij}, x_{ij} – žinomos konstantos, $\beta_1, ..., \beta_m, \gamma_1, ..., \gamma_k$ – nežinomi parametrai; e_i – vienodai pasiskirstę nepriklausomi normalieji a. d. $e_i \sim N(0, \sigma^2)$.

Visų pirma detaliau aptarsime vienfaktorės vieno kintamojo kovariacinės analizės modelį, o paskui pereisime prie bendro atvejo.

4.2. Vienfaktorė vieno kintamojo kovariacinė analizė

4.2.1. Mažiausiųjų kvadratų įvertiniai

Nagrinėsime modelį (4.1.1), kuris užrašomas ir (4.1.5) pavidalu. Matrica $\boldsymbol{A}$ (žr. 2.1 skyrelį) turi $n = J_1 + \cdots + J_I$ eilučių ir I stulpelių. Pirmosios J_1 eilutės turi pavidalą $(1,\ 0, ..., 0)$, paskui J_2 eilučių turi pavidalą $(0,\ 1, ..., 0)$, pagaliau paskutinės J_I eilutės turi pavidalą $(0,\ 0, ..., 1)$. Plano matrica $\boldsymbol{B}$ modelyje (4.1.5) gaunama iš plano matricos $\boldsymbol{A}$ papildžius ją stulpeliu $\boldsymbol{C} = (x_{11}, ..., x_{1J_1}, ..., x_{I1}, ..., x_{IJ_I})^T$. Taigi $Rang(\boldsymbol{B}) = I + 1$. Parametrų μ, α_i ir γ MK įvertiniai gaunami minimizuojant kvadratinę formą

$$SS(\boldsymbol{\delta}) = \sum_i \sum_j (Y_{ij} - \mu - \alpha_i - \gamma x_{ij})^2. \tag{4.2.1}$$

Diferencijuodami pagal parametrus, prilyginę išvestines 0 ir pasinaudoję lygybe $\sum_i \alpha_i = 0$, gauname normaliųjų lygčių sistemą

$$\begin{cases} \hat{\mu} + \hat{\gamma}\bar{x}_{..} = \bar{Y}_{..}, \\ \hat{\mu} + \hat{\alpha}_i + \hat{\gamma}\bar{x}_{i.} = \bar{Y}_{i.}, \quad i = 1, ..., I, \\ n\hat{\mu}\bar{x}_{..} + \sum_i \bar{x}_{i.}\hat{\alpha}_i + \hat{\gamma}\sum_i \sum_j x_{ij}^2 = \sum_i \sum_j Y_{ij} x_{ij}. \end{cases}$$

Iš pirmosios lygties išreiškę $\hat{\mu} = \bar{Y}_{..} - \hat{\gamma}\bar{x}_{..}$ ir įstatę į antrąją, paskui iš pastarosios išreiškę $\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..} - \hat{\gamma}(\bar{x}_{i.} - \bar{x}_{..})$ ir įstatę į paskutiniąją, gauname parametro γ mažiausiųjų kvadratų įvertinį

$$\hat{\gamma} = \frac{\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})(x_{ij} - \bar{x}_{..})}{\sum_i (x_{ij} - \bar{x}_{..})^2}. \tag{4.2.2}$$

Liekamoji kvadratinė forma

$$SS_E = \sum_i \sum_j (Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\gamma}x_{ij}))^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.} - \hat{\gamma}(x_{ij} - \bar{x}_{i.}))^2$$

$$= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 - \hat{\gamma} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})(x_{ij} - \bar{x}_{i.}) \sim \sigma^2 \chi^2_{n-I-1}. \tag{4.2.3}$$

Toks SS_E skirstinio pavidalas gaunamas iš 1.3.1 teoremos, nes $Rang(\boldsymbol{B}) = I+1$.

Pažymėkime

$$G_{xx} = \sum_i J_i(\bar{x}_{i.}-\bar{x}_{..})^2, \quad G_{xy} = \sum_i J_i(\bar{x}_{i.}-\bar{x}_{..})(\bar{Y}_{i.}-\bar{Y}_{..}), \quad G_{yy} = \sum_i J_i(\bar{Y}_{i.}-\bar{Y}_{..})^2,$$

$$R_{xx} = \sum_i \sum_j (x_{ij}-\bar{x}_{i.})^2, \; R_{xy} = \sum_i \sum_j (x_{ij}-\bar{x}_{i.})(Y_{ij}-\bar{Y}_{i.}), \; R_{yy} = \sum_i \sum_j (Y_{ij}-\bar{Y}_{i.})^2,$$

$$T_{xx} = \sum_i \sum_j (x_{ij}-\bar{x}_{..})^2, \; T_{xy} = \sum_i \sum_j (x_{ij}-\bar{x}_{..})(Y_{ij}-\bar{Y}_{..}), \; T_{yy} = \sum_i \sum_j (Y_{ij}-\bar{Y}_{..})^2.$$

Gausime vadinamąją kovariacinės analizės lentelę.

4.2.1 lentelė. Kovariacinės analizės lentelė

Šaltinis	xx	xy	yy
Tarp grupių	G_{xx}	G_{xy}	G_{yy}
Grupių viduje	R_{xx}	R_{xy}	R_{yy}
Visas	T_{xx}	T_{xy}	T_{yy}

Reikia pažymėti, kad paskutinio stulpelio nariai G_{yy}, R_{yy}, T_{yy} sutampa su vienfaktorėje dispersinėje analizėje naudotomis kvadratų sumomis (žr. 2.1.2 lentelę) SS_A, SS_E ir SS_T, gautomis neatsižvelgiant į trukdančiąją kovariantę.

Jei vienfaktorės dispersinės analizės atveju liekamoji kvadratinė forma SS_E sutapo su R_{yy}, tai iš (4.2.3) išplaukia, kad kovariacinės analizės atveju liekamoji kvadratinė forma SS_E išreiškiama lentelės antrosios eilutės nariais:

$$SS_E = R_{yy} - \hat{\gamma}R_{xy} = R_{yy} - R_{xy}^2/R_{xx}, \quad \hat{\gamma} = \frac{R_{xy}}{R_{xx}}. \tag{4.2.4}$$

Matome, kad liekamoji kvadratų suma R_{yy}, kuri gaunama neatsižvelgiant į kovariantę x, sumažėja dydžiu

$$\hat{\gamma}R_{xy} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})(x_{ij} - \bar{x}_{..}),$$

t. y. sumažėjimas proporcingas porų (x_{ij}, Y_{ij}) empirinei kovariacijai. Nuo to ir kilęs kovariacinės analizės pavadinimas.

4.2.2. Faktoriaus įtakos nebuvimo hipotezės tikrinimas

Tikrinant hipotezę

$$H_A : \alpha_1 = ... = \alpha_I = 0,$$

kad nėra faktoriaus A įtakos, reikia rasti sąlyginį kvadratinės formos minimumą

$$SS_{EH_A} = \min_{\mu,\gamma} \sum_i \sum_j (Y_{ij} - \mu - \gamma x_{ij})^2.$$

Tokią kvadratinę formą minimizavome nagrinėdami vieno kintamojo tiesinę regresiją (žr. 3.2.1 teoremą). Gauname, kad šis minimumas išreiškiamas 4.2.1 lentelės nariais:

$$SS_{EH_A} = \sum_i \sum_j (Y_{ij} - \tilde{\mu} - \tilde{\gamma} X_{ij})^2 = T_{yy} - \tilde{\gamma} T_{xy},$$

$$\tilde{\mu} = \bar{Y}_{\cdot\cdot} - \tilde{\gamma} \bar{X}_{\cdot\cdot}, \quad \tilde{\gamma} = \frac{T_{xy}}{T_{xx}} = \frac{R_{xy} + G_{xy}}{R_{xx} + G_{xx}}.$$

Kvadratų suma, apibūdinanti faktoriaus A įtaką

$$\begin{aligned} SS_A = SS_{EH_A} - SS_E &= T_{yy} - \frac{T_{xy}^2}{T_{xx}} - (R_{yy} - \frac{R_{xy}^2}{R_{xx}}) \\ &= G_{yy} - \frac{T_{xy}^2}{T_{xx}} + \frac{R_{xy}^2}{R_{xx}} \end{aligned} \tag{4.2.5}$$

Hipotezė H_A turi pavidalą (3.3.20) su $k = I - 1$ ir $m = I$. Iš 3.3.7 skyrelio išplaukia, kad hipotezė H_A atmetama, kai

$$F_A = \frac{SS_A(n - I - 1)}{(I - 1)SS_E} = \frac{MS_A}{MS_E} > F_\alpha(I - 1, n - I - 1). \tag{4.2.6}$$

4.2.3. Trukdančių parametrų įtakos nebuvimo hipotezės tikrinimas

Taikant kovariacinę analizę prasminga patikrinti hipotezę $H_\gamma : \gamma = 0$. Jei ši hipotezė atmetama, reiškia, kad, tiriant faktoriaus įtaką, atsižvelgti į kovariantę x prasminga.

Kai teisinga hipotezė H_γ, modelis (4.1.1) yra tiesiog vienfaktorės dispersinės analizės modelis. Todėl

$$SS_{EH_\gamma} = R_{yy}, \quad SS_\gamma = R_{yy} - SS_E = \frac{R_{xy}^2}{R_{xx}}. \tag{4.2.7}$$

Hipotezė H_γ turi pavidalą (3.3.21) (tai atskiras (3.3.20) atvejis, kai $k = 1$), todėl pagal (4.3.27) hipotezė H_γ atmetama, kai

$$F_\gamma = \frac{SS_\gamma}{MS_E} = \frac{R_{xy}^2(n - I - 1)}{R_{xx}R_{yy} - R_{xy}^2} > F_\alpha(1, n - I - 1). \tag{4.2.8}$$

4.2.4. Regresijos tiesių lygiagretumo hipotezės tikrinimas

Modelyje (4.1.1) priimama prielaida, kad regresijos tiesės krypties koeficientas γ yra tas pats esant bet kuriam faktoriaus A lygmeniui, t. y. regresijos tiesės prie įvairių faktoriaus A lygmenų yra lygiagrečios.

Nagrinėkime bendresnį vienfaktorės vieno kintamojo kovariacinės analizės modelį

$$Y_{ij} = \mu + \alpha_i + \gamma_i x_{ij} + e_{ij}, \tag{4.2.9}$$

kuriame regresijos parametrai γ_i gali būti skirtingi.

MK metodu randame įvertinius

$$\hat{\gamma}_i = \frac{\sum_j (x_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{Y}_{i.})}{(x_{ij} - \bar{x}_{i.})^2}, \quad i = 1, ..., I,$$

ir liekamąją kvadratinę formą šiame bendresniame modelyje

$$\tilde{SS}_E = R_{yy} - \sum_i \hat{\gamma}_i \sum_j (x_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{Y}_{i.}). \tag{4.2.10}$$

Tariant, kad hipotezė $\tilde{H}_\gamma : \gamma_1 = ... = \gamma_I$ yra teisinga, sąlyginis kvadratinės formos minimumas sutampa su surastuoju SS_E iš (4.2.3).

Hipotezė $\tilde{H}_\gamma$ turi pavidalą, nagrinėtą 1.3.2 poskyrio 1.3.2 pavyzdyje (šiuo atveju $k = 2I$, $m = I$). Iš šio pavyzdžio išeina, kad hipotezė $\tilde{H}_\gamma$ atmetama, kai

$$\tilde{F}_\gamma = \frac{(SS_E - \tilde{SS}_E)(n - 2I)}{(I - 1)\tilde{SS}_E} > F_\alpha(I - 1, n - 2I). \tag{4.2.11}$$

4.3. Bendrasis kovariacinės analizės atvejis

Nagrinėkime bendrąjį kovariacinės analizės modelį (4.1.5):

$$\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A\beta} + \boldsymbol{C\gamma}. \tag{4.3.1}$$

4.3.1. Parametrų įvertiniai ir liekamoji kvadratinė forma

Normaliųjų lygčių sistema nežinomiems parametrams pagal MK metodą vertinti turi tokį pavidalą (žr. (1.2.2)):

$$\begin{cases} \boldsymbol{A}^T\boldsymbol{A\beta} + \boldsymbol{A}^T\boldsymbol{C\gamma} = \boldsymbol{A}^T\boldsymbol{Y}, \\ \boldsymbol{C}^T\boldsymbol{A\beta} + \boldsymbol{C}^T\boldsymbol{C\gamma} = \boldsymbol{C}^T\boldsymbol{Y}. \end{cases} \tag{4.3.2}$$

Jei $rang(\boldsymbol{B}) = m + k$, $\boldsymbol{B} = (\boldsymbol{A}\vdots\boldsymbol{C})$, tai jos sprendinys

$$\hat{\boldsymbol{\delta}} = (\hat{\beta}_1, ..., \hat{\beta}_m, \hat{\gamma}_1, ..., \hat{\gamma}_k)^T = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{Y}.$$

Parodysime, kad vietoje $m + k$ lygčių sistemos su $m + k$ nežinomųjų (4.3.2) užtenka spręsti $k + 1$ lygčių sistemą, kiekvienoje iš kurių yra m nežinomųjų, ir vieną k lygčių sistemą su k nežinomųjų.

Pažymėkime $\hat{\boldsymbol{\beta}}_0$ MK įvertinį dispersinės analizės modelyje $\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{e}$. Pagal (1.2.2) jis tenkina lygčių sistemą

$$\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}}_0 = \boldsymbol{A}^T\boldsymbol{Y}. \tag{4.3.3}$$

Tarkime, kad $\hat{\boldsymbol{\beta}}_i$, $\quad i = 1, ..., k$, yra lygčių sistemos

$$\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}}_i = \boldsymbol{A}^T\boldsymbol{C}_i \tag{4.3.4}$$

sprendinys; čia $\boldsymbol{C}_i$ yra matricos $\boldsymbol{C}$ i-asis stulpelis. Jei kovariantės būtų atsitiktinės, tai $\boldsymbol{\beta}_i$ būtų MK įvertinys dispersinės analizės modelyje $\boldsymbol{C}_i = \boldsymbol{A}\boldsymbol{\beta}_i + \boldsymbol{e}_i$. Pažymėkime

$$R_{yy} = (\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_0)^T(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_0) = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}_0^T\boldsymbol{A}^T\boldsymbol{Y},$$

$$R_{x_ix_i} = (\boldsymbol{C}_i - \boldsymbol{A}\hat{\boldsymbol{\beta}}_i)^T(\boldsymbol{C}_i - \boldsymbol{A}\hat{\boldsymbol{\beta}}_i) = \boldsymbol{C}_i^T\boldsymbol{C}_i - \hat{\boldsymbol{\beta}}_i^T\boldsymbol{A}^T\boldsymbol{C}_i, \quad i = 1, ..., k, \quad (4.3.5)$$

liekamąsias kvadratines formas.

Įveskime analogiškas mišrias sumas

$$R_{yx_i} = (\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_0)^T(\boldsymbol{C}_i - \boldsymbol{A}\hat{\boldsymbol{\beta}}_i) = \boldsymbol{Y}^T\boldsymbol{C}_i - \hat{\boldsymbol{\beta}}_0^T\boldsymbol{A}^T\boldsymbol{C}_i = \boldsymbol{Y}^T\boldsymbol{C}_i - \hat{\boldsymbol{\beta}}_i\boldsymbol{A}^T\boldsymbol{Y},$$

$$R_{x_ix_j} = (\boldsymbol{C}_i - \boldsymbol{A}\hat{\boldsymbol{\beta}}_i)^T(\boldsymbol{C}_j - \boldsymbol{A}\hat{\boldsymbol{\beta}}_j) = \boldsymbol{C}_i^T\boldsymbol{C}_j - \hat{\boldsymbol{\beta}}_i^T\boldsymbol{A}^T\boldsymbol{C}_j, \quad i, j = 1, ..., k. \quad (4.3.6)$$

Pažymėkime $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, ..., \hat{\gamma}_k)^T$ lygčių sistemos

$$R_{x_1x_i}\hat{\gamma}_1 + ... + R_{x_kx_i}\hat{\gamma}_k = R_{yx_i}, \quad i = 1, ..., k, \quad (4.3.7)$$

sprendinį ir tegu

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - \hat{\gamma}_1\hat{\boldsymbol{\beta}}_1 - ... - \hat{\gamma}_k\hat{\boldsymbol{\beta}}_k. \quad (4.3.8)$$

4.3.1 teorema. *(Rao). Įvertinys $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ yra lygčių sistemos (4.3.2) sprendinys, o liekamoji kvadratinė forma SS_E turi pavidalą*

$$SS_E = (\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{C}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{C}\hat{\boldsymbol{\gamma}})$$

$$= R_{yy} - \hat{\gamma}_1R_{yx_1} - ... - \hat{\gamma}_kR_{yx_k} \sim \sigma^2\chi^2_{n-m-k}. \quad (4.3.9)$$

Įrodymas. Įstatę $\hat{\boldsymbol{\beta}}$ ir $\hat{\boldsymbol{\gamma}}$ į pirmosios (4.3.2) lygties kairiąją pusę, pasinaudoję (4.3.3), (4.3.4), (4.3.8), ir lygybe $\hat{\gamma}_1\boldsymbol{C}_1 + ... + \hat{\gamma}_k\boldsymbol{C}_k = \boldsymbol{C}\hat{\boldsymbol{\gamma}}$, gauname

$$\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\beta}} + \boldsymbol{A}^T\boldsymbol{C}\hat{\boldsymbol{\gamma}} = \boldsymbol{A}^T\boldsymbol{A}(\hat{\boldsymbol{\beta}}_0 - \hat{\gamma}_1\hat{\boldsymbol{\beta}}_1 - ... - \hat{\gamma}_k\hat{\boldsymbol{\beta}}_k) + \boldsymbol{A}^T\boldsymbol{C}\hat{\boldsymbol{\gamma}}$$

$$= \boldsymbol{A}^T\boldsymbol{Y} - \hat{\gamma}_1\boldsymbol{A}^T\boldsymbol{C}_1 - ... - \hat{\gamma}_k\boldsymbol{A}^T\boldsymbol{C}_k) + \boldsymbol{A}^T\boldsymbol{C}\hat{\boldsymbol{\gamma}} = \boldsymbol{A}^T\boldsymbol{Y}.$$

Taigi įvertinys $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ tenkina pirmąją (4.3.1) lygtį.

Iš paskutiniųjų lygybių gauname, kad $\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\gamma}})$, todėl

$$\boldsymbol{C}^T(\boldsymbol{A}\hat{\boldsymbol{\beta}} + \boldsymbol{C}\hat{\boldsymbol{\gamma}} - \boldsymbol{Y}) = \boldsymbol{C}^T(\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T - \mathbf{I})(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\gamma}}) = \boldsymbol{0},$$

o tai yra matricine forma užrašyta (4.3.7) lygčių sistema. Taigi $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ tenkina ir antrąją (4.3.1) lygtį.

Remiantis tiesinių modelių teorija (žr. 1.2.2 pastabą), liekamąją kvadratų sumą galima užrašyti tokiu pavidalu:

$$SS_E = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{A}^T\boldsymbol{Y} - \hat{\boldsymbol{\gamma}}^T\boldsymbol{C}^T\boldsymbol{Y}.$$

Įstatę $\hat{\boldsymbol{\beta}}$ išraišką (4.3.8) ir pasinaudoję (4.3.5) bei (4.3.6) žymėjimais, gauname

$$SS_E = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}_0^T\boldsymbol{A}^T\boldsymbol{Y} + \hat{\gamma}_1\hat{\boldsymbol{\beta}}_1^T\boldsymbol{A}^T\boldsymbol{Y} + ... + \hat{\gamma}_k\hat{\boldsymbol{\beta}}_k^T\boldsymbol{A}^T\boldsymbol{Y} - \hat{\boldsymbol{\gamma}}^T\boldsymbol{C}^T\boldsymbol{Y}$$

$$= (\boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}_0^T\boldsymbol{A}^T\boldsymbol{Y}) - \hat{\gamma}_1(\boldsymbol{C}_1^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}_1^T\boldsymbol{A}^T\boldsymbol{Y}) - ... - \hat{\gamma}_k(\boldsymbol{C}_k^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}_k^T\boldsymbol{A}^T\boldsymbol{Y})$$

$$= R_{yy} - \hat{\gamma}_1 R_{yx_1} - ... - \hat{\gamma}_k R_{yx_k}.$$

▲

4.3.2. Hipotezių tikrinimas

Patikrinsime hipotezę $H_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = \boldsymbol{0}$, t. y. prielaidą, kad kovariantės $X_1, ..., X_k$ nedaro įtakos priklausomo kintamojo Y skirstiniui. Jei ši hipotezė priimama, tai vietoje kovariacinės analizės modelio galima nagrinėti dispersinės analizės modelį. Sąlyginis kvadratinės formos minimumas

$$SS_{EH_\gamma} = \min_{\boldsymbol{\beta},\boldsymbol{\gamma}=\boldsymbol{0}} (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{C}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{C}\boldsymbol{\gamma})$$

yra tiesiog $R_{yy} = (\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_0)^T(\boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_0)$. Kvadratų suma $SS_{\boldsymbol{\gamma}} = SS_{EH_\gamma} - SS_E$ yra

$$SS_{\boldsymbol{\gamma}} = R_{yy} - SS_E = \hat{\gamma}_1 R_{yx_1} + ... + \hat{\gamma}_k R_{yx_k}.$$

Hipotezė $H_{\boldsymbol{\gamma}}$ turi pavidalą (3.3.20) (šiuo konkrečiu atveju imame $k := k$, $m := m + k$). Hipotezė $H_{\boldsymbol{\gamma}}$ atmetama, kai

$$F_{\boldsymbol{\gamma}} = \frac{SS_{\boldsymbol{\gamma}}(n - m - k)}{kSS_E} = \frac{MS_{\boldsymbol{\gamma}}}{MS_E} > F_\alpha(k, n - m - k). \tag{4.3.10}$$

Analogiškai galima patikrinti prielaidą, kad tiktai dalis kovariančių, pavyzdžiui, $X_{i_1}, ..., X_{i_l}$, neturi įtakos priklausomo kintamojo Y skirstiniui, t. y. patikrinti hipotezę $H_{i_1,...,i_l} : \gamma_{i_1} = ... = \gamma_{i_l} = 0$. Šiuo atveju

$$SS_{EH_{i_1,...,i_l}} = \min_{\boldsymbol{\beta},\boldsymbol{\gamma}:\gamma_{i_1}=...=\gamma_{i_l}=0} (\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{C}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{C}\boldsymbol{\gamma}),$$

$SS_{\gamma_{i_1},...,\gamma_{i_l}} = SS_{EH_{i_1,...,i_l}} - SS_E$ ir hipotezė $H_{i_1,...,i_l}$ atmetama, kai

$$F_{\gamma_{i_1},...,\gamma_{i_l}} = \frac{SS_{\gamma_{i_1},...,\gamma_{i_l}}(n - m - k)}{lSS_E} > F_\alpha(l, n - m - k). \tag{4.3.11}$$

Patikrinsime hipotezę

$$H : \beta_{i_1} = \cdots = \beta_{i_s},$$

kuri dažnai tikrinama dispersinėje analizėje.

Vienfaktorėje kovariacinėje analizėje $\beta_{i_j} := \alpha_j$, $j = 1, \ldots, s = I$. Dvifaktorėje kovariacinėje analizėje gali būti $\beta_{i_j} := \alpha_j$, $j = 1, \ldots, s = I$ arba $\beta_{i_j} := \beta_j$, $j = 1, \ldots, s = J$, arba β_{i_j} gali reikšti sąveiką apibūdinančius parametrus γ_{ij} priklausomai nuo tikrinamos hipotezės.

Remiantis bendra mažiausiųjų kvadratų teorija reikia rasti sąlyginį minimumą

$$SS_{EH} = \min_{\beta_{i_1}=\cdots=\beta_{i_s}} (\boldsymbol{Y} - \boldsymbol{A\beta} - \boldsymbol{C\gamma})^T (\boldsymbol{Y} - \boldsymbol{A\beta} - \boldsymbol{C\gamma}).$$

Šį sąlyginį minimumą galima rasti taikant 4.3.1 teoremą. Jis sutampa su liekamąja kvadratine forma modelyje, kuriame vektorius $\boldsymbol{\beta}$ pakeistas vektoriumi $\boldsymbol{\theta} = (\theta_1, ..., \theta_{m-s+1})$. Šis vektorius susideda iš $m-s$ likusių koordinačių $\beta_{i_j}, j \neq 1, ..., s$ ir parametro θ_{m-s+1}, įrašyto vietoje koordinačių $\beta_{i_j}, j = 1, ..., s$.

Pažymėkime $\tilde{\theta}_0$ ir $\tilde{\theta}_i$, $i = 1, ..., m-s+1$, lygčių (4.3.3) ir (4.3.4) sprendinius, kai teisinga hipotezė H. Kvadratines formas (4.3.4) ir kovariacijas (4.3.5), kai naudojami įvertiniai $\tilde{\theta}_0$ ir $\tilde{\theta}_{i_j}$, žymėsime $\tilde{R}_{yy}$, $\tilde{R}_{x_i x_i}$, $\tilde{R}_{yx_i}$, $\tilde{R}_{x_i x_j}$. Lygčių sistemos

$$\tilde{R}_{x_1 x_i}\gamma_1 + ... + \gamma_k \tilde{R}_{x_k x_i} = \tilde{R}_{yx_i}, \quad i = 1, ..., k,$$

sprendinį žymėsime $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, ..., \tilde{\gamma}_k)^T$. Tada analogiškai (4.3.9) sąlyginė liekamoji kvadratinė forma yra

$$SS_{EH} = \tilde{R}_{yy} - \tilde{\gamma}_1 \tilde{R}_{yx_1} - ... - \tilde{\gamma}_k \tilde{R}_{yx_k}. \tag{4.3.12}$$

Kvadratinė forma, apibūdinanti hipotezę H, yra

$$SS_H = SS_{EH} - SS_E, \tag{4.3.13}$$

kuri, kai hipotezė H teisinga, turi tokį pat skirstinį kaip a. d. $\sigma^2 \chi^2_s$ ir nepriklauso nuo SS_E.

Hipotezė H yra atmetama, kai

$$F_H = \frac{SS_H(n-m-k)}{s(SS_E)} = \frac{MS_H}{MS_E} > F_\alpha(s, n-m-k). \tag{4.3.14}$$

4.3.1 pavyzdys Tiriant azoto dujų N_2 susidarymą žmogaus organizme, buvo registruojamas iškvepiamo azoto kiekis Y priklausomai nuo maitinimosi režimų (faktorius A), kurie skyrėsi baltymų kiekiu. Kartu buvo registruojamas įkvepiamo azoto kiekis X. Atsitiktinai atrinkta po $J_i = 9$ individus ir jiems buvo taikoma dieta $A_i, i = 1, ..., 4$. [1]

4.3.1 lentelė. Statistiniai duomenys

A_1		A_2		A_3		A_4	
Y	X	Y	X	Y	X	Y	X
4,079	4,158	4,368	4,322	4,169	4,102	4,928	4,829
4,859	4,877	5,668	5,617	5,709	5,582	5,608	5,400
3,540	3,576	3,752	3,720	4,416	4,339	4,940	4,799
5,047	5,078	5,848	5,797	5,666	5,585	5,291	5,167
3,298	3,315	3,802	3,773	4,123	4,049	4,674	4,565
4,679	4,702	4,844	4,800	5,059	4,987	5,038	4,933
2,870	2,901	3,578	3,539	4,403	4,322	4,905	4,762
4,648	4,718	5,393	5,317	4,496	4,383	5,208	5,080
3,847	3,880	4,374	4,343	4,688	4,623	4,806	4,709

Tirdami Y priklausomybę nuo faktoriaus A ir atlikę vienfaktorę dispersinę analizę, gauname statistikos $F_A = MS_A/MS_E$ reikšmę, lygią 3,211. Esant normalumo prielaidai ir teisingai hipotezei H_A, statistika F_A turi Fišerio skirstinį su 3 ir 32 laisvės laipsniais. Hipotezė neatmetama kriterijumi, kurio reikšmingumo lygmuo $\alpha < \mathbf{P}(F_{3;\,32} > 3,211) = 0,0359$.

Atlikime duomenų analizę, remdamiesi kovariacinės analizės metodais, atsižvelgiant į kintamąjį X. Tikrindami regresijos koeficientų lygybės hipotezę $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma$, gauname santykio (4.2.18) reikšmę 3,39. Taigi hipotezė apie jų lygybę kriterijumi su reikšmingumo lygmeniu $\alpha < 0,0316$ neatmetama.

Tikrindami hipotezę apie bendro regresijos koeficiento γ lygybę 0, gauname statistikos (4.2.8) reikšmę yra 25300. Taigi hipotezė apie kintamojo X nereikalingumą prognozuojant Y atmetama labai aukštu reikšmingumo lygmens kriterijumi.

Pagaliau, tikrinant hipotezę H_A eliminavę kintamojo X įtaką, gausime statistikos (4.2.6) reikšmę 60,5. Kai yra normalumo prielaida ir teisinga hipotezė H_A, ši statistika turi Fišerio skirstinį su 3 ir 31 laisvės laipsniu. Taigi, eliminavus kintamojo X įtaką, gaunama priešinga išvada ir hipotezė H_A atmetama.

4.4. Dispersinė ir kovariacinė analizė – atskiri regresinės analizės atvejai

Dispersinė analizė naudojama, kai norima ištirti įvairių kokybinių faktorių (nominaliųjų kovariančių) įtaką priklausomam kintamajam. Pagrindinės dispersinės analizės hipotezės - faktorių ir jų sąveikų įtakos nebuvimo hipotezės. Regresinėje analizėje labiau pabrėžiamas priklausomo kintamojo prognozavimas pagal kovariančių (tiek tolydžių, tiek nominalių) reikšmes. Dispersinės analizės uždaviniai gali būti suformuluoti regresinės analizės terminais.

Skyrelyje 4.3.7 buvo parodyta, kad tiesinės regresijos modelyje

$$Y_i = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_m x_m^{(i)} + e_i \tag{4.4.1}$$

reikšmingumo lygmens α kritinė sritis hipotezei

$$H_{j_1 \ldots j_k} : \beta_{j_1} = \cdots = \beta_{j_k} = 0, \quad 1 \le j_1 \le j_k \le m \tag{4.4.2}$$

tikrinti turi pavidalą

$$F_{j_1 \ldots j_k} = \frac{SS_E^{(m-k)} - SS_E}{k s^2} > F_\alpha(k, n-m-1); \tag{4.4.3}$$

čia $SS_E^{(m-k)}$ yra SS_E modelio be kovariančių $x_{j_1}, \ldots, x_{j_k}$ analogas; $s^2 = SS_E/(n-m-1)$. Be to, parodyta, kad hipotezės $H_{1,\ldots,m}$ atveju

$$SS_E^{(0)} - SS_E = SS_R = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_\cdot)^2, \quad \hat{Y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}^{(i)}. \tag{4.4.4}$$

Priminsime, jei kovariantė (faktorius) yra nominalioji ir įgyja I reikšmių, tai šią kovariantę tiesinės regresijos modelyje atitinka $I-1$ fiktyvių kovariančių vektorius $(z_1, \ldots, z_{I-1})^T$, įgyjantis reikšmes $(0,0,\ldots,0)^T$, $(1,0,\ldots,0)^T$, $\ldots$, $(0,0,\ldots,1)^T$, kai nominalioji kovariantė įgyja atitinkamai pirmąją, antrąją ir t. t., I-ąją reikšmę.

4.4.1. Vienfaktorė dispersinė analizė

Stebėjimus aprašantis matematinis modelis pateiktas 2.1.1 skyrelyje. Nagrinėjama kintamojo Y priklausomybė nuo faktoriaus A, kuris gali būti I skirtingų lygmenų $A_1, ..., A_I$. Su kiekvienu lygmeniu A_i gaunama paprastoji imtis

$$Y_{i1}, ..., Y_{iJ_i}, \quad i = 1, ..., I.$$

Tariama, kad stebėjimus galime užrašyti šitaip

$$Y_{ij} = \mu_i + e_{ij}, \quad j = 1, ..., J_i, \quad i = 1, ..., I, \tag{4.4.5}$$

čia e_{ij} – nepriklausomi vienodai pasiskirstę normalieji a. d. $e_{ij} \sim N(0, \sigma^2)$. Pagrindinis analizės tikslas – tikrinti vidurkių lygybės hipotezę $H_A : \mu_1 = ... = \mu_I$.

Traktuokime šį modelį kaip regresinės analizės modelį, kuriame kintamasis Y prognozuojamas remiantis viena nominaliąja kovariante A, įgyjančia I skirtingų reikšmių $A_1, ..., A_I$. Pagal pirmiau naudotą nominaliosios kovariantės kodavimą gauname tiesinį regresijos modelį:

$$Y_{1j} = \beta_0 + e_{1j}, \quad j = 1, ..., J_1; \quad Y_{ij} = \beta_0 + \beta_i + e_{ij}, \quad j = 1, ..., J_i, \quad i = 2, ..., I. \tag{4.4.6}$$

Palyginę su (4.4.5) matome, kad nauji parametrai susiję su vidurkiais $\mu_1, ..., \mu_I$ lygybėmis

$$\beta_0 = \mu_1, \quad \beta_0 + \beta_{i-1} = \mu_i, \quad i = 2, ..., I.$$

Naujųjų parametrų MK įvertiniai yra

$$\hat{\beta}_0 = \bar{Y}_{1.}, \quad \hat{\beta}_{i-1} = \bar{Y}_{i.} - \bar{Y}_{1.}, \quad i = 2, ..., I, \tag{4.4.7}$$

o liekamoji kvadratų suma

$$SS_E = \sum_{j=1}^{J_1} (Y_{1j} - \hat{\beta}_0)^2 + \sum_{i=2}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \hat{\beta}_0 - \hat{\beta}_{i-1})^2 =$$

$$\sum_{i=2}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2 \sim \sigma^2 \chi^2_{n-I}, \quad n = J_1 + ... + J_I$$

yra ta pati kaip ir dispersinės analizės schemoje.

Vidurkių lygybės hipotezė $H : \mu_1 = ... = \mu_I$ naujų parametrų terminais ekvivalenti hipotezei $H : \beta_1 = ... = \beta_{I-1} = 0$, kuri yra atskiras dalies regresijos koeficientų lygybę 0 hipotezės (3.4.2) atvejis. Liekamoji kvadratų suma, kai H teisinga, yra

$$SS_E^{(0)} = \min_{\beta_0} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \hat{\beta}_0)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{..})^2 = SS_T,$$

o skirtumas

$$SS_E^{(0)} - SS_E = \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 = SS_A$$

sutampa su kvadratų suma SS_A, gauta 2.1 skyrelyje. Taigi kriterijus (4.4.3):

$$F_{1,...,I-1} = \frac{(SS_E^{(0)} - SS_E)}{(I-1)s^2} > F_\alpha(I-1, n-I)$$

sutampa su vienfaktorėje dispersinėje analizėje gautuoju kriterijumi (2.1.10).

Atliekant duomenų analizę matematinės statistikos programų paketais (SAS, SPSS ir kt.), paprastai automatiškai pateikiami regresijos koeficientų įverčiai (4.4.7), taip pat patikrinamos kiekvieno regresijos koeficiento reikšmingumo prognozuojant kintamąjį Y hipotezės $H_i : \beta_i = 0$. Šiuos rezultatus galima traktuoti kaip dalinę kontrastų analizę (žr. 2.1.4 skyrelį), kurioje tikrinamos kontrastų lygybės nuliui hipotezės

$$\varphi_i = \mu_i - \mu_1, \quad i = 2, ..., I.$$

Nesunku regresinės analizės modelį pertvarkyti taip, kad kriterijus (4.4.3) būtų pritaikomas kitokių kontrastų lygybės 0 hipotezėms tikrinti. Pavyzdžiui, turima papildomos informacijos, kad vidurkių lygybės hipotezės atmetimo priežastis gal būt yra ta, kad kintamajam Y skirtingą poveikį turi faktoriaus A lygmenų grupės $A_1, ..., A_m$ ir $A_{m+1}, ..., A_I$, o grupių viduje vidurkiai nesiskiria. Kitaip sakant, norima patikrinti hipotezę $H' : \mu_1 = ... = \mu_m; \mu_{m+1} = ... = \mu_I$. Nagrinėkime regresinės analizės modelį, kuriame naujieji parametrai apibrėžti tokiu būdu:

$$\mu_1 = \beta_0, \quad \mu_i = \beta_0 + \beta_{i-1}, \quad i = 2, ..., m;$$

$$\mu_{m+1} = \gamma_0, \quad \mu_{m+1+i} = \gamma_0 + \gamma_{i-1}, \quad i = 2, ..., I-m-1.$$

Tada hipotezė H' yra ekvivalenti tokiai regresijos koeficientų lygybės nuliui hipotezei:

$$H' : \beta_1 = ... = \beta_{m-1} = 0; \quad \gamma_1 = ... = \gamma_{I-m-1} = 0$$

ir gali būti patikrinta remiantis (4.4.3) kriterijumi.

4.4.2. Vienfaktorė kovariacinė analizė

Nagrinėkime vienfaktorės kovariacinės analizės modelį

$$Y_{ij} = \mu_i + \gamma_1 x_{ij}^{(1)} + ... + \gamma_k x_{ij}^{(k)} + e_{ij}, \quad j = 1, ..., J_i, \quad i = 1, ..., I, \qquad (4.4.8)$$

kuriame yra viena nominalioji kovariantė (faktorius A), įgyjanti I skirtingų reikšmių $A_1, ..., A_I$ ir k kiekybinių (trukdančiųjų) kovariančių.

Analogiškai kaip pirmesniame skyrelyje, sukodavę nominaliąją kovariantę kodavimą;

$$\mu_1 = \beta_0, \quad \mu_i = \beta_0 + \beta_{i-1}, \quad i = 2, ..., I,$$

gausime tiesinės regresijos modelį su parametrais $\beta_0, \beta_1, ..., \beta_{I-1}, \gamma_1, ..., \gamma_k$.

Kaip ir atliekant vienfaktorę dispersinę analizę, pagrindinė vienfaktorės kovariacinės analizės hipotezė yra

$$H : \mu_1 = \cdots = \mu_I,$$

kuri reiškia, kad tiriamas faktorius (nominalioji kovariantė) neturi įtakos priklausomojo kintamojo Y skirstiniui. Be to, kuriant kriterijų pageidautina eliminuoti kitų kovariančių įtaką.

Regresinės analizės schemoje hipotezė H ekvivalenti tvirtinimui, kad dalis regresijos koeficientų lygi nuliui: $H : \beta_1 = ... = \beta_{I-1} = 0$. Šią hipotezę tikriname pagal (4.4.3) kriterijų

$$F_{1,...,I-1} = \frac{SS_E^{(k)} - SS_E}{ks^2} > F_\alpha(I-1, n-I-k), \tag{4.4.9}$$

čia $s^2 = MS_E = SS_E/(n-k-I)$, SS_E – besąlyginė liekamoji kvadratų suma, o $SS_E^{(k)}$ – liekamoji kvadratų suma, apskaičiuota tarus, kad $\beta_1 = ... = \beta_{I-1} = 0$.

Norint įsitikinti, ar trukdančios kovariantės yra reikšmingos, reikia patikrinti hipotezę $H_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T = \mathbf{0}$. Jei ši hipotezė priimama, tai vietoje kovariacinės analizės modelio galima nagrinėti dispersinės analizės modelį.

Kadangi vėl nagrinėjame regresinį modelį (4.4.2), kuriame $m := I - 1 + k$, tiktai $k := k$, tai kritinė sritis (4.4.3) hipotezei $H_{\boldsymbol{\gamma}}$ tikrinti užrašoma šitaip

$$F_{\boldsymbol{\gamma}} = \frac{SS_E^{(I-1)} - SS_E}{ks^2} > F_\alpha(k, n-I-k). \tag{4.4.10}$$

Analogiškai galima patikrinti prielaidą, kad tiktai dalis kovariančių, pavyzdžiui $x_{i_1}, ..., x_{i_l}$, neturi įtakos priklausomo kintamojo Y skirstiniui, t. y. patikrinti hipotezę $H_{\gamma_{i_1},...,\gamma_{i_l}} : \gamma_{i_1} = ... = \gamma_{i_l} = 0$. Kadangi vėl nagrinėjame regresinį modelį (4.4.2), kuriame $m := I - 1 + k$, tiktai $k := l$, tai kritinė sritis (4.4.3) hipotezei $H_{\boldsymbol{\gamma}}$ tikrinti užrašoma tokiu pavidalu

$$F_{\gamma_{i_1},...,\gamma_{i_l}} = \frac{SS_E^{(I-1+k-l)} - SS_E}{ks^2} > F_\alpha(l, n-I-k). \tag{4.4.11}$$

Atveju $k = 1$ gaunamos hipotezių tikrinimo kritinės sritys išreikštiniu pavidalu pateiktos 4.1.2 skyrelyje.

4.4.3. Dvifaktorė dispersinė analizė

Imkime dvifaktorės dispersinės analizės modelį iš 2.2.1 skyrelio. Jame nagrinėjama kintamojo Y priklausomybė nuo faktoriaus A, kurio lygmenys $A_1, ..., A_I$, ir nuo faktoriaus B, jo lygmenys $B_1, ..., B_J$. Kiekvienam faktorių lygmenų rinkiniui (A_i, B_j) gaunama po $K > 1$ nepriklausomų kintamojo Y matavimų. Taigi stebėjimo rezultatai yra Y_{ijk}, $i = 1, ..., I$, $j = 1, ..., J$, $k = 1, ..., K$. Stebėjimus aprašo tiesinis modelis

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, ..., I, \quad j = 1, ..., J, \quad k = 1, ..., K. \tag{4.4.12}$$

čia μ_{ij} – nežinomi parametrai, o e_{ijk} – normalieji n. a. d. $e_{ijk} \sim N(0, \sigma^2)$.

Vidurkis μ_{ij} išskaidomas į komponentes

$$\mu_{ij} = \bar{\mu}_{..} + (\bar{\mu}_{i.} - \bar{\mu}_{..}) + (\bar{\mu}_{.j} - \bar{\mu}_{..}) + (\bar{\mu}_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}) =$$
$$\mu + \alpha_i^A + \alpha_j^B + \alpha_{ij}^{AB}. \tag{4.4.13}$$

Nauji parametrai tenkina sąlygas

$$\sum_i \alpha_i^A = \sum_i \alpha_j^B = \sum_i \alpha_{ij}^{AB} = \sum_j \alpha_{ij}^{AB} = 0, \tag{4.4.14}$$

todėl jų skaičius yra IJ, kaip ir pradinių parametrų μ_{ij}.

Pagrindinės dispersinės analizės faktorių ir jų sąveikos įtakos nebuvimo hipotezės formuluojamos šitaip:

$$H_A : \alpha_i^A = 0, \quad \forall\, i = 1, ..., I, \quad H_B : \alpha_j^B = 0, \quad \forall\, j = 1, ..., J,$$
$$H_{AB} : \alpha_{ij}^{AB} = 0, \quad \forall\, i = 1, ..., I, \quad \forall\, j = 1, ..., J. \tag{4.4.15}$$

Traktuokime šį modelį kaip regresinės analizės modelį, kuriame kintamasis Y prognozuojamas remiantis dviem nominaliosiomis kovariantėmis A ir B, įgyjančiomis atitinkamai I reikšmių $A_1, ..., A_I$ ir J reikšmių $B_1, ..., B_J$.

Tarkime, kad modelis yra adityvusis, t. y. hipotezė H_{AB} yra teisinga. Taikykime pirmiau naudotą nominaliųjų kovariančių kodavimą. Nominaliąją kovariantę A keiskime vektoriumi $(y_1, ..., y_{I-1})^T$, kuris įgyja reikšmes $(0, 0, ..., 0)^T$, $(1, 0, ..., 0)^T$, $(0, 1, ..., 0)^T$,...,$(0, 0, ..., 1)^T$, kai faktorius A yra $A_1, ..., A_I$ lygmenyse. Analogiškai nominaliąją kovariantę B keiskime vektoriumi $(z_1, ..., z_{J-1})^T$, kuris įgyja reikšmes $(0, 0, ..., 0)^T$, $(1, 0, ..., 0)^T$, $(0, 1, ..., 0)^T$,...,$(0, 0, ..., 1)^T$, kai faktorius B yra $B_1, ..., B_J$ lygmenyse. Regresijos koeficientus žymėkime β_i^A, $i = 1, ..., I-1$ ir β_j^B, $\quad j = 1, ..., J-1$, laisvąjį narį žymėkime β_0.

Pernumeruokime stebėjimus Y_{ijk} bet kuria tvarka nuo 1 iki $n = IJK$. Tada gautąjį regresinės analizės modelį galime užrašyti taip:

$$Y_i = \beta_0 + (\beta_1^A y_1^{(i)} + ... + \beta_{I-1}^A y_{I-1}^{(i)}) + (\beta_1^B z_1^{(i)} + ... + \beta_{J-1}^B z_{J-1}^{(i)}) + e_i, \quad i = 1, ..., n. \tag{4.4.16}$$

Regresijos koeficientai susieti su dispersinės analizės modelio parametrais lygybėmis:

$$\mu_{11} = \beta_0, \quad \mu_{1j} = \beta_0 + \beta_{j-1}^B, \quad j = 2, ..., J, \quad \mu_{i1} = \beta_0 + \beta_{i-1}^A, \quad i = 2, ..., I;$$
$$\mu_{ij} = \beta_0 + \beta_{i-1}^A + \beta_{j-1}^B, \quad i = 2, ..., I, \quad j = 2, ..., J;$$

arba lygybėmis

$$\beta_i^A = \alpha_i^A - \alpha_1^A, \quad i = 2, ..., I; \quad \beta_j^B = \alpha_j^B - \alpha_1^B, \quad j = 2, ..., J;$$

Dispersinės analizės hipotezės H_A, H_B yra ekvivalenčios gautojo regresijos modelio dalies koeficientų lygybės nuliui hipotezėms:

$$H_A : \beta_i^A = 0, \quad \forall\, i = 1, ..., I-1, \quad H_B : \beta_j^B = 0, \quad \forall\, j = 1, ..., J-1. \tag{4.4.17}$$

Todėl jas galime patikrinti kriterijumi (4.4.3), kuriame imame $m = I + J - 2$, o $k = I - 1$ ir $J - 1$ atitinkamai pirmosios ir antrosios hipotezių atveju. Gautos statistikos turi atitinkamai Fišerio skirstinius $F(I-1, n-I-J+1)$ ir $F(J-1, n-I-J+1)$ Nesunku patikrinti, kad statistikų skaitikliuose esančios kvadratų sumos sutampa su dispersinėje analizėje apibrėžtomis sumomis SS_A ir SS_B, t. y. gautieji kriterijai sutampa su atitinkamais dispersinės analizės kriterijais.

Jeigu modelis nėra adityvusis ir pereidami prie regresijos modelio dešiniąją lygybės (4.4.16) pusę papildysime dėmeniu

$$\beta_{11}^{AB} y_1^{(i)} z_1^{(i)} + ... + \beta_{I-1,J-1}^{AB} y_{I-1}^{(i)} z_{J-1}^{(i)},$$

tai nesunku patikrinti, kad hipotezės

$$H_A : \alpha_i^A = 0, \quad \forall\, i = 1, ..., I, \quad H_B : \alpha_j^B = 0, \quad \forall\, j = 1, ..., J,$$

nėra ekvivalenčios tvirtinimams $H'_A : \beta_i^A = 0, \forall\ i = 1, ..., I-1$, $H'_B : \beta_j^B = 0\ \forall\, j = 1, ..., J-1$.

Todėl, pereinant prie regresijos modelio, reikėtų išreikšti kuriuos nors parametrus remiantis sąryšiais (4.4.14) taip, kad jų liktų minimalus galimas skaičius $IJ - 1$.

4.4.1 pastaba. Primename, kad nesubalansuoto modelio atveju, kai stebėjimų skaičiai langeliuose K_{ij} yra skirtingi, išvados iš dalies priklauso nuo naudojamos svorių sistemos (žr. 2.3.3 skyrelį). Todėl, pereinant prie regresinės analizės modelio reikia išreikšti kuriuos nors parametrus paliekant jų minimalų galimą skaičių iš sąryšių (2.3.10), (2.3.13), kuriuose atsižvelgiama į naudojamų svorių sistemą.

4.4.4. Dvifaktorė kovariacinė analizė

Tarkime, kad kartu su 4.4.3 skyrelio kintamojo Y matavimais gauname ir trukdančių kovariančių matavimus. Jeigu faktorių įtaka yra adityvi, tai naudojamas *adityvus dvifaktorės kovariacinės analizės modelis*:

$$Y_i = \beta_0 + (\beta_1^A y_1^{(i)} + ... + \beta_{I-1}^A y_{I-1}^{(i)}) + (\beta_1^B z_1^{(i)} + ... + \beta_{J-1}^B z_{J-1}^{(i)}) +$$

$$(\gamma_1 x_1^{(i)} + ... + \gamma_k x_k^{(i)}) + e_i, \quad i = 1, ..., n. \tag{4.4.18}$$

Naudodami kriterijų (4.4.3) hipotezėms

$$H_A : \beta_i^A = 0, \quad \forall\, i = 1, ..., I-1, \quad H_B : \beta_j^B = 0, \quad \forall\, j = 1, ..., J-1, \tag{4.4.19}$$

tikrinti, imame $m = I + J + k - 2$, o $k = I - 1$ ir $J - 1$ atitinkamai pirmosios ir antrosios hipotezių atveju. Gautos statistikos turi atitinkamai Fišerio skirstinius $F(I-1, n-I-J-k+1)$, $F(J-1, n-I-J-k+1)$.

Tikrinant hipotezę $H_{\boldsymbol{\gamma}} : \gamma_1 = ... = \gamma_k = 0$, kriterijaus statistika turi Fišerio skirstinį $F(k, n-I-J-k+1)$, o hipotezę $H_{\gamma_{i_1},...,\gamma_{i_l}} : \gamma_{i_1} = ... = \gamma_{i_l} = 0$ - skirstinį $F(l, n-I-J-k+1)$.

Jeigu modelis nėra adityvus, tai pereidami prie regresinės analizės modelio, parametrus gauname naudodami sąryšius (4.4.14).

Analogiškai regresinės analizės schemose galima nagrinėti ir kitus dispersinės ir kovariacinės analizės modelius.

4.4.2 pastaba. Dispersinės analizės su atsitiktiniais faktoriais arba mišriuosius modelius taip pat galima nagrinėti remiantis regresine analize. Tuo tikslu nagrinėjamas fiktyvus modelis, tariant, kad visi faktoriai yra pastovūs. Gautoji dispersinės analizės lentelė (žr. 2.6 skyrelį) skiriasi tik paskutiniuoju stulpeliu, kuriame pateikiami vidurkiai $\mathbf{E}(MS)$. Todėl, remiantis stulpeliu $\mathbf{E}(MS)$, tereikia patikrinti, ar hipotezių tikrinimo statistikų vardikliuose parašytos teisingos kvadratų sumos (primename, kad statistikų vardikliai gali būti skirtingi priklausomai nuo modelio). Suprantama, kad schemoje su atsitiktiniais faktoriais regresijos modelio koeficientų reikšmingumo tikrinimas neturi prasmės.

4.4.3 pastaba. Tai kad įvairias dispersinės ir kovariacinės analizės schemas galima traktuoti kaip regresinės analizės modelius, leidžia sukurti bendrą programų sistemą, tinkamą statistiniams duomenims analizuoti naudojant bet kurį tiesinį modelį. Pavyzdžiui, SAS programų sistemoje yra procedūra GLM, kuri tinka tiesinio modelio statistiniams duomenims analizuoti (dispersinė analizė, regresinė analizė, kovariacinė analizė) netgi tuo atveju, kai nepriklausomas kintamasis Y yra vektorinis. Greta to SAS sistemoje yra procedūra ANOVA ir procedūra REG, skirtos duomenims analizuoti dispersinės ir regresinės analizės schemose. Pastarosiose procedūrose yra numatyta daugiau galimybių negu bendroje procedūroje GLM, atsižvelgiant į dispersinėje ir regresinėje analizėje sprendžiamų uždavinių specifiką (žr.[13]).

4.5. Faktoriniai eksperimentai 2^m

Baigdami šį skyrelį detaliau panagrinėsime eksperimentų schemą, kai kintamasis Y priklauso nuo gana didelio skaičiaus m nominalių ar tolydžių kovariančių $X_1, ..., X_m$. Tarkime, kad eksperimentą atlikti yra brangu ir jam reikia daug laiko. Pavyzdžiui, Y gali reikšti produkcijos išeigą, o kovariantės $X_1, ..., X_m$ yra technologinio proceso charakteristikos. Kintamasis Y gali reikšti plieno tvirtumą, o kovariantės $X_1, ..., X_m$ – įkrovos sudėtį ir lydymo krosnies darbo režimo charakteristikas ir pan. Suprantama, jeigu visuose atliktuose eksperimentuose kuri nors kovariantė X_i įgis tą pačią reikšmę, tai gauti duomenys nesuteikia jokios informacijos apie Y skirstinio priklausomybę nuo kovariantės X_i pasikeitimo. Todėl kiekviena kovariantė turėtų įgyti bent jau dvi skirtingas reikšmes. Mažindami eksperimentų skaičių tarsime, kad visos kovariantės įgyja dvi reikšmes.

Tokie eksperimentai vadinami *faktoriniais eksperimentais* 2^m. Duomenis galime analizuoti pagal daugiafaktorės dispersinės analizės schemą, kai kiekvienas faktorius gali būti dviejų lygmenų, arba pagal regresinės analizės schemą, kai yra m nominaliųjų kovariančių, įgyjančių dvi skirtingas reikšmes.

Jeigu kovariančių vektoriaus dimensija m yra didelė, tai kryžminės klasi-

fikacijos schemoje, netgi kai kiekviena kovariantė įgyja tik dvi reikšmes, eksperimentų skaičius 2^m gali būti nepriimtinas. Pavyzdžiui, jeigu $m = 10$, tai kryžminės klasifikacijos schemoje rekėtų atlikti $2^{10} = 1024$ eksperimentus. Todėl eksperimentus galėsime atlikti tiktai kai kuriuose iš šių taškų. Kyla klausimas, kuriuose didumo 2^m galimų reikšmių aibės taškuose reikia atlikti eksperimentus, kad gautieji duomenys leistų spręsti iškeltą uždavinį pagal galimybę sumažinant eksperimentų skaičių.

Bene svarbiausioji faktorinių eksperimentų 2^m taikymo sritis yra tokio uždavinio sprendimas. Tarkime, kintamasis Y tam tikru tikslumu gali būti išreikštas $X_1, ..., X_m$ funkcija

$$Y = f(X_1, ..., X_m) + e.$$

Reikia rasti tokias kovariančių $X_1, ..., X_m$ reikšmes, kurioms esant Y įgyja artimą maksimumui reikšmę. Pavyzdžiui, taip parinkti technologinio proceso charakteristikas, kad geros produkcijos išeiga būtų maksimali, arba taip parinkti lydymo krosnies įkrovą ir jos režimą, kad gautojo plieno kažkoks parametras būtų kuo didesnis, ir pan.

Tradicinis metodas yra toks. Fiksuojamos visos kovariantės, išskyrus vieną, ir bandoma rasti tos atskiros kovariantės geriausioji reikšmė. Paskui keičiama kita kovariantė esant fiksuotoms likusioms ir t. t.

Pastaruoju metu plačiai naudojamas *greičiausiojo pakilimo* arba *evoliucinio planavimo* metodas, kuriame naudojami regresinės (ar dispersinės analizės) metodai. Jo esmė yra tokia. Pradinio eksperimentinio taško aplinkoje funkcija $f(X_1, ..., X_m)$ aproksimuojama tiesine

$$f(X_1, ..., X_m) \approx \beta_0 + \beta_1 X_1 + ... + \beta_m X_m. \tag{4.5.1}$$

Atliekamas nedidelis eksperimentas, kuris leistų įvertinti parametrus $\beta_0, \beta_1, ..., \beta_m$. Atsižvelgiant į gautos plokštumos lygtį, nustatoma kryptis, kuria reikia judėti funkcijos maksimumo link. Kitas eksperimentas atliekamas taške, kuris parenkamas funkcijos maksimumo kryptimi. Kai pasiekiamas taškas arti funkcijos maksimumo, t. y. visų regresijos parametrų įvertiniai reikšmingai nesiskiria nuo nulio, gali būti atliktas didesnės apimties eksperimentas, leidžiantis tiksliau apibūdinti funkcijos f pavidalą, pavyzdžiui, aproksimuojant jį antrojo laipsnio polinomu

$$f(X_1, ..., X_m) \approx \beta_0 + \sum_i \beta_i X_i + \sum_i \sum_j \beta_{ij} X_i X_j. \tag{4.5.2}$$

Nesunku suvokti, kad net ir dvimačiu atveju greičiausiojo pakilimo metodui reikės gerokai mažiau eksperimentų. Šio metodo efektyvumas ypač padidėja, kai kovariančių skaičius m didelis.

4.5.1. Faktoriniai eksperimentai 2^2

Aptarsime paprasčiausią atvejį, kai kovariančių vektoriaus dimensija $m = 2$. Tegu pirmoji kovariantė X_1 igyja reikšmes $X_{11} < X_{12}$, o antroji kovariantė X_2 – reikšmes $X_{21} < X_{22}$. Analizės patogumui pažymėkime $Z_1 = (2X_1 - X_{11} -$

$X_{12})/(X_{12}-X_{11})$, $Z_2=(2X_2-X_{21}-X_{22})/(X_{22}-X_{21})$. Tada naujos kovariantės Z_1 ir Z_2 įgyja reikšmę -1, kai X_1 ar X_2 yra apatinio lygmens ir įgyja reikšmę $+1$, kai X_1 ar X_2 yra viršutinio lygmens. Eksperimento rezultatai pateikti 4.5.1 lentelėje.

4.5.1 lentelė. Perkoduoti eksperimentų rezultatai

Y_i	Z_{0i}	Z_{1i}	Z_{2i}	Kodas	$Z_{1i}Z_{2i}$	Replika 2^{3-1}
Y_1	$+1$	-1	-1	(1)	$+1$	c
Y_2	$+1$	$+1$	-1	a	-1	a
Y_3	$+1$	-1	$+1$	b	-1	b
Y_4	$+1$	$+1$	$+1$	ab	$+1$	abc

Pirmame stulpelyje pateikti nepriklausomo kintamojo Y matavimai; Z_{0i} yra koeficientai prie laisvojo nario. Tiesinės regresijos modelis

$$Y_i=\beta_0 Z_{0i}+\beta_1 Z_{1i}+\beta_2 Z_{2i}+e_i,\quad i=1,2,3,4; \tag{4.5.3}$$

čia tariama, kad e_i nepriklausomi normalieji a. d. $e_i\sim N(0,\sigma^2)$.

Trumpumo sumetimais eksperimentas užrašomas koduotu pavidalu. Pirmajam kintamajam priskirkime raidę a, antrajam – raidę b. Jeigu kintamasis yra žemutinio lygmens, atitinkama raidė nerašoma. Kai visi kintamieji yra žemutiniame lygmenyje, tai tokį tašką žymėsime (1). Remiantis šiomis taisyklėmis eksperimentų planas iš 4.5.1 lentelės užrašomas kaip $((1),a,b,ab)$; žr. 4.5.1 lentelės penktą stulpelį.

Eksperimentų planas iš 4.5.1 lentelės yra ortogonalus ir $\boldsymbol{A}^T\boldsymbol{A}$ yra diagonali matrica, kurios diagonaliniai elementai vienodi ir lygūs 4. Gauname parametrų įvertinius

$$\hat\beta_0=\frac{Y_1+Y_2+Y_3+Y_4}{4},\quad \hat\beta_1=\frac{-Y_1+Y_2-Y_3+Y_4}{4},$$

$$\hat\beta_2=\frac{-Y_1-Y_2+Y_3+Y_4}{4}.$$

Įvertiniai yra nepriklausomi ir turi vienodas dispersijas

$$\boldsymbol{V}(\hat\beta_0)=\boldsymbol{V}(\hat\beta_1)=\boldsymbol{V}(\hat\beta_2)=\sigma^2/4.$$

Liekamoji kvadratų suma

$$SS_E=\sum_i Y_i^2-4\hat\beta_0^2-4\hat\beta_1^2-4\hat\beta_2^2\sim\sigma^2\chi_1^2$$

gali būti panaudota dispersijai įvertinti

$$\hat\sigma^2=SS_E.$$

Jeigu dispersijos vertinti nereikia, tai galima nagrinėti regresijos modelį

$$Y_i=\beta_0 Z_{0i}+\beta_1 Z_{1i}+\beta_2 Z_{2i}+\beta_{12}Z_{1i}Z_{2i}+e_i$$

ir papildomai įvertinti parametrą β_{12} prie sandaugos $Z_{1i}Z_{2i}$. Nesunku patikrinti, kad plano matricos stulpelis, atitinkantis sandaugą $Z_{1i}Z_{2i}$ (žr. 4.5.1 lentelę), taip pat ortogonalus su kitais stulpeliais. Todėl

$$\hat{\beta}_{12} = \frac{Y_1 - Y_2 - Y_3 + Y_4}{4}, \quad \boldsymbol{V}(\hat{\beta}_{12}) = \sigma^2/4.$$

4.5.2. Faktoriniai eksperimentai 2^3

Tarkime, kad kovariančių vektoriaus dimensija yra $m = 3$. Kiekviena kovariantė įgyja dvi skirtingas reikšmes. Įvedę analogišką kodavimą gausime, kad eksperimentai atliekami kubo viršūnėse (Z_1, Z_2, Z_3), $Z_i = +1, -1$; $i = 1, 2, 3$. Šio eksperimento kodinį planą galima gauti taip. Priskirkime trečiajai kovariantei raidę c. Iš pradžių atlikime 4 eksperimentus, kai trečioji kovariantė yra žemutinio lygmens $((1), a, b, ab)$, paskui 4 eksperimentus, kai trečioji kovariantė yra viršutinio lygmens (c, ac, bc, abc). Sujungę gausime faktorinio eksperimento 2^3 kodinį planą. Stebėjimus surašykime į analogišką 4.5.1 lentelę (pakanka nurodyti tik kovariančių ženklą).

4.5.2 lentelė. Faktorinis eksperimentas 2^3 .

Y_i	Z_{0i}	Z_{1i}	Z_{2i}	Z_{3i}	Kodas	$Z_{1i}Z_{2i}$	$Z_{1i}Z_{3i}$	$Z_{2i}Z_{3i}$	$Z_{1i}Z_{2i}Z_{3i}$
Y_1	+	-	-	-	(1)	+	+	+	-
Y_2	+	+	-	-	a	-	-	+	+
Y_3	+	-	+	-	b	-	+	-	+
Y_4	+	+	+	-	ab	+	-	-	-
Y_5	+	-	-	+	c	+	-	-	+
Y_6	+	+	-	+	ac	-	+	-	-
Y_7	+	-	+	+	bc	-	-	+	-
Y_8	+	+	+	+	abc	+	+	+	+

Vertindami regresijos

$$Y_i = \beta_0 Z_{0i} + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + e_i, \quad i = 1, ..., 8,$$

parametrus pažymėsime, kad plano matrica ortogonali, o $\boldsymbol{A}^T\boldsymbol{A}$ yra diagonali, turi vienodus diagonalinius elementus lygius 8. Parametrų įvertiniai

$$\hat{\beta}_j = \frac{1}{8}\sum_i Y_i Z_{ji}, \quad \boldsymbol{V}(\hat{\beta}_j) = \frac{\sigma^2}{8}, \quad j = 0, 1, 2, 3.$$

Dispersijai įvertinti lieka 4 laisvės laipsniai

$$\hat{\sigma}^2 = \frac{1}{4}SS_E = \frac{1}{4}\{\sum_i Y_i^2 - 8\sum_{j=0}^{3}\hat{\beta}_j^2\}, \quad \frac{SS_E}{\sigma^2} \sim \chi^2(4).$$

Papildžius plano matricą trimis papildomais ortogonaliais stulpeliais, atitinkančiais sandaugas $Z_{1i}Z_{2i}$, $Z_{1i}Z_{3i}$, $Z_{2i}Z_{3i}$, galima papildomai įvertinti dar tris

parametrus

$$\hat{\beta}_{jl} = \frac{1}{8}\sum_i Y_i Z_{ji} Z_{li}, \quad \boldsymbol{V}(\hat{\beta}_{jl}) = \frac{\sigma^2}{8},$$

paliekant vieną laisvės laipsnį dispersijai vertinti.

Pagaliau, jei nereikia vertinti dispersijos, tai galima papildomai įvertinti parametrą β_{123} prie sandaugos $Z_{1i}Z_{2i}Z_{3i}$

$$\hat{\beta}_{123} = \frac{1}{8}\sum_i Y_i Z_{1i} Z_{2i} Z_{3i}, \quad \boldsymbol{V}(\hat{\beta}_{123}) = \frac{\sigma^2}{8}.$$

4.5.3. Faktoriniai eksperimentai 2^m

Jeigu kovariančių vektoriaus dimensija $m > 3$ ir kiekviena kovariantė įgyja po dvi skirtingas reikšmes, tai, atlikę kodavimą, gausime, kad eksperimentai atliekami m-mačio kubo viršūnėse $\{(Z_1, ..., Z_m) : Z_i = \pm 1, i = 1, ..., m\}$. Bendras stebėjimų skaičius $n = 2^m$. Tokio eksperimento koduotą planą gauname analogiškai kaip ir pirmiau iš vienetu mažesnės dimensijos plano. Pavyzdžiui, koduotą faktorinio eksperimento 2^4 planą gauname taip: surašome 8 kodus iš 4.5.2 lentelės 6 stulpelio; paskui prie kiekvieno iš jų pirašome raidę d, priskirtą ketvirtajai kovariantei. Gausime pilną faktorinį planą, kai stebėjimų skaičius $n = 2^4 = 16$. Analogiškai išnagrinėtiems atvejams $m = 2$, $m = 3$, tiesinės regresijos

$$Y_i = \beta_0 Z_{0i} + \beta_1 Z_{1i} + ... + \beta_m Z_{mi} + e_i, \quad i = 1, ..., n,$$

parametrai įvertinami labai paprastai, nes plano matricos stulpeliai yra ortogonalūs:

$$\sum_{i=1}^n Z_{ji} Z_{li} = 0, \quad j \neq l, \quad \sum_{i=1}^n Z_{ji}^2 = n, \quad j = 1, ..., m.$$

Gauname parametrų $\beta_0, \beta_1, ..., \beta_m$ įvertinius

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^n Y_i Z_{ji}, \quad j = 0, 1, ..., m. \tag{4.5.4}$$

Įvertiniai $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$ yra nepriklausomi ir įvertinami vienodu tikslumu

$$\boldsymbol{V}(\hat{\beta}_j) = \frac{\sigma^2}{n}, \quad j = 0, 1, ..., m.$$

Liekamoji kvadratų suma

$$SS_E = \sum_{i=1}^n Y_i^2 - n\sum_{j=1}^m \hat{\beta}_j^2 \sim \sigma^2 \chi^2_{n-m-1}. \tag{4.5.5}$$

Laisvės laipsnių skaičius dispersijai vertinti yra

$$n - m - 1 = 2^m - m - 1. \tag{4.5.6}$$

4.5.1 pastaba. Faktoriniai planai pasižymi šia svarbia savybe. Pažymėkime $\hat{Y}$ kintamojo Y prognozę taške $(\tilde{Z}_0, \tilde{Z}_1, ..., \tilde{Z}_m)$. Tada

$$\boldsymbol{V}(\hat{Y}) = \frac{\sigma^2}{n}(1 + \sum_{j=1}^{m} \tilde{Z}_j^2) = \frac{\sigma^2}{n}(1 + \rho^2). \tag{4.5.7}$$

Prognozės tikslumas visuose taškuose, kurie yra nutolę vienodu atstumu nuo eksperimento centro, yra vienodas. Tokie eksperimentų planai vadinami *rototabiliais*. Kadangi praktiškai nežinoma, kurioje pusėje yra maksimumo taškas, tai ši plano savybė yra pageidautina.

4.5.2 pastaba. Faktoriniai eksperimentai 2^m yra atskiras dispersinės analizės atvejis, kai m faktorių turi po du lygmenis ir dalyvauja eksperimente pagal pilną kryžminės klasifikacijos schemą. Kiekvieno faktoriaus lygmenys gali būti kokios nors kokybinės ar kiekybinės kovariantės 2 skirtingos reikšmės. Duomenų analizę galima atlikti dispersinės analizės terminais, apibrėžiant kvadratų sumas SS, charakterizuojančias atskirų faktorių ar jų rinkinių įtaką, taip pat liekamąją kvadratų sumą dispersijai vertinti (žr. 4.3.1 skyrelį). Norėdami išskirti evoliucinio planavimo metodą, duomenų analizę atlikome regresinės analizės schemoje. Šių schemų sąryšis aptartas 4.4 skyrelyje (žr. 4.11–4.19 pratimus). Įvertinus regresijos lygties parametrus, dispersinės analizės lentelę galima užpildyti tokiu būdu: kvadratų sumos, apibūdinančios atskirų faktorių A_i įtaką, yra $SS_{A_i} = n\hat{\beta}_i^2, i = 1, ..., m$; kvadratų sumos, apibūdinančios dviejų faktorių A_i ir A_j, $i \neq j$ sąveiką, yra $SS_{A_iA_j} = n\hat{\beta}_{ij}^2$ ir t. t.; laisvės laipsnių skaičius lygūs 1.

Nors faktoriniai planai 2^m turi gerų savybių, tačiau juos įgyvendinti, kai eksperimentai brangūs, o kovariančių vektoriaus dimensija didelė, gali būti nepriimtina dėl per didelio eksperimentų skaičiaus. Pavyzdžiui, jei $m = 10$, tai pilno faktorinio plano eksperimentų skaičius yra $n = 2^{10} = 1024$. Minėjome, kad naudojant evoliucinio planavimo metodą, reikia atlikti minimalų eksperimentų skaičių, kurie leistų įvertinti regresijos plokštumą. Kai m didelis, iš (4.5.6) matome, kad tik nedidelė laisvės laipsnių dalis panaudojama regresijos parametrams vertinti, o didžioji dalis tenka dispersijai vertinti. Pavyzdžiui, jeigu $m = 7$, tai tik 8 laisvės laipsniai bus panaudota vertinant tiesinės regresijos parametrus, o likusieji $\nu = 128 - 8 = 120$ laisvės laipsnių bus panaudota vertinant dispersiją.

Natūralu sumažinti faktorinio plano 2^m eksperimentų skaičių, įgyvendinant pusę, ketvirtadalį ir pan. plano, kad laisvės laipsnių skaičius, tenkantis dispersijai vertinti, būtų mažesnis. Taškai, kuriuose bus atliekamas mažesnis eksperimentas, turi būti parenkami specialiai, nes, pavyzdžiui, jeigu atliksime pusę faktorinių eksperimentų 2^m taip, kad kuri nors kovariantė bus to paties lygmens, tai koeficiento prie šios kovariantės regresijos lygtyje įvertinti negalėsime.

4.5.4. Faktorinių eksperimentų 2^m replikos

Tarkime, kad nuspręsta atlikti pusę faktorinio eksperimento 2^3 iš 4.5.2 lentelės. Tuo tikslu 4.5.1 lentelėje sandaugai Z_1Z_2 priskirkime trečiąją kovariantę

$$Z_3 = Z_1Z_2. \tag{4.5.8}$$

Tada 4.5.1 lentelę galime traktuoti kaip pusę eksperimento iš 4.5.2 lentelės. Sakoma, kad atliekame faktorinio eksperimento 2^3 *pusinę repliką* 2^{3-1}, susidedančią iš keturių eksperimentų. Lygybė (4.5.8) vadinama *repliką generuojančiu sąryšiu*. Ši replika kodiniu pavidalu užrašyta 4.5.1 lentelės paskutiniame stulpelyje. Prilyginę

$$Z_3 = -Z_1 Z_2, \tag{4.5.9}$$

gautume kitą repliką 2^{3-1}, kurios kodinis pavidalas $((1), ac, bc, ab)$.

Įveskime simbolinę daugybą, tardami, kad $Z_i Z_i = Z_i^2 = 1$. Padauginę lygybių (4.5.8) ir (4.5.9) abi puses iš Z_3 gausime

$$1 = Z_1 Z_2 Z_3, \qquad 1 = -Z_1 Z_2 Z_3. \tag{4.5.10}$$

Šie sąryšiai vadinami *replikas generuojančiais kontrastais*. Jais galima nustatyti, kurie parametrai regresijos lygtyje neatskiriami atliekant ne visą eksperimentą.

Daugindami (4.5.10) abi lygybės puses iš Z_1 gausime

$$Z_1 = Z_2 Z_3, \qquad Z_1 = -Z_2 Z_3. \tag{4.5.11}$$

Tai reiškia, kad, vertindami parametrą β_1 prie kovariantės Z_1, faktiškai pirmoje replikoje vertiname parametrą $\beta_1 + \beta_{23}$; antroje replikoje – parametrą $\beta_1 - \beta_{23}$. Analogiškai gauname, kad pirmoje replikoje vietoje parametrų β_2 ir β_3 vertiname parametrus $\beta_2 + \beta_{13}$ ir $\beta_3 + \beta_{12}$, o antroje replikoje – parametrus $\beta_2 - \beta_{13}$ ir $\beta_3 - \beta_{12}$.

Vadinasi, jeigu regresijos koeficientai β_{jl} prie sandaugų $Z_j Z_l$ nelygūs 0, tai jų atskirai įvertinti iš atskiros replikos stebėjimų negalime. Sujungę abi pusines replikas į vieną visumą, koeficientus β_j galėsime atskirti nuo β_{jl}, nes pastarieji skirtingose pusiau replikose įeina su priešingais ženklais. Tai ir buvo atlikta 4.5.2 skyrelyje.

Iš replikos 2^{3-1}, pateiktos 4.5.1 lentelėje, galima įvertinti regresijos lygties

$$Y_i = \beta_0 Z_{0i} + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + e_i$$

koeficientus $\beta_j, j = 0, 1, 2, 3$. Skyrelyje 4.5.1 reikia pasirinkti $\beta_{12} = \beta_3$. Dispersijai įvertinti laisvės laipsnių nelieka. Norint įvertinti dispersiją reikėtų atlikti visą 2^3 eksperimentų planą. Arba dispersijai įvertinti galima papildomai atlikti ne mažiau kaip du stebėjimus eksperimentų centre $(Z_1, Z_2) = (0, 0)$. Tokių taškų prijungimas nepažeidžia plano matricos $\boldsymbol{A}$ ortogonalumo.

Kai $m > 3$, tai iš pusinės replikos galima įvertinti ir regresijos lygties parametrus, ir dispersiją. Pavyzdžiui, eksperimentą 4.5.2 lentelėje traktuokime kaip pusinę faktorinio eksperimento 2^4 repliką. Generuojančiu repliką sąryšiu natūralu pasirinkti

$$Z_4 = Z_1 Z_2 Z_3.$$

Replika kodiniu pavidalu yra $((1), ad, bd, ab, cd, ac, bc, abcd)$.

Iš šios replikos eksperimentų galima įvertinti regresijos lygties

$$Y_i = \beta_0 Z_{0i} + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + e_i$$

parametrus ir trys laisvės laipsniai lieka dispersijai vertinti.

Dispersijai vertinti galima palikti vieną ar du laisvės laipsnius, įvertinant papildomai du ar vieną parametrą β_{jl}.

Jeigu m didelis, tai ir pusinės replikos eksperimentų skaičius gali būti per didelis. Tada faktorinį planą galima dalyti į smulkesnes replikas: ketvirtines, aštuntines ir t. t. Replika 2^{m-k} reiškia, kad visas faktorinis 2^m planas dalijamas pusiau k kartų, t. y. vietoje 2^m eksperimentų atliekama tik 2^{m-k} eksperimentų. Norint apibrėžti tokią repliką, reikia sudaryti k skirtingų generuojančių sąryšių.

Eksperimentą 4.5.2 lentelėje traktuokime kaip faktorinio eksperimento 2^6 aštuntinę repliką 2^{6-3}. Pasirinkime generuojančius sąryšius

$$Z_4 = Z_1 Z_2, \quad Z_5 = Z_1 Z_3, \quad Z_6 = Z_2 Z_3. \tag{4.5.12}$$

Priskyrus kintamiesiems $Z_1, ..., Z_6$ atitinkamai raides a, b, c, d, e, f, šios replikos planas kodiniu pavidalu yra $(def, af, be, abd, cd, ace, bcf, abcdef)$.

Pagal šį dalinį eksperimentų planą įvertinami regresinės lygties parametrai $\beta_0, \beta_1, ..., \beta_6$ ir vienas laisvės laipsnis lieka dispersijai įvertinti, vietoje $2^6 = 64$ viso faktorinio plano eksperimentų atliekant tik 8 eksperimentus.

Kontrastai, atitinkantys generuojančius sąryšius (4.5.12), yra

$$1 = Z_1 Z_2 Z_4 = Z_1 Z_3 Z_5 = Z_2 Z_3 Z_6 = Z_1 Z_2 Z_5 Z_6 = Z_1 Z_3 Z_4 Z_6 = Z_2 Z_3 Z_4 Z_5.$$

Todėl, pavyzdžiui, vertindami parametrą β_1, faktiškai vertiname parametrą

$$\beta_1 + \beta_{24} + \beta_{35} + \beta_{256} + \beta_{346} + \beta_{1236} + \beta_{12345}.$$

Viso eksperimentų plano 2^m padalijimas į replikas gali būti reikalingas ir tada, kai visų eksperimentų negalime atlikti per pakankamai trumpą laiką ir galbūt negalime užtikrinti identiškų eksperimento atlikimo sąlygų. Tada eksperimentas 2^m atliekamas dalimis, padalijant jį į blokus (replikas). Suprantama, padalijimas į blokus turėtų būti atliekamas taip, kad trukdantysis faktorius (blokas) neiškreiptų tų parametrų, kuriuos laikome reikšmingais, įvertinių. Pavyzdžiui, jeigu visi eksperimentai, kai kovariantė $Z_1 = -1$ pateks į vieną bloką, o visi eksperimentai, kai $Z_1 = +1$, – į kitą, tai parametro β_1 įvertinys bus iškreiptas, jeigu skirtingų blokų eksperimentai atlikti skirtingomis sąlygomis (faktoriaus A_1 įtaka yra sumaišyta su tarpblokiniu efektu).

Panagrinėkime konkrečius pavyzdžius. Tarkime, kad, pavyzdžiui, vienu metu galime atlikti tik 8 faktorinio plano $2^4 = 16$ eksperimentus. Tegu pirmosios replikos generuojantis kontrastas yra

$$Z_1 Z_2 Z_3 Z_4 = 1.$$

Tada pirmąją repliką sudaro eksperimentai, kurių kodai $\{(1), ab, ac, ad, bc, bd, cd, abcd\}$, o antrąją – likusieji eksperimentai $\{a, b, c, d, abc, abd, acd, bcd\}$. Replikos gaunamos dauginant pagal pirmiau įvestos simbolinės sandaugos taisykles kintamųjų Z_i kombinacijas, atitinkančias eksperimentų kodus, ir į vieną repliką

įtraukiant tuos eksperimentus, kuriems gautos sandaugos turi lyginį daugiklių skaičių, o į kitą – nelyginį.

Tarkime, kad sąveikos po tris faktorius yra lygios 0. Tada, atlikus trifaktorę visų $n = 16$ eksperimentų analizę, galima sudaryti kvadratų sumas $SS_{A_i}, i = 1, ..., 4$, apibūdinančias atskirų faktorių įtaką, kvadratų sumas $SS_{A_iA_j}$, $i \neq j$, apibūdinančias dviejų faktorių įtaką; laisvės laipsnių skaičius lygūs 1. Kvadratų sumą $SS_{A_1A_2A_3} + SS_{A_1A_3A_4} + SS_{A_2A_3A_4}$ galima panaudoti dispersijai vertinti; laisvės laipsnių skaičius 3. O kvadratų suma $SS_{A_1A_2A_3A_4}$ yra sumaišyta su tarpblokiniu efektu ir jos naudoti be papildomos analizės dispersijai vertinti nerekomenduojama.

Jeigu iš $n = 2^4 = 16$ eksperimentų vienu metu galime atlikti tik 4 eksperimentus, tai visus eksperimentus teks padalyti į 4 replikas. Tada su tarpblokiniu efektu bus sumaišytos trys sąveikos. Tegu pirmosios replikos generuojantys kontrastai yra

$$Z_1Z_2Z_3 = 1, \quad Z_2Z_3Z_4 = 1.$$

Į pirmąją repliką įtraukiame tuos eksperimentus, kuriems atitinkamos kintamųjų kombinacijos, padaugintos iš kontrastų duoda lygines sandaugas. Gauname repliką $\{(1), bc, acd, abd\}$. Kitos replikos gaunamos, kai atitinkamos sandaugos yra lyginė ir nelyginė; nelyginė ir lyginė; nelyginė ir nelyginė: $\{ab, ac, bcd, d\}$; $\{a, abc, cd, bd\}$; $\{b, c, abcd, ad\}$.

Tarkime, kad sąveikos po tris ir keturis faktorius yra lygios 0. Tada, atlikus trifaktorę visų $n = 16$ eksperimentų analizę, galima sudaryti kvadratų sumas $SS_{A_i}, i = 1, ..., 4$, apibūdinančias atskirų faktorių įtaką, kvadratų sumas $SS_{A_iA_j}$, $i \neq j$, apibūdinančias dviejų faktorių įtaką (išskyrus A_1 ir A_4 sąveiką, kuri sumaišyta su tarpblokiniu efektu, nes $Z_1Z_2Z_3\dot{Z}_2Z_3Z_4 = Z_1Z_4 = 1$); laisvės laipsnių skaičius lygūs 1. Kvadratų sumą $SS_{A_1A_2A_4} + SS_{A_1A_3A_4} + SS_{A_1A_2A_3A_4}$ galima panaudoti dispersijai vertinti; laisvės laipsnių skaičius 3. O kvadratų sumos $SS_{A_1A_4}$, $SS_{A_1A_2A_3}$, $SS_{A_2A_3A_4}$ yra sumaišytos su tarpblokiniu efektu ir jų naudoti be papildomos analizės dispersijai vertinti nerekomenduojama.

4.6. Pratimai

4.1. Tegu $Y_{ij} = \mu_i + \gamma_1 X_{ij}^{(1)} + \gamma_2 X_{ij}^{(2)} + e_{ij}$, $i = 1, ..., I, j = 1, ..., J$, o a. d. $\{e_{ij}\}$ nepriklausomi ir normalieji $e_{ij} \sim N(0, \sigma^2)$. a) Raskite parametrų γ_1 ir γ_2 mažiausiųjų kvadratų įvertinius $\hat{\gamma}_1$ ir $\hat{\gamma}_2$. b) Raskite įvertinių $\hat{\gamma}_1$ ir $\hat{\gamma}_2$ kovariacijų matricą. Kokiomis sąlygomis ivertiniai $\hat{\gamma}_1$ ir $\hat{\gamma}_2$ nekoreliuoti?

4.2. Tegu $Y_{ijk} = \mu_{ij} + \gamma_{ij} X_{ijk} + e_{ijk}$, $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$, o a. d. $\{e_{ijk}\}$ nepriklausomi ir normalieji $e_{ijk} \sim N(0, \sigma^2)$. a) Raskite kriterijų hipotezei $H : \gamma_{ij} = \gamma$ tikrinti. b) Tardami, kad hipotezė H teisinga, sudarykite parametro γ pasikliovimo intervalą.

4.3. Tegu $Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma X_{ijk} + e_{ijk}$, $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$, o a. d. $\{e_{ijk}\}$ nepriklausomi ir normalieji $e_{ijk} \sim N(0, \sigma^2)$; $\sum_i \alpha_i = 0$, $\sum_j \beta_{ij} = 0, i = 1, ..., I$. a) Raskite kriterijų hipotezei $H : \gamma = 0$ tikrinti. b) Raskite kriterijų hipotezei $H_A : \alpha_i = 0, i = 1, ..., I$, tikrinti.

4.4. Tegu $Y_{ij} = \mu_i + \gamma_i X_j + e_{ij}$, $i = 1, 2, j = 1, ..., J$, o a. d. $\{e_{ij}\}$ nepriklausomi ir normalieji $e_{ij} \sim N(0, \sigma^2)$. Remdamiesi kovariacine analize Raskite kriterijų hipotezei H :

$\gamma_1 = \gamma_2$ tikrinti. Įsitikinkite, kad gautasis kriterijus yra ekvivalentus dviejų regresijos tiesių lygiagretumo kriterijui.

4.5. Lentelėje pateiktos bandelių, iškeptų iš 100 (g) tešlos, apimtis Y priklausomai nuo 17 miltų rūšių ir nuo juose esančio bromistinio kalio kiekio X (g), kai $X = 0, 1, 2, 3, 4$ (žr. [14]).

A	X					A	X				
	0	1	2	3	4		0	1	2	3	4
A_1	950	1075	1055	975	880	A_{10}	885	1000	1015	960	895
A_2	890	980	955	865	825	A_{11}	895	935	965	950	920
A_3	830	850	820	770	735	A_{12}	685	835	870	875	880
A_4	770	815	765	725	700	A_{13}	615	665	650	680	660
A_5	860	1040	1065	975	945	A_{14}	885	910	890	835	785
A_6	835	960	985	915	845	A_{15}	985	1075	1070	1015	1005
A_7	795	900	905	880	785	A_{16}	710	750	740	725	720
A_8	800	860	870	850	850	A_{17}	785	845	865	825	820
A_9	750	940	1000	960	960						

a) Atlikite vienfaktorę dispersinę analizę neatsižvelgdami į kintamąjį X. b) Atlikite kovariacinę analizę, tarę, kad X yra kiekybinis kintamasis, ir pasirinkdami antrojo ir trečiojo laipsnio polinomus kintamojo X atžvilgiu.

4.6. Lentelėje pateikti duomenys, gauti atlikus eksperimentą pagal keturfaktorės dispersinės analizės pilną kryžminės klasifikacijos schemą. Registruojamas tam tikro maisto produkto drėgnumas priklausomai nuo druskos rūšies (faktorius A), druskos kiekio (faktorius B), rūgšties lygio (faktorius C) ir dviejų skirtingų priemaišų (faktorius D) [14].

		A_1			A_2			A_3		
		B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
C_1	D_1	8	17	22	7	26	34	10	24	39
C_1	D_2	5	11	16	3	17	32	5	14	33
C_2	D_1	8	13	20	10	24	34	9	24	36
C_2	D_2	4	10	15	5	19	29	4	16	34

a) Atlikite keturfaktorę dispersinę analizę, kai trijų ir keturių faktorių sąveikos nereikšmingos. b) Atlikite kovariacinę analizę tarę, kad faktorius B kiekybinis ($B_1 = 1$, $B_2 = 2$, $B_3 = 3$).

4.7. Lentelėje pateikti duomenys apie krakmolo plėvelės tvirtumą Y priklausomai nuo krakmolo tipo (faktorius A); A_1 – iš kviečių; A_2 – iš ryžių; A_3 – iš kukurūzų; A_4 – iš bulvių; A_5 – iš saldžiųjų bulvių) ir nuo plėvelės storio X [14].

A_1		A_2		A_3		A_4		A_5	
Y	X	Y	X	Y	X	Y	X	Y	X
263,7	5,0	556,7	7,1	731,0	8,0	983,3	13,0	837,1	9,4
130,8	3,5	552,5	6,7	710,0	7,3	958,8	13,3	901,2	10,6
382,9	4,7	397,5	5,6	604,7	7,2	747,8	10,7	595,7	9,0
302,5	4,3	532,3	8,1	508,8	6,1	866,0	12,2	510,0	7,6
213,3	3,8	587,8	8,7	393,0	6,4	810,8	11,6		
132,1	3,0	520,9	8,3	416,0	6,4	950,0	9,7		
292,0	4,2	574,3	8,4	400,0	6,9	1282,0	10,8		
315,5	4,5	505,0	7,3	335,6	5,8	1233,8	10,1		
262,4	4,3	604,6	8,5	306,4	5,3	1660,0	12,7		
314,4	4,1	522,5	7,8	426,0	6,7	746,0	9,8		
310,8	5,5	555,0	8,0	382,5	5,8	650,0	10,0		
280,8	4,8	561,1	8,4	340,8	5,7	992,5	13,8		
331,7	4,8			436,7	6,1	896,5	13,3		
672,5	8,0			333,3	6,2	873,9	12,4		
496,0	7,4			382,3	6,3	924,4	12,2		
311,9	5,2			397,7	6,0	1050,0	14,1		
276,7	4,7			619,1	6,8	973,3	13,7		
325,7	5,4			857,3	7,9				
310,8	5,4			592,5	7,2				
288,0	5,4								
269,3	4,9								

a) Atlikti kintamojo Y vienfaktorę dispersinę analizę priklausomai nuo faktoriaus A; b) Atlikite kintamojo Y regresinę analizę neatsižvelgdami į faktorių A. c) Atlikite kovariacinę analizę ir palyginkite gautus rezultatus.

4.8. Lentelėje pateiktas trejų metų (faktorius A) kviečių derlingumas Y šešiose skirtingose Anglijos žemės ūkio stotyse (faktorius B). Kartu užregistruotas augalo aukštis varpų atsiradimo metu (kintamasis X_1) ir vidutinis augalų iš vieno kelmelio skaičius (kintamasis Z) [14].

Metai	Kintamasis	B_1	B_2	B_3	B_4	B_5	B_6
1933	Y	19,0	22,2	35,3	32,8	25,3	35,8
	X	25,6	25,4	30,8	33,0	28,5	28,0
	Z	14,9	13,3	4,6	14,7	12,8	7,5
1934	Y	32,4	32,2	43,7	35,7	28,3	35,2
	X	25,4	28,3	35,3	32,4	25,9	24,2
	Z	7,2	9,5	6,8	9,7	9,2	7,5
1935	Y	26,2	34,7	40,0	29,6	20,6	47,2
	X	27,2	34,4	32,5	27,5	23,7	32,9
	Z	18,6	22,2	10,0	17,6	14,4	7,9

a) Atlikite dvifaktorę dispersinę analizę neatsižvelgdami į kintamuosius X ir Z. b) Atlikite kovariacinę analizę atsižvelgdami į kovariantes X ir Z. c) 1934 metais stoties B_5 apylinkėje užregistruotas vidutinis augalų aukštis $X = 27$ ir augalų iš vieno krūmelio skaičius $Z = 10$. Gaukite taškinį numatomo derliaus įvertį.

4.9. Lyginami keturi vaistai (faktorius A), mažinantys kraujo spaudimą. Registruojamas kraujo spaudimas po gydymo Y. Kartu užregistruotas kraujo spaudimas prieš gydymą X [1].

	X	Y		X	Y
A_1	194	157	A_3	172	136
	162	136		196	182
	183	145		158	134
	180	153			
A_2	154	124	A_4	158	124
	184	123		165	124
	173	143		186	132
	170	136		182	133

a) Atlikite dispersinę analizę, neatsižvelgiant į kintamąjį X.
b) Atlikite vienfaktorę vieno kintamojo kovariacinę analizę.
c) Palyginkite a) ir b) gautus rezultatus.

4.10. Lentelėje pateikta 30 kiaulių svorio prieaugis Y priklausomai nuo fermos (faktorius A), maitinimo tipo (faktorius B), lyties (faktorius C). Eksperimentas suplanuotas pagal trifaktorės dispersinės analizės kryžminės klasifikacijos schemą su vienu stebėjimu langelyje. Daroma prielaida, kad svorio prieaugis gali priklausyti nuo tolydžios kovariantės – pradinio svorio X, kurio reikšmės taip pat pateiktos lentelėje [12].

A	B	C	X	Y	A	B	C	X	Y
A_1	B_1	C_1	48	9,94	A_3	B_3	C_1	33	7,63
A_1	B_2	C_1	48	10,00	A_3	B_1	C_1	35	9,32
A_1	B_3	C_1	48	9,75	A_3	B_2	C_1	41	9,34
A_1	B_3	C_2	48	9,11	A_3	B_2	C_2	46	8,43
A_1	B_2	C_2	39	8,51	A_3	B_3	C_2	42	8,90
A_1	B_1	C_2	38	9,52	A_3	B_1	C_2	41	9,32
A_2	B_2	C_1	32	9,24	A_4	B_3	C_1	50	10,37
A_2	B_3	C_1	28	8,66	A_4	B_1	C_2	48	10,56
A_2	B_1	C_1	32	9,48	A_4	B_2	C_1	46	9,68
A_2	B_3	C_2	37	8,50	A_4	B_1	C_1	46	10,98
A_2	B_1	C_2	35	8,21	A_4	B_2	C_2	40	8,86
A_2	B_2	C_2	38	9,95	A_4	B_3	C_2	42	9,51
A_5	B_2	C_1	37	9,67	A_5	B_2	C_2	40	9,20
A_5	B_1	C_1	32	8,82	A_5	B_3	C_2	40	8,76
A_5	B_3	C_1	30	8,57	A_5	B_1	C_2	43	10,42

a) Tarę, kad nėra faktorių sąveikos, atlikite stebėjimų trifaktorę dispersinę analizę.
b) Atlikite trifaktorę analizę eliminuodami kintamojo X įtaką.
c) Palyginkite a) ir b) punktuose gautus rezultatus.

4.11. Pasirinkę naujus parametrus perkelkite **3.10** pratimo duomenis į regresinės analizės schemą.

4.12. Pasirinkę naujus parametrus perkelkite **3.16** pratimo duomenis į regresinės analizės schemą.

4.13. Pasirinkę naujus parametrus perkelkite **3.26** pratimo duomenis į regresinės analizės schemą.

4.14. Pasirinkę naujus parametrus perkelkite **3.31** pratimo duomenis į regresinės analizės schemą.

4.15. Pasirinkę naujus parametrus perkelkite **3.39** pratimo duomenis į regresinės analizės schemą.

4.16. Pasirinkę naujus parametrus perkelkite **3.45** pratimo duomenis į regresinės analizės schemą.

4.17. Pasirinkę naujus parametrus perkelkite **3.47** pratimo duomenis į regresinės analizės schemą.

4.18. Atliktas visas dvifaktorės dispersinės analizės planas 2^2. Pasirinkę naujus parametrus, perkelkite dispersinės analizės modelį į regresinės analizės schemą. Patikrinkite, kad gautojo modelio plano matrica turi ortogonalius stulpelius ir sutampa su 4.5.1 lentelėje pateikta plano matrica. Gaukite dispersinės analizės kvadratų sumų išraiškas regresijos koeficientų įvertiniais.

4.19. Atliktas visas trifaktorės dispersinės analizės planas 2^3. Pasirinkę naujus parametrus perkelkite dispersinės analizės modelį į regresinės analizės schemą. Patikrinkite, kad gautojo modelio plano matrica turi ortogonalius stulpelius ir sutampa su 4.5.2 lentelėje pateikta plano matrica. Gaukite dispersinės analizės kvadratų sumų išraiškas regresijos koeficientų įvertiniais.

4.20. Faktoriniame eksperimente 2^2 su trimis stebėjimais langelyje gauti rezultatai pateikti lentelėje eksperimentus žymint kodiniu pavidalu.

Kodas	(1)			a			b			ab		
Y	0;	2;	1	4;	6;	2	-1;	-3;	1	-1;	-3;	-7

Įvertinkite Y tiesinės regresijos parametrus kintamųjų Z_1 ir Z_2 atžvilgiu. Priėmę normalumo prielaidą, patikrinkite regresijos koeficientų lygybės 0 hipotezes.

4.21. Tiriamos galingumo sąnaudos Y pjaustant metalą keraminiu instrumentu priklausomai nuo instrumento tipo (kintamasis X_1), rėžiklio briaunelės kampo (kintamasis X_2) ir nuo pjovimo tipo (kintamasis X_3). Atliktas visas faktorinis eksperimentas 2^3. Rezultatai (sąlyginiais vienetais), eksperimentus žymint kodiniu pavidalu, pateikti lentelėje [7].

Kodas	(1)	a	b	ab	c	bc	ac	abc
Y	2	-5	15	13	-12	-2	-17	-7

a) Įvertinkite Y tiesinės regresijos parametrus kintamųjų Z_1, Z_2, Z_3 atžvilgiu. Priėmę normalumo prielaidą, patikrinkite šių parametrų lygybės 0 hipotezes. b) Papildomai įvertinkite regresijos koeficientus β_{ij} prie sandaugų Z_iZ_j, $i \neq j$. Patikrinkite šių koeficientų lygybės 0 hipotezes.

4.22. Lentelėje pateikti tam tikro cheminio eksperimento duomenys. Atliktas visas faktorinis eksperimentas 2^3 su dviem stebėjimais langelyje.

Kodas	(1)	a	b	ab	c	bc	ac	abc
Y	1595	1573	1835	1700	1745	1838	2184	1717
	1578	1592	1823	1815	1689	1614	1538	1806

Atlikite duomenų analizę.

4.23. Lentelėje pateikti tam tikro faktorinio eksperimento 2^4 su dviem stebėjimais langelyje.

Kodas	(1)	a	b	ab	c	bc	ac	abc
Y	1985	1595	1765	1835	1694	1806	2243	1614
	1592	2067	1700	1823	1712	1758	1745	1838
Kodas	d	ad	bd	abd	cd	bcd	acd	$abcd$
Y	2156	1578	1923	1863	2184	1957	1745	1917
	2032	1733	2007	1910	1921	1717	1818	1922

a) Atlikite duomenų analizę, tarę, kad regresijos koeficientai prie trijų ir keturių kovariančių sandaugų lygūs 0. b) Priėmę normalumo prielaidą, patikrinkite regresijos koeficientų lygybės 0 hipotezes. c) Raskite tolesnio nepriklausomo Y stebėjimo prognozės intervalą, jeigu žinoma,

kad jis bus atliktas taške $\boldsymbol{z} = (z_1, ..., z_4)^T$, kurio koordinatės tenkina sąlygą

$$\sum_i z_i^2 + \sum_{i \neq j} z_i^2 z_j^2 = \rho^2.$$

4.24. Atliekamas visas faktorinis eksperimentas 3^2, kai kiekviena kovariantė Z_i (jeigu reikia atlikus transformavimą) įgyja reikšmes $-1; 0; +1$, t.y. eksperimentas atliktas kvadrato viršūnėse, centre ir kraštinių vidurio taškuose. Patikrinkite, kad regresijos lygties

$$Y_j = \beta_0 + \beta_1 Z_{1j} + \beta_2 Z_{2j} + \beta_{11} U_{1j}^2 + \beta_{22} U_{2j}^2 + \beta_{12} U_{1j}^2 U_{2j}^2 +$$

$$+\beta_{211} Z_{2j} U_{1j}^2 + \beta_{122} Z_{1j} U_{2j}^2 + e_j, \quad U_{ij}^2 = Z_{ij}^2 - \bar{Z^2}_{i.}, \quad i = 1, 2; \quad j = 1, ..., 9.$$

plano matrica turi ortogonalius stulpelius. Raskite regresijos parametrų įvertinius ir jų dispersijas. Tardami, kad a.d. $\{e_j\}$ yra nepriklausomi ir normalieji $e_j \sim N(0, \sigma^2)$, raskite kriterijus regresijos koeficientų lygybės 0 hipotezėms tikrinti.

4.25 (**4.24 tęsinys.**) Norint įvertinti regresijos koeficientus prie kovariančių sandaugų ir kvadratų, eksperimentas 2^2 atliktas kvadrato viršūnėse papildomas stebėjimais keturiuose taškuose $(\pm a, 0), (0, \pm a)$. Norint įvertinti dispersiją, eksperimentą du kartus pakartojame eksperimento centre, t.y. taške $(0, 0)$. Parinkite a taip, kad **4.24** pratimo regresijos lygties plano matrica turėtų ortogonalius stulpelius.

4.26 (**4.25** tęsinys.) Apibendrinkite **4.25** pratimą 3 kovariančių atveju.

4.27. Lentelėje pateiktos jėgos Y, reikalingos nustumti gaminį nuo konvejerio juostos priklausomai nuo temperatūros (kovariantė X_1) ir nuo drėgnumo (kovariantė X_2). Eksperimentas atliktas pagal visą faktorinio eksperimento 3^2 planą su dviem stebėjimais langelyje. Pateikiamos lentelės pirmoje eilutėje yra transformuotos kovariantės X_1 reikšmės, o pirmajame stulpelyje – kovariantės X_2 reikšmės.

	-1	-1	0	0	+1	+1
-1	0,8;	2,8	1,5;	3,2	2,5;	4,2
0	1,0;	1,6	1,6;	1,8	1,8;	1,0
+1	2,0;	2,2	1,5;	0,8	2,5;	4,0

Įvertinkite **4.24** pratime pateiktos regresijos lygties parametrus. Priėmę normalumo prielaidą patikrinkite regresijos koeficientų lygybės 0 hipotezes.

4.28. Tarkime, kad vienu metu galime atlikti tik keturis **4.22** pratimo eksperimentus. Sudalinkite eksperimentus į dvi replikas su tarpblokiniu efektu sumaišydami a) visų trijų faktorių sąveiką; b) faktorių A_1 ir A_3 sąveiką.

4.29. Atlikite **4.23** pratimo duomenų analizę suskaidę eksperimentus į 4 blokus po 4 eksperimentus. Su tarpblokiniu efektu sumaišykite sąveikas $A_1 \times A_3 \times A_4$, $A_1 \times A_2 \times A_4$, $A_1 \times A_2$.

4.30. Atlikite viso faktorinio eksperimento 2^5 suskaidymą į du blokus po 16 stebėjimų su tarpblokiniu efektu sumaišydami a) sąveiką $A_1 \times A_2 \times A_3$; b) sąveiką $A_1 \times A_2 \times A_3$.

4.31. Atlikite viso faktorinio eksperimento 2^5 suskaidymą į 4 blokus po 8 stebėjimus su tarpblokiniu efektu sumaišydami įvairius sąveikų rinkinius.

4.32. Tarkime, kad **4.22** pratimo sąlygomis galima atlikti tik vieną pusinę repliką. Parinkę repliką atlikite duomenų analizę ir aptarkite rezultatus.

4.33. Tarkime, kad **4.23** pratimo sąlygomis galima atlikti tik vieną ketvirtinę repliką. Parinkę repliką atlikite duomenų analizę ir aptarkite rezultatus.

4.7. Atsakymai ir nurodymai

4.1. Pažymėkime $R_{0r} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})(X_{ij}^{(r)} - \bar{X}_{i.}^{(r)})$, $r = 1, 2$; $R_{rs} = \sum_i \sum_j (X_{ij}^{(r)} - \bar{X}_{i.}^{(r)})(X_{ij}^{(s)} - \bar{X}_{i.}^{(s)})$, $r, s = 1, 2$. Tada a) $\hat{\gamma}_1 = (R_{01}R_{22} - R_{02}R_{12})/\Delta$, $\hat{\gamma}_2 = (R_{02}R_{11} - R_{01}R_{12})/\Delta$, $\Delta = R_{11}R_{22} - R_{12}^2$; b) $\boldsymbol{V}\hat{\gamma}_1 = \sigma^2 I(J-1)R_{22}/\Delta$, $\boldsymbol{V}\hat{\gamma}_2 = \sigma^2 I(J-1)R_{11}/\Delta$, $\mathbf{Cov}\,(\hat{\gamma}_1, \hat{\gamma}_2) = -\sigma^2 I(J-1)R_{12}/\Delta$; $\hat{\gamma}_1$ ir $\hat{\gamma}_2$ nekoreliuoti, kai $R_{12} = 0$. **4.2.** a) Randame $R_{yy}(i,j) = \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$, $R_{xx}(i,j) = \sum_k (X_{ijk} - \bar{X}_{ij.})^2$, $R_{yx}(i,j) = \sum_k (Y_{ijk} - Y_{ij.})(X_{ijk} - \bar{X}_{ij.})$; $SS_E(i,j) = R_{yy}(i,j) - R_{yx}^2(i,j)/R_{xx}(i,j)$, $SS_E = \sum_i \sum_j SS_E(i,j) \sim \sigma^2 \chi^2_{IJ(K-2)}$. Kai hipotezė H teisinga, tai $\hat{\gamma} = R_{yx}/R_{xx}$, $SS_{EH} = R_{yy} - R_{yx}^2/R_{xx}$, $R_{yy} = \sum_i \sum_j R_{yy}(i,j)$, $R_{xx} = \sum_i \sum_j R_{xx}(i,j)$, $R_{yx} = \sum_i \sum_j R_{yx}(i,j)$; hipotezė H atmetama reikšmingumo lygmens α kriterijumi, kai $(SS_{EH} - SS_E)(IJ(K-2))/(SS_E(IJ-1)) > F_\alpha(IJ-1, IJ(K-2))$. b) Kai H teisinga, tai a. d. $\sqrt{R_{xx}}(\hat{\gamma} - \gamma)/\sqrt{SS_{EH}/(IJK - IJ - 1)} \sim S(IJK - IJ - 1)$. **4.3.** Randame $R_{yy} = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$, $R_{xx} = \sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{ij.})^2$, $R_{yx} = \sum_i \sum_j \sum_k (Y_{ijk} - Y_{ij.})(X_{ijk} - \bar{X}_{ij.})$; $\hat{\gamma} = R_{yx}/R_{xx}$, $SS_E = R_{yy} - \hat{\gamma} R_{yx} \sim \sigma^2 \chi^2_{IJK - IJ - 1}$. a) Hipotezė H atmetama reikšmingumo lygmens α kriterijumi, kai $\hat{\gamma} R_{xx}(IJK - IJ - 1)/SS_E > F_\alpha(1, IJK - IJ - 1)$. b) Randame $\tilde{R}_{yy} = R_{yy} + JK \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2$, $\tilde{R}_{xx} = R_{xx} + JK \sum_i (\bar{X}_{i..} - \bar{X}_{...})^2$, $\tilde{R}_{yx} = R_{yx} + JK \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})(\bar{X}_{i..} - \bar{X}_{...})$, $SS_{EH} = \tilde{R}_{yy} - \hat{\gamma} \tilde{R}_{yx}$; hipotezė H_A atmetama reikšmingumo lygmens α kriterijumi, kai $(SS_{EH} - SS_E)(IJK - IJ - 1)/((I-1)SS_E) > F_\alpha(I-1, IJK - IJ - 1)$. **4.4.** Remdamiesi kovariacine analize randame $SS_E = \sum_i [R_{yy}^{(i)} - (R_{yx}^{(i)})^2/R_{xx}]$, $R_{yy}^{(i)} = \sum_j (Y_{ij} - \bar{Y}_{i.})^2$, $R_{yx}^{(i)} = \sum_j (Y_{ij} - \bar{Y}_{i.})(X_j - \bar{X}_{.})$, $i = 1, 2$, $R_{xx} = \sum_j (X_j - \bar{X}_{.})^2$; $SS_{EH} = R_{yy}^{(1)} + R_{yy}^{(2)} - (R_{yx}^{(1)} + R_{yx}^{(2)})^2/(2R_{xx}$. Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai $F = (SS_{EH} - SS_E)2(J-1)/SS_E = [R_{yx}^{(1)} - R_{yx}^{(2)}]^2 2(J-1)/(2R_{xx} SS_E) > F_\alpha(1, 2(J-1))$. Tikrinant dviejų regresijos tiesių lygiagretumo hipotezę remiamės statistika $T = (\hat{\gamma}_1 - \hat{\gamma}_2)\sqrt{R_{xx}}/\sqrt{2SS_E/(2(J-1))}$. Nesunku patikrinti, kad $F = T^2$. **4.5. a)** Kadangi statistikos reikšmė $F_A = 14,6$, tai hipotezė atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu. **b)** Imant antro laipsnio polinomą gauname, kad kovariantė ir jos kvadratas yra reikšmingi, nes statistikų reikšmės atitinkamai yra 48,83 ir 53,75. Imant trečio laipsnio polinomą gauname, kad kovariantė, jos kvadratas ir trečias laipsnis yra reikšmingi, nes statistikų reikšmės atitinkamai yra 35,23, 17,89 ir 8,87. Kadangi statistikos reikšmė yra 28,78, tai hipotezė apie faktoriaus įtakos nebuvimą atmetama su gana aukštu reikšmingumo lygmeniu. **4.6. a)** Atlikus dispersinę analizę gauta, kad faktorius C ($F_C = 1,17$; atitinkama P reikšmė yra $0,2947$) ir faktorių sąveikos su faktoriumi C ($F_{AC} = 1,35$, $F_{BC} = 1,09$, $F_{CD} = 1,17$; atitinkamos P reikšmės yra $0,2878$; $0,3609$; $0,2947$) nereikšmingos. Atlikus dispersinę analizę nejtraukiant faktoriaus C ir sąveikų su faktoriumi C gautos tokios statistikų realizacijos: $F_A = 124,60$, $F_B = 728,98$, $F_D = 118,78$, $F_{AB} = 41,75$, todėl hipotezės atmetamos su gana aukštu reikšmingumo lygmeniu. Gauta, kad $F_{AD} = 0,87$, $F_{BD} = 3,09$, atitinkamos P reikšmės $0,4348$, $0,0657$. **b)** Atlikus analizę, kai eliminuota kovariantės B įtaka, gauta: faktorių A ir C, A ir D, C ir D sąveikos nereikšmingos (statistikų reikšmės yra $0,17$; $0,11$; $0,15$; atitinkamos P reikšmės $0,8469$; $0,8933$; $0,7059$). Faktorių A ir D įtakos nebuvimo hipotezės atmetamos su gana aukštu reikšmingumo lygmeniu. Faktorius C nereikšmingas (statistikos reikšmė $0,15$, atitinkama P reikšmė $0,7059$). **4.7. a)** Kadangi $F_A = 43,67$, tai hipotezė atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu. **b)** $\hat{\beta}_0 = -127,17$, $\hat{\beta}_1 = 90,827$. Hipotezė $H: \beta_1 = 0$ atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu, nes statistikos reikšmė yra $16,27$. **c)** Statistikos (4.2.18) reikšmė $3,43$; regresijos koeficientų lygybės hipotezės neatmetame kriterijumi

su reikšmingumo lygmeniu $\alpha < 0,0134$. Tikrinant hipotezę H_A, eliminavus kintamojo X įtaką, statistikos (4.2.6) reikšmė yra $3,65$; hipotezė atmetama kriterijumi su reikšmingumo lygmeniu $\alpha > 0,0098$. **4.8. a)** $F_A = 2,72$, $F_B = 5,50$; P reikšmės $0,1139$, $0,0108$. **b)** Eliminavus kintamųjų X ir Z įtaką, gaunama $F_A = 13,96$, $F_B = 5,96$; P reikšmės $0,0025$, $0,0137$. Hipotezė, kad kovariantė X nereikšminga (statistikos reikšmė $57,58$) atmetama kriterijumi su gana aukštu reikšmingumo lygmeniu. Hipotezė, kad kovariantė Z nereikšminga (statistikos reikšmė $8,30$) atmetama kriterijumi su reikšmingumo lygmeniu $\alpha > 0,0205$. **c)** $29,6$. **4.9. a)** $F_A = 2,46$; P reikšmė $0,1173$. **b)** Statistikos (4.2.18) reikšmė $2,56$; regresijos koeficientų lygybės hipotezės neatmetame kriterijumi su reikšmingumo lygmeniu $\alpha < 0,1376$. Tikrinant hipotezę H_A, eliminavus kintamojo X įtaką, statistikos (4.2.6) reikšmė yra $2,88$; hipotezė atmetama kriterijumi su reikšmingumo lygmeniu $\alpha > 0,0896$. **4.10. a)** $F_A = 3,08$, $F_B = 2,88$, $F_C = 1,13$; P reikšmės $0,0373$, $0,0773$, $0,3001$. **b)** $F_A = 2,58$, $F_B = 5,08$, $F_C = 5,63$; P reikšmės $0,0667$, $0,0159$, $0,0273$. Hipotezė, kad kovariantė X nereikšminga (statistikos reikšmė $16,53$) atmetama kriterijumi su reikšmingumo lygmeniu $\alpha > 0,0006$. **4.18.** $SS_{A_1} = 4\hat{\beta}_1^2$, $SS_{A_2} = 4\hat{\beta}_2^2$, $SS_E = \sum_i Y_i^2 - 4(\hat{\beta}_1^2 + \hat{\beta}_2^2)$. **4.19.** $SS_{A_i} = 8\hat{\beta}_i^2$, $i = 1,2,3$; $SS_E = \sum_i Y_i^2 - 8\sum_i \hat{\beta}_i^2$. **4.20.** $\hat{\beta}_0 = 1/12; \hat{\beta}_1 = -1/12$, $\hat{\beta}_2 = -29/12$; $\hat{\sigma}^2 = s^2 = 6,75$. Statistikos, kurios esant teisingoms hipotezėms $H_i : \beta_i = 0$ turi Fišerio skirstinius su 1 ir 9 laisvės laipsniais įgijo reikšmes 0,012; 0,012, 10,383; atmesti parametrų β_0 ir β_1 lygybės 0 hipotezes nėra pagrindo; hipotezė $H_2 : \beta_2 = 0$ atmetama, kai kriterijaus reikšmingumo lygmuo viršija 0,01. **4.21.** a) $\hat{\beta}_0 = -13/8$, $\hat{\beta}_1 = -19/8$, $\hat{\beta}_2 = 51/8$, $\hat{\beta}_3 = -63/8$, $\hat{\sigma} = s = 2,318$. Statistikos, kurios esant teisingoms hipotezėms $H_i : \beta_i = 0$ turi Fišerio skirstinius su 1 ir 4 laisvės laipsniais, įgijo reikšmes 3,930; 8,395; 60,488; 92,302; atitinkamos P-reikšmės yra 0,118; 0,044; 0,0015; 0,0007. b) $\hat{\beta}_{12} = 5/8$, $\hat{\beta}_{13} = -1/8$, $\hat{\beta}_{23} = -11/8$, $\hat{\sigma} = s = 1,768$; atmesti hipotezes nėra pagrindo. **4.22.** $\hat{\beta}_0 = 1727,63$, $\hat{\beta}_1 = 13$, $\hat{\beta}_2 = 40,875$, $\hat{\beta}_3 = 38,75$, $\hat{\sigma} = s = 167,138$. Hipotezės $H_i : \beta_i = 0, i = 1,2,3$, neatmetamos. $\hat{\beta}_{12} = -22$, $\hat{\beta}_{13} = 31,875$, $\hat{\beta}_{23} = -63,5$, $\hat{\sigma} = s = 165,564$; atmesti hipotezes $H_{ij} : \beta_{ij} = 0, i \neq j = 1,2,3$, nėra pagrindo. **4.23.** a) Parametrų įverčiai: $\hat{\beta}_0 = 1848,59$, $\hat{\beta}_1 = -20,7188$, $\hat{\beta}_2 = -13,9063$, $\hat{\beta}_3 = 0,8438$, $\hat{\beta}_4 = 50,3438$; $\hat{\beta}_{12} = 26,2813$, $\hat{\beta}_{13} = 26,5313$, $\hat{\beta}_{14} = -67,4688$, $\hat{\beta}_{23} = -19,4063$, $\hat{\beta}_{24} = 16,9688$, $\hat{\beta}_{34} = -2,1563$, $\hat{\sigma} = 169,803$. b) Hipotezė $H_0 : \beta_0 = 0$ atmetama; hipotezė $H_{14} : \beta_{14} = 0$ atmetama, kai kriterijaus reikšmingumo lygmuo viršija 0,0355; kitos hipotezės neatmetamos. c) Prognozės intervalas su pasikliovimo lygmeniu $Q = 1 - \alpha$ yra $\hat{Y} \pm \hat{\sigma} t_{\alpha/2}(21)\sqrt{(33 + \rho^2)/32}$. **4.25.** $a = \sqrt{\sqrt{10} - 2}$. **4.26.** $a = 2\sqrt{\sqrt{2} - 1}$. **4.27.** Parametrų įverčiai: $\hat{\beta}_0 = 2,0444$, $\hat{\beta}_1 = 0,4667$, $\hat{\beta}_2 = -0,1667$, $\hat{\beta}_{11} = 0,4667$, $\hat{\beta}_{22} = 0,8667$, $\hat{\beta}_{12} = 0,1361$, $\hat{\beta}_{211} = 0,650$, $\hat{\beta}_{122} = 0,625$, $\hat{\sigma}^2 = s^2 = 0,8066$. Hipotezė $H_0 : \beta_0 = 0$ atmetama: P-reikšmė $pv = 2 \times 10^{-6}$. Hipotezei $H_{22} : \beta_{22} = 0$ P-reikšmė $pv = 0,082$; atmesti kitas hipotezes nėra pagrindo. **4.28.** Abiem atvejais pirmoji replika ((1), ab, ac, bc).

5 skyrius

Apibendrintieji tiesiniai modeliai

Pirmesniuose skyreliuose tariama, kad imties $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ narys Y_i turi tokią struktūrą

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_m x_{mi} + e_i,$$

čia $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ yra nežinomų parametrų vektorius, $\boldsymbol{x}_i = (1, x_{1i}, ..., x_{mi})^T$ – kovariančių vektoriaus $\boldsymbol{x} = (x_0, x_1, ..., x_m)^T$, $x_0 = 1$ žinoma i-oji reikšmė, $\boldsymbol{e} = (e_1, e_2, ..., e_n)^T$ – paklaidų vektorius. Tariama, kad paklaidų vektoriaus koordinatės nepriklausomi normalieji a. d. $e_i \sim N(0, \sigma^2)$ (kartu $Y_i \sim N(\boldsymbol{\beta}^T \boldsymbol{x}_i, \sigma^2)$).

Imties $\boldsymbol{Y}$ narių normalumo prielaida ne visada priimtina. Pavyzdžiui, Y_i gali būti diskretieji a. d. ir jų skirstiniais reikėtų imti kuriuos nors diskrečiuosius modelius (Puasono, binominį ir kt.). Išgyvenamumo analizėje, kai Y_i reiškia i-ojo individo gyvenimo trukmę, a. d. Y_i igyja neneigiamas reikšmes ir jų skirstiniais reikėtų imti tikimybinius modelius intervale $(0, \infty)$ (pavyzdžiui, eksponentinį, gama, Veibulo ir kt.).

Apibendrintųjų tiesinių modelių aptarimą galima rasti [15], [6]. Jų taikomuosius aspektus ir duomenų analizę populiariais matematinės statistikos TPP (žr. [4], III dalis; [5]).

Šiame skyriuje trumpai aptarsime apibendrintųjų tiesinių modelių sudarymo ir analizės metodus apsiribodami vienparametriais eksponentinio tipo skirstiniais.

5.1. Vienparametrių eksponentinio tipo skirstinių tiesiniai modeliai

Eksponentinio tipo skirstinių šeimų apibrėžimas ir jų savybės pateiktos I dalies 4.3 skyrelyje. Tariama, kad a. d. Y skirstinys priklauso vienparametrei kanoninio pavidalo eksponentinio tipo skirstinių šeimai, jeigu jo tankis σ-baigtinio

mato $\boldsymbol{\mu}$ atžvilgiu yra toks:

$$f(x|\theta) = h(x)e^{T(x)\theta - B(\theta)}, \quad \theta \in \Theta \subset \mathbf{R}, \tag{5.1.1}$$

čia $h(x)$ nepriklauso nuo θ, T(Y) – pakankamoji statistika. Pirmosios dalies 4.3.2 teoremoje įrodyta, kad

$$\mathbf{E}_\theta T(Y) = \dot{B}(\theta), \quad \boldsymbol{V}_\theta T(Y) = \ddot{B}(\theta). \tag{5.1.2}$$

Tegu imties $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ nariai yra n. a. d., turintys (5.1.1) skirstinį, kai parametrai yra $\theta_1, ..., \theta_n$. Tikėtinumo funkcija

$$L(\theta_1, ..., \theta_n) = \prod_{i=1}^{n} \{h(Y_i)e^{T(Y_i)\theta_i - B(\theta_i)}\}. \tag{5.1.3}$$

Tai *prisotintas* modelis, kai nežinomų parametrų $\theta_1, ..., \theta_n$ skaičius lygus imties didumui n.

Apibendrintasis tiesinis modelis. Tarkime, kad kartu su a. d. Y_i gaunama kovariančių vektoriaus $\boldsymbol{x} = (x_0, x_1, ..., x_m)^T$, $x_0 = 1$ reikšmė $\boldsymbol{x}_i = (x_{0i}, x_{1i}, ..., x_{mi})^T$, $i = 1, ..., n$. Parametrų skaičius sumažinamas nuo n iki $m+1 < n$ imant kovariančių vektoriaus tiesinius darinius $\boldsymbol{\beta}^T\boldsymbol{x}_i = \beta_0 x_{0i} + \beta_1 x_{1i} + ... + \beta_m x_{mi}$; čia $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ naujas nežinomų parametrų vektorius. Nauji parametrai įvedami prilyginant sąlyginio vidurkio $\mu_i = \mathbf{E}(T(Y)|\boldsymbol{x}_i)$ funkciją tiesiniam dariniui $\boldsymbol{\beta}^T\boldsymbol{x}_i$:

$$g(\mu_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i. \tag{5.1.4}$$

Funkcija g vadinama *jungties funkcija.* Tariama, kad funkcija g turi diferencijuojamą atvirkštinę

$$g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i) = \mu_i. \tag{5.1.5}$$

Jungties funkcija vadinama *kanonine*, jeigu $\mu_i = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i) = \dot{B}(\boldsymbol{\beta}^T\boldsymbol{x}_i)$, t. y. parametras $\theta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i$. Tačiau kartais tenka funkciją g parinkti kitokio pavidalo. Natūralu ją parinkti taip, kad $g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ priklausytų vidurkio $\mu = \dot{B}(\theta), \theta \in \Theta$ kitimo sričiai, kad ir kokia būtų tiesinio darinio $\boldsymbol{\beta}^T\boldsymbol{x}$ reikšmė $\boldsymbol{\beta}^T\boldsymbol{x}_i$.

Parametrų įvertiniai. Imties

$$(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$$

tikėtinumo funkcija

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \{h(Y_i)e^{T(Y_i)\theta_i - B(\theta_i)}\}, \tag{5.1.6}$$

čia remiantis (5.1.2) ir (5.1.5)

$$\theta_i = \dot{B}^{-1}(g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i)).$$

Kanoninės funkcijos atveju $\theta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i$.

Logtikėtinumo funkcija

$$l = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [T(Y_i)\theta_i - B(\theta_i) + \ln(h(Y_i))]. \tag{5.1.7}$$

Diferencijuojant pagal β_j gaunama informančių vektoriaus j-oji koordinatė

$$\dot{l}_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} x_{ji} \frac{T(Y_i) - \dot{B}(\theta_i)}{\dot{g}(\mu_i)\ddot{B}(\theta_i)}, \quad j = 0, 1, ..., m. \tag{5.1.8}$$

Kai jungtis kanoninė, tai $\dot{g}(\mu_i)\ddot{B}(\theta_i) = 1$, t. y. (5.1.8) lygybėje vardiklis lygus 1.

Prilyginę informančių vektoriaus koordinates nuliui, gauname lygčių sistemą

$$\dot{l}_j = 0, \quad j = 0, 1, ..., m, \tag{5.1.9}$$

parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ DT įvertiniui $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m)^T$ rasti. Turint $\hat{\boldsymbol{\beta}}$, galima įvertinti sąlyginį vidurkį $\mu(\boldsymbol{x}) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x})$ arba kitas tiesinio darinio $\boldsymbol{\beta}^T\boldsymbol{x}$ funkcijas.

Randame Fišerio informacinės matricos $\boldsymbol{I}(\boldsymbol{\beta}) = [I_{js}]_{(m+1)\times(m+1)}$ elementus

$$I_{js}(\boldsymbol{\beta}) = \mathbf{E}(\dot{l}_j\dot{l}_s) = \sum_{i=1}^{n} x_{ji}x_{si} \frac{\mathbf{E}(T(Y_i) - \dot{B}(\theta_i))^2}{[\dot{g}(\mu_i)]^2[\ddot{B}(\theta_i)]^2}$$

$$= \sum_{i=1}^{n} x_{ji}x_{si} \frac{1}{[\dot{g}(\mu_i)]^2\ddot{B}(\theta_i)} = \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\mathbf{X}, \tag{5.1.10}$$

čia

$$\mathbf{X} = \begin{pmatrix} x_{01} & ... & x_{m1} \\ ... & ... & ... \\ x_{0n} & ... & x_{mn} \end{pmatrix},$$

o $\boldsymbol{W}(\boldsymbol{\beta})$ – diagonalioji matrica su diagonaliniais elementais

$$\frac{1}{[\dot{g}(\mu_1)]^2\ddot{B}(\theta_1)}, ..., \frac{1}{[\dot{g}(\mu_n)]^2\ddot{B}(\theta_n)}.$$

Kai jungtis kanoninė, tai matricos $\boldsymbol{W}(\boldsymbol{\beta})$ diagonaliniai elementai yra $\ddot{B}(\boldsymbol{\beta}^T\boldsymbol{x}_1), ...,$ $\ddot{B}(\boldsymbol{\beta}^T\boldsymbol{x}_n)$.

Jeigu įvykdytos I dalies 4.5.4 teoremos sąlygos, tai DT įvertinys yra pagrįstasis ir asimptotiškai ($n \to \infty$) normalusis

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{Z} \sim N_{m+1}(\mathbf{0}, \boldsymbol{i}^{-1}(\boldsymbol{\beta})), \tag{5.1.11}$$

čia $\boldsymbol{i}^{-1}(\boldsymbol{\beta}) = \lim_{n\to\infty} \boldsymbol{I}(\boldsymbol{\beta})/n$.

Remiantis įvertinio $\hat{\boldsymbol{\beta}}$ asimptotiniu normalumu, naudojant Fišerio informacinės matricos įvertinį $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$, galima rasti parametrų $\beta_0, ..., \beta_m$ ar jų tiesinių darinių asimptotinius pasikliovimo intervalus.

Hipotezių tikrinimas. Kovariantės x_j reikšmingumo hipotezės $H_j : \beta_j = 0$ asimptotinį tikrinimo kriterijų sudarome remdamiesi įvertinio $\hat{\beta}_j$ asimptotiniu normalumu. Jei hipotezė H_j teisinga, tai remiantis (5.1.11) statistika

$$\sqrt{n}\hat{\beta}_j / \sqrt{c_{jj}(\hat{\boldsymbol{\beta}})} \xrightarrow{d} Z \sim N(0,\ 1), \tag{5.1.12}$$

čia $c_{jj}(\hat{\boldsymbol{\beta}})$ yra matricos $\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}}) = [c_{js}(\hat{\boldsymbol{\beta}})]_{(m+1)\times(m+1)}$ diagonalinis elementas.

Hipotezė atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$\sqrt{n}|\hat{\beta}_j| / \sqrt{c_{jj}(\hat{\boldsymbol{\beta}})} > z_{\alpha/2}, \quad j = 0, 1, ..., m.$$

Tikrindami visų kovariančių reikšmingumo hipotezę

$$H_{1,...,m} : \beta_1 = ... = \beta_m = 0,$$

naudojame tikėtinumų santykio kriterijų. Tikėtinumų santykis

$$\Lambda = \frac{\max_{\boldsymbol{\beta}:\beta_1=...=\beta_m=0} L(\boldsymbol{\beta})}{\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})} = \frac{\exp\{\sum_i [T(Y_i)\hat{\theta} - B(\hat{\theta})]\}}{\exp\{\sum_i [T(Y_i)\hat{\theta}_i - B(\hat{\theta}_i)]\}}, \tag{5.1.13}$$

čia $\hat{\theta}_i = \dot{B}^{-1}(g^{-1}(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i))$ yra DT įvertiniai, kai $\hat{\boldsymbol{\beta}}$ gaunamas iš lygčių sistemos (5.1.9); $\hat{\theta}$ – DT įvertinys, kai modelis nusakomas vieninteliu parametru θ:

$$\hat{\theta} = \dot{B}^{-1}(\bar{T}), \quad \bar{T} = \frac{1}{n}\sum_{i=1}^{n} T(Y_i). \tag{5.1.14}$$

Jeigu įvykdytos I dalies 4.5.4 teoremos sąlygos, tai, remiantis I dalies 4.5.4 skyreliu, gauname, kad kai hipotezė $H_{1,...,m}$ teisinga, asimptotiškai ($n \to \infty$)

$$D_R = -2\ln\Lambda = 2\sum_{i=1}^{n}[T(Y_i)(\hat{\theta}_i - \hat{\theta}) - (B(\hat{\theta}_i) - B(\hat{\theta}))] \xrightarrow{d} \chi^2_m. \tag{5.1.15}$$

Hipotezė atmetama asimptotiniu reikšmingumo lygmens α kriterijumi ,kai

$$D_R > \chi^2_\alpha(m). \tag{5.1.16}$$

Tikrindami kovariančių $x_{i_1}, ..., x_{i_l}$ įtakos nebuvimo hipotezę $H_{i_1,...,i_l} : \beta_{i_1} = ... = \beta_{i_l} = 0$ vėl naudojame tikėtinumų santykio kriterijų. Pažymėkime $D_R^{(m)}$ ir $D_R^{(m-l)}$ statistiką D_R atitinkamai modeliui su visais parametrais $\beta_0, ..., \beta_m$ ir modeliui be parametrų $\beta_{i_1}, ..., \beta_{i_l}$. Kai $H_{i_1,...,i_l}$ teisinga, statistikos $D_R^{(m)} - D_R^{(m-l)}$ skirstinys aproksimuojamas χ^2 skirstiniu su $m - (m - l) = l$ laisvės laipsnių. Hipotezė $H_{i_1,...,i_l}$ atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$D_R^{(m)} - D_R^{(m-l)} > \chi^2_\alpha(l). \tag{5.1.17}$$

Determinacijos koeficiento analogas. Aptarsime, kaip galima apibrėžti determinacijos koeficiento R^2 ir kvadratų sumų (pilnosios SS_T, liekamųjų paklaidų SS_E ir regresijos SS_R), nagrinėtų 3.3.8 skyrelyje, analogus. Turime tris vienas į kitą įdėtus modelius. Pirmas plačiausias (prisotintas) modelis (5.1.3), kai nežinomų parametrų $\theta_1, ..., \theta_n$ skaičius lygus imties didumui n. Antras, siauresnis už pirmą regresijos modelis, kai nežinomų parametrų $\beta_0, ..., \beta_m$ skaičius $m+1 < n$. Pagaliau trečias dar siauresnis modelis, kai imties $Y_1, ..., Y_n$ nariai vienodai pasiskirstę ir priklauso nuo vienintelio parametro θ.

Sudarykime tikėtinumų santykį palygindami pirmą ir trečią modelius

$$\Lambda_{13} = \frac{\max_{\theta_1=...=\theta_n=\theta} L(\theta_1, ..., \theta_n)}{\max_{\theta_1,...,\theta_n} L(\theta_1, ..., \theta_n)} = \frac{\exp\{\sum_i [T(Y_i)\hat{\theta} - B(\hat{\theta})]\}}{\exp\{\sum_i [T(Y_i)\tilde{\theta}_i - B(\tilde{\theta}_i)]\}}, \qquad (5.1.18)$$

čia $\hat{\theta}$ yra DT įvertinys (5.1.14); $\tilde{\theta}_1, ..., \tilde{\theta}_n$ yra DT įvertiniai prisotintame modelyje (5.1.3):

$$\tilde{\theta}_i = \dot{B}^{-1}(T(Y_i)), \quad i = 1, 2, ..., n.$$

Jeigu trečias modelis teisingas ir n didelis, tai statistikos

$$D_T = -2 \ln \Lambda_{13} = 2 \sum_{i=1}^{n} [T(Y_i)(\tilde{\theta}_i - \hat{\theta}) - (B(\tilde{\theta}_i) - B(\hat{\theta}))]$$

skirstinys artimas χ^2 skirstiniui su $n-1$ laisvės laipsniu. Statistika D_T yra tiesinės regresijos pilnosios kvadratų sumos SS_T analogas.

Tikėtinumų santykis lyginant antrą ir trečią modelius sudarytas (5.1.13). Gautoji statistika D_R (5.1.14) yra tiesinės regresijos kvadratų sumos SS_R analogas.

Sudarę tikėtinumų santykį pirmam ir antram modeliui palyginti, gausime tiesinės regresijos kvadratų sumos SS_E analogą

$$D_E = D_T - D_R.$$

Determinacijos koeficiento analogą R^2 apibrėžiame analogiškai kaip skyrelyje 3.3.8

$$R^2 = \frac{D_R}{D_T} = 1 - \frac{D_E}{D_T}. \qquad (5.1.19)$$

Jeigu regresijos modelis teisingas, tai $D_R \approx D_T$ ir R^2 įgyja reikšmes arti vieneto. Jeigu regresijos nėra, tai $D_E \approx D_T$ ir R^2 įgyja mažas reikšmes. Taigi R^2 yra regresijos modelio tinkamumo matas. Kitus regresijos kokybės matus galima rasti, pavyzdžiui, [13].

5.2. Apibendrintųjų tiesinių modelių pavyzdžiai

5.2.1. Puasoninė regresija

Atsitiktinio dydžio $Y \sim \mathcal{P}(\lambda)$ skirstinys nusakomas tikimybėmis

$$P\{Y = k|\lambda\} = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{1}{k!} e^{k \ln \lambda - \lambda}, \quad k = 0, 1, 2, ...$$

Reparametrizavę, t. y. imdami $\theta = \ln\lambda$, matome, kad Puasono skirstinys priklauso kanoninio pavidalo vienparametrių ekponentinio tipo skirstinių šeimai. Tankis (skaičiuojančiojo mato atžvilgiu) yra

$$f(y|\theta) = \frac{1}{y!}e^{y\theta - e^{\theta}}, \quad \theta = \ln\lambda, \quad y = 0, 1, 2, ...$$

Nagrinėsime bendresnį modelį, kai $Y \sim \mathcal{P}(\lambda t)$, o $t > 0$ žinomas. Pavyzdžiui, jeigu Y yra puasoninio srauto įvykių ilgio t intervale skaičius, kai srauto intensyvumas pastovus ir lygus λ (įvykių/laiko vienete), tai a. d. $Y \sim \mathcal{P}(\lambda t)$.

Imties $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ su nepriklausomomis koordinatėmis, kai $Y_i \sim \mathcal{P}(\lambda_i t_i)$, $\lambda_i > 0$ – nežinomi parametrai, o $t_i > 0$ žinomi, tikėtinumo funkcija

$$L(\theta_1, ..., \theta_n) = \prod_{i=1}^{n}(h(Y_i)e^{Y_i\theta_i - B(\theta_i)}), \tag{5.2.1}$$

$$h(Y_i) = \frac{t_i^{Y_i}}{Y_i!}, \quad \theta_i = \ln\lambda_i, \quad B(\theta_i) = t_i e^{\theta_i}.$$

Tai prisotintas modelis, kai nežinomų $\theta_1, ..., \theta_n$ parametrų skaičius lygus imties didumui n.

Jeigu kartu su Y_i gaunama kovariančių vektoriaus $\boldsymbol{x} = (x_0, x_1, ..., x_m), x_0 = 1$ reikšmė $\boldsymbol{x}_i = (x_{0i}, x_{1i}, ..., x_{mi})^T$, tai parametrų skaičius sumažinamas iki $m+1 < n$ imant tiesinius darinius $\boldsymbol{\beta}^T\boldsymbol{x}_i = \beta_0 x_{0i} + ... + \beta_m x_{mi}$. Natūralu parinkti kanoninę jungties funkciją, kai θ_i keičiamas tiesiniu dariniu $\boldsymbol{\beta}^T\boldsymbol{x}_i$. Tada

$$\mu_i = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i) = \dot{B}(\boldsymbol{\beta}^T\boldsymbol{x}_i) = t_i e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}, \tag{5.2.2}$$

$$g(\mu_i) = \ln\mu_i - \ln t_i, \quad \theta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i.$$

Lygčių sistema (5.1.9) parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ DT įvertiniui rasti turi tokį pavidalą:

$$\dot{l}_j = \sum_{i=1}^{n} x_{ji}[Y_i - t_i e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}] = 0, \quad j = 0, 1, ..., m. \tag{5.2.3}$$

Fišerio informacinė matrica (5.1.10) yra

$$\boldsymbol{I}(\boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\mathbf{X},$$

kai $\boldsymbol{W}(\boldsymbol{\beta})$ yra diagonalioji matrica su diagonaliniais elementais

$$\mu_i = \ddot{B}(\boldsymbol{\beta}^T\boldsymbol{x}_i) = t_i e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}, \quad i = 1, 2, ..., n.$$

Tikrinant hipotezę $H_{1,...,m} : \beta_1 = ... = \beta_m = 0$ tikėtinumų santykio kriterijumi (5.1.16), parametro θ įvertinys modelyje su vienu parametru yra

$$\hat{\theta} = \ln(\sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} t_i) = \ln\hat{\lambda}.$$

5.2.1 pavyzdys. Tiriama bakterijų augimo priklausomybė nuo aplinkos sąlygų. Paruošiamos Petri lėkštelės su įvairios sudėties maitinamosiomis terpėmis, kuriose pasėjamos bakterijų kultūros. Prabėgus laikui t užfiksuojamas bakterijų kolonijų skaičius Y. Natūralu i-ojo eksperimento rezultatą Y_i aprašyti Puasono skirstiniu su parametru λ_i, kuris priklauso nuo kovariančių vektoriaus $\boldsymbol{x}_i$, apibūdinančio maitinamosios terpės sudėtį, oro temperatūrą, drėgnumą ir kt. Šio pavyzdžio imčiai $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$ analizuoti taikytina puasoninė regresija.

5.2.2. Gama regresija

Atsitiktinio dydžio $Y \sim G(\theta, \eta)$, turinčio gama skirstinį su nežinomu parametru $\theta > 0$ ir žinomu $\eta > 0$, skirstinys priklauso kanoninio pavidalo vienparametrių ekponentinio tipo skirstinių šeimai. Tankis (Lebego mato atžvilgiu) yra

$$f(y|\theta) = \frac{\theta^\eta}{\Gamma(\eta)} y^{\eta-1} e^{-y\theta} = h(y) e^{T(y)\theta - B(\theta)},$$

$$h(y) = \frac{y^{\eta-1}}{\Gamma(\eta)}, \quad T(y) = -y, \quad B(\theta) = -\eta \ln \theta.$$

Tarkime, kad imties $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ su nepriklausomomis koordinatėmis narys Y_i turi gama skirstinį $G(\theta_i, \eta_i)$, kai $\theta_i > 0$ – nežinomi parametrai, o $\eta_i > 0$ žinomi. Tikėtinumo funkcija

$$L(\theta_1, ..., \theta_n) = \prod_{i=1}^{n} (h(Y_i)) e^{\sum_i (T(Y_i)\theta_i - B(\theta_i))}, \tag{5.2.4}$$

$$h(Y_i) = \frac{Y_i^{\eta_i - 1}}{\Gamma(\eta_i)}, \quad B(\theta_i) = -\eta_i \ln \theta_i.$$

Tai prisotintas modelis, kai nežinomų $\theta_1, ..., \theta_n$ parametrų skaičius lygus imties didumui n.

Jeigu kartu su Y_i gaunama kovariančių vektoriaus $\boldsymbol{x} = (x_0, x_1, ..., x_m), x_0 = 1$ reikšmė $\boldsymbol{x}_i = (x_{0i}, x_{1i}, ..., x_{mi})^T$, tai parametrų skaičius sumažinamas iki $m+1 < n$ imant tiesinius darinius $\boldsymbol{\beta}^T \boldsymbol{x}_i = \beta_0 x_{0i} + ... + \beta_m x_{mi}$. Parinkti kanoninę jungties funkciją, kai θ_i keičiamas tiesiniu dariniu $\boldsymbol{\beta}^T \boldsymbol{x}_i$, negalima, nes $\ln \theta_i$ neapibrėžtas, kai $\boldsymbol{\beta}^T \boldsymbol{x}_i < 0$. Kadangi vidurkis $\mathbf{E} Y_i = \eta_i / \theta_i > 0$, tai naujus parametrus galima įtraukti keičiant $1/\theta_i$ neneigiamu reiškiniu $e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}$. Tada gauname

$$\mu_i = g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}_i) = -\eta_i e^{\boldsymbol{\beta}^T \boldsymbol{x}_i},$$

$$g(\mu_i) = \ln(-\mu_i / \eta_i), \quad \theta_i = e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i}.$$

Lygčių sistema (5.1.9) parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ DT įvertiniui rasti turi tokį pavidalą:

$$\dot{l}_j = \sum_{i=1}^{n} (-x_{ji} [T(Y_i) e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i} + \eta_i]) = 0, \quad j = 0, 1, ..., m. \tag{5.2.5}$$

Fišerio informacinė matrica (5.1.10) yra

$$\boldsymbol{I}(\boldsymbol{\beta}) = \mathbf{X}^T \boldsymbol{W} \mathbf{X},$$

kai diagonalioji matrica $\boldsymbol{W}$ nuo parametro $\boldsymbol{\beta}$ nepriklauso. Jos diagonaliniai elementai yra $\eta_1, ..., \eta_n$.

Tikrinant hipotezę $H_{1,...,m} : \beta_1 = ... = \beta_m = 0$ tikėtinumų santykio kriterijumi (5.1.16), parametro θ įvertinys modelyje su vienu parametru yra

$$\hat{\theta} = \sum_{i=1}^{n} \eta_i / \sum_{i=1}^{n} Y_i = \bar{\eta}/\bar{Y}.$$

5.2.2 pavyzdys. Keltas kelia per upę, kai ant jo užvažiuoja 5 automobiliai. Tarkime, kad automobilių srautas yra puasoninis ir, kol sukomplektuojamas i-asis reisas, srauto intensyvumas θ_i yra pastovus. Tada laikas Y_i nuo $(i-1)$-ojo iki i-ojo kelto išvykimo turi gama skirstinį $G(\theta_i, 5)$ su parametrais $\theta_i > 0$ ir $\eta_i = 5$. Tiriant Y skirstinio priklausomybę nuo kovariančių: paros laikas, savaitės diena, kalendorinis laikas, oro sąlygos ir kt., taikytina gama regresija.

5.2.1 pastaba. Jeigu η_i yra sveikasis skaičius, tai $Y_i \sim G(\theta_i, \eta_i)$ (Erlango skirstinys) yra suma $Y_i = Y_{i1} + ... + Y_{i\eta_i}$, vienodai pasiskirsčiusių n. a. d. $Y_{ij} \sim \mathcal{E}(\theta_i)$, turinčių eksponentinį skirstinį su parametru θ_i. Todėl šiuo atveju gama regresija faktiškai yra eksponentinio skirstinio regresija remiantis imtimi $(Y_{ij}, \boldsymbol{x}_i)$, $j = 1, ..., \eta_i, i = 1, ..., n$, kai imties didumas $N = \eta_1 + ... + \eta_n$. Kovariantės $\boldsymbol{x}$ reikšmė $\boldsymbol{x}_i$ kartojasi η_i kartų. Nesunku patikrinti, kad parametrams $\beta_0, ..., \beta_m$ vertinti pakanka žinoti sumas $Y_1, ..., Y_n$.

5.2.3. Neigiamoji binominė regresija

Atsitiktinio dydžio $Y \sim B^-(k, p)$ su nežinomu parametru $0 < p < 1$ ir žinomu parametru k skirstinys nusakomas tikimybėmis

$$P\{Y = m|p\} = \frac{\Gamma(k+m)}{\Gamma(k)m!}(1-p)^m p^k = \frac{\Gamma(k+m)}{\Gamma(k)m!} e^{m\ln(1-p)+k\ln p}, \quad m = 0, 1, 2, ...$$

Reparametrizavus, t. y. imant $\theta = \ln(1-p)$, matoma, kad neigiamasis binominis skirstinys priklauso kanoninio pavidalo vienparametrių ekponentinio tipo skirstinių šeimai. Tankis (skaičiuojančiojo mato atžvilgiu) yra

$$f(y|\theta) = h(y)e^{y\theta - B(\theta)}, \quad y = 0, 1, 2, ...,$$

$$h(y) = \frac{\Gamma(k+y)}{\Gamma(k)y!}, \quad B(\theta) = -k\ln(1-e^\theta), \quad \theta = \ln(1-p).$$

Imties $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ su nepriklausomomis koordinatėmis, kai $Y_i \sim B^-(k_i, p_i)$, $0 < p_i < 1$ – nežinomi parametrai, o $k_i > 0$ žinomi, tikėtinumo funkcija

$$L(p_1, ..., p_n) = \prod_{i=1}^{n} h(Y_i) e^{\sum_i [Y_i\theta_i - B(\theta_i)]}, \quad \theta_i = \ln(1-p_i), \quad B(\theta_i) = k_i \ln(1-e^{\theta_i}). \tag{5.2.6}$$

Tai prisotintas modelis, kai nežinomų parametrų $\theta_1, ..., \theta_n$ skaičius lygus imties didumui n.

Jeigu kartu su Y_i gaunama kovariančių vektoriaus $\boldsymbol{x} = (x_0, x_1, ..., x_m), x_0 = 1$ reikšmė $\boldsymbol{x}_i = (x_{0i}, x_{1i}, ..., x_{mi})^T$, tai parametrų skaičius sumažinamas iki $m+1 < n$ imant tiesinius darinius $\boldsymbol{\beta}^T\boldsymbol{x}_i = \beta_0 x_{0i} + ... + \beta_m x_{mi}$. Parinkti kanoninę jungties funkciją, kai θ_i keičiamas tiesiniu dariniu $\boldsymbol{\beta}^T\boldsymbol{x}_i$, negalima, nes $\ln(1-e^{\theta_i})$ neapibrėžtas, kai $\boldsymbol{\beta}^T\boldsymbol{x}_i > 1$. Kadangi vidurkis

$$\mathbf{E}_{\theta_i} Y_i = \dot{B}(\theta_i) = \frac{k_i e^{\theta_i}}{1 - e^{\theta_i}} = \frac{k_i(1-p_i)}{p_i} > 0,$$

tai naujus parametrus galima įvesti keičiant $(1 - p_i)/p_i$ neneigiamu reiškiniu $e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}$. Tada gauname

$$\mu_i = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i) = k_i e^{\boldsymbol{\beta}^T\boldsymbol{x}_i},$$

$$g(\mu_i) = \ln(\mu_i/k_i), \quad \theta_i = \ln\frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}.$$

Lygčių sistema (5.1.9) parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ DT įvertiniui rasti turi tokį pavidalą:

$$\dot{l}_j = \sum_{i=1}^{n} x_{ji} \frac{Y_i - k_i e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}\boldsymbol{x}_i}} = 0, \quad j = 0, 1, ..., m. \tag{5.2.7}$$

Fišerio informacinė matrica (5.1.10) yra

$$\boldsymbol{I}(\boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\mathbf{X},$$

kai diagonaliosios matricos $\boldsymbol{W}(\boldsymbol{\beta})$ diagonaliniai elementai yra

$$k_i \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}\boldsymbol{x}_i}}, \quad i = 1, 2, ..., n.$$

Tikrinant hipotezę $H_{1,...,m} : \beta_1 = ... = \beta_m = 0$ tikėtinumų santykio kriterijumi (5.1.16), parametro θ įvertinys modelyje su vienu parametru yra

$$\hat{\theta} = \sum_{i=1}^{n} k_i / \sum_{i=1}^{n} Y_i = \bar{k}/\bar{Y}.$$

5.2.3 pavyzdys. Atliekant ištisinę produkcijos kontrolę nuo konvejerio juostos imamas kiekvienas l-asis gaminys ir nustatoma, ar jis geras ar defektinis. Ciklas užbaigiamas, kai bus surasta k defektinių gaminių. Atsižvelgiant į gerų gaminių skaičių Y iš patikrintųjų priimami tam tikri sprendimai. Pavyzdžiui, sprendimas tikrinti ir reguliuoti technologinį procesą (Y įgijo mažą reikšmę), arba silpninti kontrolę padidinant l (Y įgijo didelę reikšmę). Jeigu tarsime, kad i-ojo ciklo metu defektinio gaminio pasirodymo tikimybė p_i yra pastovi ir defektiniai gaminiai pasirodo nepriklausomai vienas nuo kito (Bernulio eksperimentų schema), tai i-ojo ciklo metu rastų gerų gaminių skaičius Y_i turi neigiamąjį binominį skirstinį $B^-(k, p_i)$.

Jeigu reikia ištirti Y skirstinio priklausomybę nuo kovariančių vektoriaus $\boldsymbol{x}$ (tecnologinio proceso charakteristikos, žaliavų parametrai ir kt.), tai imties $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$ analizę reikėtų atlikti naudojant neigiamąją binominę regresiją.

5.2.2 pastaba. Jeigu k_i yra sveikasis skaičius, tai $Y_i \sim B^-(k_i, p_i)$ (Paskalio skirstinys) yra suma $Y_i = Y_{i1} + ... + Y_{ik_i}$ vienodai pasiskirsčiusių n. a. d. $Y_{ij} \sim B^-(1, p_i)$. Atsitiktinis dydis $Z_{ij} = Y_{ij} + 1$ turi geometrinį skirstinį su parametru p_i. Todėl šiuo atveju neigiamojo skirstinio regresija faktiškai yra geometrinio skirstinio regresija remiantis imtimi $(Y_{ij}, \boldsymbol{x}_i), j = 1, ..., k_i, i = 1, ..., n$, kai imties didumas $N = k_1 + ... + k_n$. Kovariantės $\boldsymbol{x}$ reikšmė $\boldsymbol{x}_i$ kartojasi k_i kartų. Nesunku patikrinti, kad parametrams $\beta_0, ..., \beta_m$ vertinti pakanka žinoti sumas $Y_1, ..., Y_n$.

5.2.4. Binominė regresija

Atsitiktinio dydžio $Y \sim B(k, p)$ su nežinomu parametru $0 < p < 1$ ir žinomu Bernulio eksperimentų skaičiumi k skirstinys nusakomas tikimybėmis

$$P\{Y = m|p\} = C_k^m p^m (1-p)^{k-m} = C_k^m e^{m \ln(p/(1-p)) + k \ln(1-p)}, \quad m = 0, 1, 2, ..., k.$$

Reparametrizavus, t. y. imant $\theta = \ln(p/(1-p))$, matome, kad binominis skirstinys priklauso kanoninio pavidalo vienparametrių ekponentinio tipo skirstinių šeimai. Tankis (skaičiuojančiojo mato atžvilgiu) yra

$$f(y|\theta) = h(y) e^{y\theta - B(\theta)}, \quad y = 0, 1, 2, ..., k,$$

$$h(y) = C_k^y, \quad B(\theta) = -k \ln(1 + e^\theta), \quad \theta = \ln \frac{p}{1-p}.$$

Kadangi $Y \sim B(k, p)$ yra suma $Y = Z_1 + ... + Z_k$ vienodai pasiskirsčiusių n. a. d. $Z_j \sim B(1, p_i)$, turinčių Bernulio skirstinius, tai, sudarant modelį, galima apsiriboti Bernulio skirstiniais. Šių skirstinių regresijos modelis vadinamas *logistine regresija*. Dėl šio modelio aktualumo ir dažno taikymo praktikoje, jis detaliai aptariamas tolesniame skyrelyje. Plačiau apie logistinę regresiją žr. [8], [9].

5.3. Logistinė regresija

5.3.1. Logistinės regresijos modelis

Tarkime, kad atsitiktinio įvykio A tikimybė gali priklausyti nuo nepriklausomų kintamųjų (kovariančių) $x_1, ..., x_m$. Pavyzdžiui, tikimybė gimti neišnešiotam kūdikiui gali priklausyti nuo motinos svorio, ligų, rūkymo ir kt.

Apibrėžkime atsitiktinį dydį Y, kuris įgyja reikšmę 1, kai įvykis A įvyksta ir įgyja reikšmę 0, kai įvykis A neįvyksta. Taigi galime sakyti, kad eksperimento metu stebimos a. d. Y reikšmės. Pažymėkime $\boldsymbol{x} = (x_0, x_1, ..., x_m)^T$ kovariančių vektorių, papildytą koordinate $x_0 = 1$. Remiantis kovariančių vektoriumi $\boldsymbol{x}$ reikia prognozuoti a. d. Y, įgyjantį tik dvi reikšmes 0 ir 1.

Remiantis 3.1 skyreliu, optimali prognozė yra a. d. Y regresija vektoriaus $\boldsymbol{x}$ atžvilgiu, t. y. Y sąlyginis vidurkis, kai $\boldsymbol{x}$ fiksuotas:

$$\pi(\boldsymbol{x}) = \mathbf{E}(Y|\boldsymbol{x}) = \mathbf{P}\{Y = 1|\boldsymbol{x}\} = \mathbf{P}\{A|\boldsymbol{x}\}.$$

Tiesinėje regresijoje sąlyginis vidurkis $\mathbf{E}(Y|\boldsymbol{x})$ aprašomas tiesine $\boldsymbol{x}$ funkcija $\mu(\boldsymbol{x})$, priklausančia nuo nežinomo parametro $\boldsymbol{\beta}$:

$$\mu(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m.$$

Nagrinėjamu atveju su bet kuria kovariantės $\boldsymbol{x}$ reikšme sąlyginio vidurkio $\mathbf{E}(Y|\boldsymbol{x})$ reikšmė priklauso intervalui [0, 1]. Todėl modelis

$$\pi(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m \tag{5.3.1}$$

turi trūkumų: įvertinę parametrus $\boldsymbol{\beta} = (\beta_0, ..., \beta_m)$, galime gauti $\pi(\boldsymbol{x})$ įvertį, nepriklausantį intervalui $[0, 1]$.

Šis trūkumas pašalinamas, nagrinėjant kitokį modelį.

Logistinės regresijos modelis:

$$\mathit{logit}(\boldsymbol{x}) = \ln\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m = \boldsymbol{\beta}^T\boldsymbol{x} = \mu(\boldsymbol{x}). \tag{5.3.2}$$

Funkcijos $\mathit{logit}(\boldsymbol{x})$ apibrėžimo sritis yra $\mathbf{R}$, tačiau su bet kuriais $\boldsymbol{\beta}$ ir $\boldsymbol{x}$ funkcija $\pi(\boldsymbol{x})$ įgyja reikšmes iš intervalo (0, 1). Kai $\mu(\boldsymbol{x}) \to \infty$, tai $\pi(\boldsymbol{x}) \to 1$, o kai $\mu(\boldsymbol{x}) \to -\infty$, tai $\pi(\boldsymbol{x}) \to 0$. Iš modelio išplaukia, kad įvykio A sąlyginė tikimybė žinant $\boldsymbol{x}$ apibrėžiama formule

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_m x_m}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_m x_m}} = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x}}}, \tag{5.3.3}$$

o priešingo įvykio $\bar{A}$ tikimybė žinant $\boldsymbol{x}$ yra

$$1 - \pi(\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x}}}.$$

5.3.2. Regresinių parametrų interpretavimas

Sudarykime santykį

$$\gamma(\boldsymbol{x}) = \frac{\mathbf{P}\{A|\boldsymbol{x}\}}{\mathbf{P}\{\bar{A}|\boldsymbol{x}\}} = \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = e^{\boldsymbol{\beta}^T\boldsymbol{x}},$$

kurį vadiname įvykio $\{Y = 1\}$ *galimybės santykiu* arba *šansu*, kai $\mathbf{X} = \boldsymbol{x}$. Šansas rodo, kiek kartų didesnė tikimybė įvykti įvykiui A, palyginti su priešingo įvykio $\bar{A}$ pasirodymo tikimybe, kai kovariančių vektorius $\mathbf{X}$ įgijo reikšmę $\boldsymbol{x}$. Pavyzdžiui, jeigu $\gamma(\boldsymbol{x}) = 4$, tai reiškia, kad ivykio A pasirodymo tikimybės ir įvykio $\bar{A}$ pasirodymo tikimybės santykis yra 4:1, t. y. $\pi(\boldsymbol{x}) = 0,8$ (būtent šia prasme sąvoką „šansas“ vartoja lažybų tarpininkai).

Imkime dvi skirtingas kovariantės $\mathbf{X}$ reikšmes $\boldsymbol{x}^{(1)}$ ir $\boldsymbol{x}^{(2)}$ ir sudarykime šansų santykį

$$\frac{\gamma(\boldsymbol{x}^{(2)})}{\gamma(\boldsymbol{x}^{(1)})} = \frac{\pi(\boldsymbol{x}^{(2)})/(1-\pi(\boldsymbol{x}^{(2)})}{\pi(\boldsymbol{x}^{(1)})/(1-\pi(\boldsymbol{x}^{(1)})} = e^{\boldsymbol{\beta}^T\boldsymbol{x}^{(2)}-\boldsymbol{\beta}^T\boldsymbol{x}^{(1)}},$$

kuris parodo, kiek kartų pasikeičia įvykio A šansas, kai kovariačių vektorius $\mathbf{X}$ pakinta nuo $\boldsymbol{x}^{(1)}$ iki $\boldsymbol{x}^{(2)}$. Regresinių parametrų interpretacija tiesiogiai susijusi su šansų santykiu.

Tarkime, kad j-oji kovariantė yra tolydžioji. Imkime du kovariančių vektorius $\boldsymbol{x}^{(1)}$ ir $\boldsymbol{x}^{(2)}$, kurių visos koordinatės, išskyrus j-ąją, yra vienodos, o $x_j^{(2)} = x_j^{(1)} + 1$.

Pagal (5.1.2) formulę

$$\frac{\pi(\boldsymbol{x}^{(2)})/(1-\pi(\boldsymbol{x}^{(2)})}{\pi(\boldsymbol{x}^{(1)})/(1-\pi(\boldsymbol{x}^{(1)})} = e^{logit(x_1,\ldots,x_j^{(2)},\ldots,x_m)-logit(x_1,\ldots,x_j^{(1)},\ldots,x_m)} = e^{\beta_j}. \quad (5.3.4)$$

Taigi *parametras e^{β_j} parodo, kiek kartų pasikeičia įvykio A šansas, kai j-oji kovariantė padidėja vienetu, kitoms kovariantėms nepakitus; parametras e^{β_j} yra šansų santykis.*

Jei j-oji kovariantė nominali, tai, norint, kad modelio parametrai turėtų prasmę, ši kovariantė koduojama lygiai taip pat kaip tiesinės regresijos atveju.

Tarkime, kad j-oji kovariantė nominali (pavyzdžiui, ligos stadija, lytis) ir įgyja k skirtingų reikšmių. Tada vietoje $\beta_j x_j$ (5.1.2) modelyje imamas narys

$$\boldsymbol{\beta}_j^T\boldsymbol{x}_j = \beta_{j1}x_{j1} + \beta_{j2}x_{j2} + \cdots + \beta_{j,k-1}x_{j,k-1};$$

čia

$$\boldsymbol{x}_j = (x_{j1},\ldots,x_{j,k-1})^T, \quad \boldsymbol{\beta} = (\beta_{j1},\ldots,\beta_{j,k-1})^T,$$

$$x_{jl} = \begin{cases} 1, & \text{jei } x_j \text{ įgyja } l\text{-ąją reikšmę} \quad (l = 2,\ldots,k); \\ 0, & \text{kitais atvejais,} \end{cases}$$

Taigi pirmąją kovariantės reikšmę atitinka vektoriaus $\boldsymbol{x}_j = (x_{j1},\ldots,x_{j,k-1})^T$ reikšmė $(0,0,\ldots,0)^T$, o $(l+1)$-ąją reikšmę $(l = 1,\ldots,k-1)$ atitinka šio vektoriaus reikšmė $(0,\ldots,1,\ldots 0)^T$, čia vienetas yra l-oje pozicijoje.

Gauname modelį:

$$logit(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \beta_0 + \beta_1 x_1 + ... + \sum_{i=1}^{k-1}\beta_{ji}x_{ji} + ... + \beta_m x_m. \quad (5.3.5)$$

Tuo atveju j-ąją kovariantę atitinkantis narys įgyja tokias reikšmes:

$$\boldsymbol{\beta}_j^T\boldsymbol{x}_j = \begin{cases} \beta_{jl}, & \text{jei } j\text{-oji kovariantė įgyja } (l+1)\text{-ąją reikšmę} \quad (l = 1,\ldots,k-1); \\ 0, & \text{jei } j\text{-oji kovariantė įgyja nulinę reikšmę.} \end{cases}$$

Jei, pavyzdžiui, x_j yra ligos stadija, įgyjanti 4 reikšmes, o įvykis A yra išgyvenimas praėjus metams po operacijos, tai modelyje (5.1.2) imti narį $\beta_j x_j$ su

x_j, įgyjančiu reikšmes 1, 2, 3 ir 4, būtų neteisinga, nes toks modelis reikštų, kad jau iš karto darome prielaidą, kad antros ir pirmos stadijų pacientų išgyvenimo šansų santykis lygus trečios ir antros stadijos pacientų bei ketvirtos ir trečios stadijos pacientų išgyvenimo šansų santykiui. Regresinės analizės tikslas yra būtent nustatyti, kaip priklauso išgyvenimo šansas nuo ligos stadijos. Šiuo atveju vietoje nario $\beta_j x_j$ imamas narys $\beta_{j1}x_{j1} + \beta_{j2}x_{j2} + \beta_{j3}x_{j3}$, x_{j1}, x_{j2} ir x_{j3} įgyja reikšmę 1 atitinkamai antros, trečios ir ketvirtos stadijos pacientams.

Panagrinėkime koeficientų ir modelio interpretaciją po kodavimo. Imkime du kovariančių vektorius $\boldsymbol{x}^{(1)}$ ir $\boldsymbol{x}^{(l+1)}$, kuriems visos kovariantės, išskyrus j-ąją nominalią kovariantę, yra vienodos, o j-osios kovariantės reikšmė pirmajam vektoriui yra pirmoji, o antrajam $(l+1)$-oji. Iš (5.2.2) formulės išplaukia, kad

$$\frac{\pi(\boldsymbol{x}^{(l+1)})/(1-\pi(\boldsymbol{x}^{(l+1)})}{\pi(\boldsymbol{x}^{(1)})/(1-\pi(\boldsymbol{x}^{(1)})} = logit(\boldsymbol{x}^{(l+1)}) - logit(\boldsymbol{x}^{(1)}) = e^{\beta_{jl}}. \tag{5.3.6}$$

Taigi *parametras $e^{\beta_{jl}}$ parodo objektų, kurių j-oji kovariantė įgyja l-ąją reikšmę, bei objektų, kurių j-oji kovariantė įgyja nulinę reikšmę, šansų santykį kitoms kovariantėms nepakitus.* Pavyzdžiui, jei x_j yra ligos stadija, įgyjanti 4 reikšmes, tai β_{j1} parodo antros ir pirmos stadijų pacientų išgyvenimo šansų santykį, o β_{j2} trečios ir pirmos stadijų pacientų išgyvenimo šansų santykį.

Jei tas pats kovariantės x_j reikšmės pokytis rodo skirtingą šansų santykį, kai yra įvairios fiksuotos kitų kovariančių reikšmės, turime x_j ir šių kovariančių sąveiką. Tada (5.1.2) modelis modifikuojamas. Pavyzdžiui, kai yra dvi tolydžiosios kovariantės ($m = 2$), naudojamas modelis

$$logit(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \tag{5.3.7}$$

kai yra trys kovariantės:

$$logit(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3. \tag{5.3.8}$$

Kai $m = 2$ gauname

$$\frac{\pi(x_1+1, x_2)/(1-\pi(x_1+1, x_2)}{\pi(x_1, x_2)/(1-\pi(x_1, x_2))} = e^{logit(x_1+1, x_2) - logit(x_1, x_2)} = e^{\beta_1 + \beta_3 x_2}. \tag{5.3.9}$$

Taigi šansų santykis, padidėjus pirmajai kovariantei vienetu, priklauso nuo antrosios kovariantės reikšmės x_2 ir lygus $e^{\beta_1+\beta_3 x_2}$.

Jei x_1 yra tolydžioji kovariantė, o x_2 yra nominali kovariantė, įgyjanti keturias reikšmes, tai naudojamas modelis

$$logit(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \beta_{22} x_{22} + \beta_{23} x_{23} + \beta_{121} x_1 x_{21} + \beta_{122} x_1 x_{22} + \beta_{123} x_1 x_{23}.$$

Tada

$$logit(x_1+1, x_{21}, x_{22}, x_{23}) - logit(x_1, x_{21}, x_{22}, x_{23}) = \beta_1 + \beta_{121} x_{21} + \beta_{122} x_{22} + \beta_{123} x_{23}.$$

Taigi šansų santykis, padidėjus pirmajai kovariantei vienetu, priklauso nuo antrosios kovariantės reikšmės x_2 ir lygus e^{β_1}, $e^{\beta_1+\beta_{121}}$, $e^{\beta_1+\beta_{122}}$ ir $e^{\beta_1+\beta_{123}}$, kai

atitinkamai yra pirmoji, antroji, trečioji ir ketvirtoji nominalios kovariantės reikšmė.

Ir atvirkščiai, jei pirmoji kovariantė fiksuota, o antroji pakeičia nulinę reikšmę pirmąja, tai

$$logit(x_1, 1, 0, 0) - logit(x_1, 0, 0, 0) = \beta_{21} + \beta_{121} x_1.$$

Taigi šansų santykis $e^{\beta_{21}+\beta_{121}x_1}$ priklauso nuo pirmosios kovariantės reikšmių.

Jei, pavyzdžiui, x_1 yra paciento amžius, x_2 – ligos stadija, įgyjanti 4 reikšmes, o įvykis A – išgyvenimas praėjus metams po operacijos, tai naudotume (5.2.5) modelį, jei pereinant nuo vienos prie kitos stadijos šansų santykis priklausytų nuo pacientų amžiaus, pavyzdžiui, penkiadešimtmečių ir septyniasdešimtmečių bei trečios ir pirmos stadijų ligonių šansų santykis gali būti skirtingas.

5.3.3. Regresinių parametrų vertinimas

Tarkime, kad nežinomam regresijos parametrui $\boldsymbol{\beta}$ (kartu tikimybei $\pi(\boldsymbol{x})$) vertinti atliekama n nepriklausomų eksperimentų; i-asis eksperimentas atliekamas, kai kovariantės $\boldsymbol{x}$ reikšmė $\boldsymbol{x}^{(i)} = (x_{i0}, ..., x_{im})^T$, $x_{i0} = 1$, $i = 1, ..., n$.

Kiekvieno eksperimento metu stebimas atsitiktinis dydis

$$Y_i = \begin{cases} 1, & \text{jei } i\text{-ojo eksperimento metu įvyksta } A; \\ 0, & \text{priešingu atveju.} \end{cases}$$

Taigi turime imtį

$$(Y_1, \boldsymbol{x}^{(1)}), ..., (Y_n, \boldsymbol{x}^{(n)}), \tag{5.3.10}$$

kuri nėra paprastoji, nes a. v. $(Y_i, \boldsymbol{x}^{(i)})^T$ nėra vienodai pasiskirstę. Atsitiktiniai dydžiai Y_i turi sąlyginius Bernulio skirstinius:

$$(Y_i | \boldsymbol{x}^{(i)}) \sim B(1, \pi(\boldsymbol{x}^{(i)})), \quad i = 1, ..., n.$$

Tikėtinumo funkcija yra

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi(\boldsymbol{x}^{(i)})]^{Y_i} [1 - \pi(\boldsymbol{x}^{(i)})]^{1-Y_i}, \tag{5.3.11}$$

jos logaritmas

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i \ln \pi(\boldsymbol{x}^{(i)}) + (1 - Y_i) \ln (1 - \pi(\boldsymbol{x}^{(i)}))]$$

$$= \sum_{i=1}^{n} [Y_i \ln \frac{\pi(\boldsymbol{x}^{(i)})}{1 - \pi(\boldsymbol{x}^{(i)})} + \ln (1 - \pi(\boldsymbol{x}^{(i)}))]$$

$$= \sum_{i=1}^{n} [Y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}) - \ln (1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}})],$$

o informančių vektoriaus $\dot{\ell}(\boldsymbol{\beta}) = (\dot{\ell}_0(\boldsymbol{\beta}), \dot{\ell}_1(\boldsymbol{\beta}), \ldots, \dot{\ell}_m(\boldsymbol{\beta}))^T$ koordinatės

$$\dot{\ell}_j(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}[Y_i - \pi(\boldsymbol{x}^{(i)})], \quad j = 0, ..., m.$$

Didžiausiojo tikėtinumo įvertinys $\hat{\boldsymbol{\beta}}$ tenkina lygčių sistemą

$$\dot{\ell}_j(\boldsymbol{\beta}) = 0, \quad j = 0, ..., m.$$

Suradus įvertinį $\hat{\boldsymbol{\beta}}$, galima įvertinti įvykio A sąlyginę tikimybę $\pi(\boldsymbol{x})$, šansų santykius e^{β_i}:

$$\hat{\pi}(\boldsymbol{x}) = \frac{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}}}{1 + e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}}}, \quad e^{\hat{\beta}_i}, \quad i = 1, ..., m,$$

ar kitas parametro $\boldsymbol{\beta}$ funkcijas.

5.3.1 pavyzdys. Atliktas skausmą mažinančių vaistų poveikio pagyvenusiems žmonėms, besiskundžiantiems neuralgija, tyrimas(žr. [13]). Priklausomas kintamasis Y: skundėsi pacientas skausmais po gydymo kurso ($Y = 0$) ar nesiskundė ($Y = 1$); kovariantė X_1 – gydymo tipas (A, B arba placebas P); kovariantė X_2 – paciento lytis; kovariantė X_3 – paciento amžius; kovariantė X_4 – laikas, kurį pacientas jautė skausmus iki gydymo kurso pradžios. Duomenys pateikti 5.3.1 lentelėje.

5.3.1 lentelė. Statistiniai duomenys

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y_i	1	1	1	0	1	1	1	1	0	0	1	0	1	1	0
X_{11i}	0	0	0	0	0	0	1	0	0	1	1	1	0	1	0
X_{12i}	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0
X_{21i}	1	0	1	0	1	1	1	1	1	0	1	1	1	0	0
X_{3i}	68	74	67	66	67	77	71	72	76	71	63	69	66	62	64
X_{4i}	1	16	30	26	28	16	12	50	9	17	27	18	12	42	1
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Y_i	1	1	1	1	1	0	1	0	0	0	1	1	1	1	1
X_{11i}	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0
X_{12i}	1	0	0	0	1	0	1	1	0	0	0	0	1	1	1
X_{21i}	0	1	0	0	1	0	1	0	0	1	0	1	1	0	0
X_{3i}	59	64	70	69	78	83	69	75	77	79	70	69	65	70	67
X_{4i}	29	30	28	1	1	1	42	30	29	20	12	12	14	1	23
i	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Y_i	1	0	0	0	1	1	0	0	1	0	0	0	1	1	0
X_{11i}	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0
X_{12i}	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
X_{21i}	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1
X_{3i}	65	60	78	75	67	72	70	75	65	68	68	67	70	65	67
X_{4i}	29	26	15	21	11	27	13	6	7	27	11	17	22	15	1
i	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Y_i	1	1	1	0	1	0	0	0	1	0	1	0	1	0	1
X_{11i}	1	0	1	0	0	1	0	0	0	0	1	0	1	0	1
X_{12i}	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0
X_{21i}	0	1	0	0	1	0	0	0	1	0	0	1	1	0	1
X_{3i}	64	74	72	70	66	76	78	77	69	66	67	72	74	80	69
X_{4i}	17	4	25	1	19	25	12	1	24	4	10	11	1	21	3

Kadangi kovariantė X_1 yra nominali ir įgyja tris reikšmes, tai ji lentelėje koduota remiantis 5.2 skyreliu (placebas – $(0,0)$, metodas A – $(1,0)$, metodas B – $(0,1)$); kovariantė X_2 taip pat nominali, įgyja dvi reikšmes, todėl ji taip pat koduota (moteris – 1, vyras – 0).

Remdamiesi logistinės regresijos modeliu (5.1.2)

$$logit(\mathbf{X}) = \beta_0 + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{21}X_{21} + \beta_3 X_3 + \beta_4 X_4 = \mu(\mathbf{X})$$

pagal 5.2.1 lentelės duomenis įvertinsime nežinomus parametrus $\beta_0, \beta_{11}, \beta_{12}, \beta_{21}, \beta_3, \beta_4$. Įverčiams rasti naudosime SAS programų paketą. Detaliau apie SAS procedūros LOGISTIC galimybes žr. [11], p. 347. Knygoje minėtos procedūros galimybės iliustruojamos būtent šiuo pavyzdžiu. Lentelėje 5.3.2 pateikta dalis LOGISTIC procedūros skaičiavimo rezultatų.

5.3.2 lentelė. Parametrų įverčiai

Parametras	Įvertis	Paklaida	χ^2 statistika	P-reikšmė	Šansų santykis
β_0	15,5744	6,5912	5,5828	0,0181	
β_{11}	3,1817	1,0161	9,8049	0,0017	24,087
β_{12}	3,7085	1,1407	10,5700	0,0011	40,794
β_{21}	1,8322	0,7963	5,2946	0,0214	6,248
β_3	-0,2621	0,0970	7,2977	0,0069	0,769
β_4	0,00586	0,0330	0,0315	0,8591	1,006

Parametrų įverčiai pateikti antrajame stulpelyje. Paskutiniame stulpelyje pateikti šansų santykio įverčiai. Šiuos įverčius galima interpretuoti taip: didėjant amžiui šansas nesiskųsti skausmais po gydymo mažėja (šansų santykio įvertis 0,769); o pereinant nuo placebo prie gydymo būdo B šansas nesiskųsti skausmais gerokai padidėja (šansų santykio įvertis 40,794) ir pan.

Tarkime, reikia įvertinti įvykio $\{Y = 1\}$ tikimybę $\pi(\boldsymbol{x})$, kai kovariančių vektorius $\mathbf{X}$ įgijo reikšmę $\boldsymbol{x} = (1; (1\ \ 0); 1; 70; 5)^T$, t.y. taikytas gydymo būdas A 70 metų moteriai, kuri skundėsi skausmais 5 metus iki gydymo pradžios. Naudodamiesi 5.3.2 lentele gauname

$$\hat{\mu}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}^T\boldsymbol{x} = 15,5744 + 3,1817 \times 1 + 1,8322 \times 1 - 0,2621 \times 70 + 0,00586 \times 5 = 2,2706.$$

$$\hat{\pi}(\boldsymbol{x}) = e^{\hat{\mu}(\boldsymbol{x})}/(1 + e^{\hat{\mu}(\boldsymbol{x})}) = 0,9064.$$

5.3.4. Regresinių parametrų įvertinių savybės

Ieškokime Fišerio informacinės matricos

$$\mathbf{I}(\boldsymbol{\beta}) = [I_{ls}(\boldsymbol{\beta})]_{(m+1)\times(m+1)}, \quad I_{ls}(\boldsymbol{\beta}) = -\mathbf{E}\left(\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_l \partial\beta_s}\right).$$

Naudojant logistinės regresijos modelį išvestinė

$$-\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_l \partial\beta_s} = \sum_{i=1}^{n} x_{il}x_{is}\pi(\boldsymbol{x}^{(i)})(1 - \pi(\boldsymbol{x}^{(i)})) \quad (l, s = 0, ..., m)$$

nėra atsitiktinė, taigi

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{V}(\boldsymbol{\beta})\mathbf{X};$$

čia

$$\mathbf{X} = \begin{pmatrix} x_{10} & ... & x_{1m} \\ ... & ... & ... \\ x_{n0} & ... & x_{nm} \end{pmatrix},$$

$$\mathbf{V}(\boldsymbol{\beta}) = \begin{pmatrix} \pi(\boldsymbol{x}^{(1)})(1-\pi(\boldsymbol{x}^{(1)})) & ... & 0 \\ 0 & ... & 0 \\ 0 & ... & \pi(\boldsymbol{x}^{(n)})(1-\pi(\boldsymbol{x}^{(n)})) \end{pmatrix}.$$

Remiantis DT įvertinių savybėmis (žr. 1 dalis, 3.5.15 pavyzdys), galima tvirtinti, kad kai n dideli, gana bendromis sąlygomis įvertinio $\hat{\boldsymbol{\beta}}$ skirstinys gali būti aproksimuotas normaliuoju:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{b.t.}{\to} \boldsymbol{U} \sim N_{m+1}(\mathbf{0}, \boldsymbol{i}^{-1}(\boldsymbol{\beta})), \tag{5.3.12}$$

čia

$$\boldsymbol{i}(\boldsymbol{\beta}) = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{I}(\boldsymbol{\beta}).$$

Tariama, kad ši riba egzistuoja, o matricos $\boldsymbol{i}(\boldsymbol{\beta})$ rangas lygus $m+1$. Matricos $\boldsymbol{i}^{-1}(\boldsymbol{\beta})$ elementus žymėsime $\sigma_{ls}(\boldsymbol{\beta})$.

Jei $\boldsymbol{x} = (x_0, ..., x_m)^T$ fiksuotas kovariančių vektorius, tai pagal delta metodą įvertinio $\hat{\pi}(\boldsymbol{x})$ skirstinys taip pat aproksimuojamas normaliuoju:

$$\sqrt{n}(\hat{\pi}(\boldsymbol{x}) - \pi(\boldsymbol{x})) \overset{b.t.}{\to} U \sim N(0, \sigma^2_{\hat{\pi}(\boldsymbol{x})}); \tag{5.3.13}$$

čia

$$\sigma^2_{\hat{\pi}(\boldsymbol{x})} = \left(\frac{\partial\pi(\boldsymbol{x})}{\partial\boldsymbol{\beta}}\right)^T_{1\times(m+1)} \boldsymbol{i}^{-1}(\boldsymbol{\beta}) \left(\frac{\partial\pi(\boldsymbol{x})}{\partial\boldsymbol{\beta}}\right)_{(m+1)\times 1} = \sum_{l=0}^{m}\sum_{s=0}^{m} \frac{\partial\pi(\boldsymbol{x})}{\partial\beta_l}\sigma_{ls}(\boldsymbol{\beta})\frac{\partial\pi(\boldsymbol{x})}{\partial\beta_s}$$

$$= \pi^2(\boldsymbol{x})(1-\pi(\boldsymbol{x}))^2 \sum_{l=0}^{m}\sum_{s=0}^{m} x_l x_s \sigma_{ls}(\boldsymbol{\beta}) = \pi^2(\boldsymbol{x})(1-\pi(\boldsymbol{x}))^2 \boldsymbol{x}^T \boldsymbol{i}^{-1}(\boldsymbol{\beta})\boldsymbol{x}. \tag{5.3.14}$$

Dispersijos $\sigma^2_{\hat{\pi}(\boldsymbol{x})}$ pagrįstas įvertinys

$$\hat{\sigma}^2_{\hat{\pi}(\boldsymbol{x})} = \hat{\pi}^2(\boldsymbol{x})(1-\hat{\pi}(\boldsymbol{x}))^2 \boldsymbol{x}^T \boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}. \tag{5.3.15}$$

5.4.1 pavyzdys (**5.3.1** pavyzdžio tęsinys). 5.3.1 pratimo sąlygomis rasime Fišerio informacinės matricos $\boldsymbol{i}(\boldsymbol{\beta}) = \boldsymbol{I}(\boldsymbol{\beta})/n$ atvirkštinės $\boldsymbol{i}^{-1}(\boldsymbol{\beta})$ įvertį $\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}}) = [\sigma_{ls}(\hat{\boldsymbol{\beta}})]_{6\times 6}$. Naudodami šią matricą rasime regresijos parametrų, šansų santykių ir tikimybės $\pi(\boldsymbol{x})$ įvertinių dispersijų įverčius.

Naudodamiesi 5.3.1 pavyzdyje surastais regresijos parametrų įverčiais gauname

$$\frac{\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})}{n} = \begin{pmatrix} 43,4483 & 2,0018 & 3,8851 & 0,3464 & -0,6322 & -0,0521 \\ 2,0018 & 1,0325 & 0,7278 & 0,3245 & -0,0386 & 0,0009 \\ 3,8851 & 0,7278 & 1,3012 & 0,3261 & -0,0658 & -0,0028 \\ 0,3464 & 0,3245 & 0,3261 & 0,6340 & -0,0123 & 0,0019 \\ -0,6322 & -0,0386 & -0,0658 & -0,0123 & 0,0094 & 0,0005 \\ -0,0521 & 0,0009 & -0,0028 & 0,0019 & 0,0005 & 0,0011 \end{pmatrix}$$

$$= \left[\frac{\sigma_{ls}(\hat{\boldsymbol{\beta}})}{n}\right]_{6\times 6}.$$

Tegu β_j yra kuris nors regresijos parametras. Tada įvertinio $\hat{\beta}_j$ dispersijos pagrįstas įvertinys yra gautos matricos diagonalinis elementas

$$\hat{\boldsymbol{V}}(\hat{\beta}_j) = \sigma_{jj}(\hat{\boldsymbol{\beta}})/n.$$

Atliekant skaičiavimus SAS procedūra LOGISTIC šie įvertiniai pateikiami automatiškai. Pagal 5.3.1 pratimo duomenis 5.3.2 lentelės trečiajame stulpelyje „Paklaidos“ yra pateikti kvadratinių nuokrypių įverčiai, t.y. $\sqrt{\sigma_{jj}(\hat{\boldsymbol{\beta}})/n}$.

Remiantis delta metodu, šansų santykio e_j^{β} įvertinio $e^{\hat{\beta}_j}$ pagrįstas dispersijos įvertinys yra

$$\hat{\boldsymbol{V}}(e^{\hat{\beta}_j}) = e^{2\hat{\beta}_j}\hat{\boldsymbol{V}}(\hat{\beta}_j) = e^{2\hat{\beta}_j}\sigma_{jj}(\hat{\boldsymbol{\beta}})/n.$$

Tikimybės $\pi(\boldsymbol{x})$ įvertinio $\hat{\pi}(\boldsymbol{x})$ dispersijos pagrįstas įvertinys randamas pagal (5.4.4) formulę. Pavyzdžiui, imdami 5.3.1 pavyzdžio kovariančių vektorių, gauname tokią dispersijos įvertinio realizaciją

$$\hat{\sigma}^2_{\hat{\pi}(\boldsymbol{x})} = \hat{\pi}^2(\boldsymbol{x})(1-\hat{\pi}(\boldsymbol{x}))^2\boldsymbol{x}^T\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x} = 0,3067.$$

5.3.5. Tikėtinumų santykiai ir determinacijos koeficientas

Aptarsime, kokios logistinės regresijos sąvokos yra determinacijos koeficiento R^2, kvadratų sumų (pilnosios SS_T, liekamųjų paklaidų SS_E ir regresijos SS_R) sąvokų, nagrinėtų skyrelyje 3.3.8, ekvivalentai?

Pažymėkime

$$\hat{Y}_i = \hat{\pi}(\boldsymbol{x}^{(i)}) = \frac{e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{x}^{(i)}}}{1+e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{x}^{(i)}}}$$

stebėtos reikšmės Y_i prognozę.

Tiesinėje regresijoje atveju prognozuojamos reikšmės buvo apibrėžtos formule $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}^{(i)}$. Prognozė gera, jei stebėtos reikšmės Y_i ir prognozuojamos reikšmės $\hat{Y}_i$ artimos; Y_i ir $\hat{Y}_i$ artumą apibūdina likutinė kvadratų suma $SS_E = \sum(Y_i - \hat{Y}_i)^2$. Kai paklaidų skirstinys normalusis, SS_E/σ^2 turi chi kvadrato skirstinį su $n-m-1$ laisvės laipsnių.

Logistinėje regresijoje nagrinėkime tris vieną į kitą įdėtus modelius.

1) Plačiausias modelis gaunamas, kai funkcija $\pi(\boldsymbol{x})$ nežinoma ir norima įvertinti tikimybes

$$\mathbf{P}\{Y_i = 1|\boldsymbol{x}^{(i)}\} = \pi(\boldsymbol{x}^{(i)}) = p_i.$$

Vertinama n nežinomų parametrų $p_1, ..., p_n$. Turime *prisotintą* modelį, nes parametrų skaičius sutampa su imties didumu n.

Tikėtinumo funkcija

$$L_0(\boldsymbol{p}) = L_0(p_1, ..., p_n) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

maksimizuojama taške $\hat{\boldsymbol{p}} = (\hat{p}_1, ..., \hat{p}_n)$; čia $\hat{p}_i = Y_i$.

Jei $Y_i = 1$, tai tikėtinumo funkcijos i-asis daugiklis yra p_i, taigi i-asis $L_0(\hat{p}_1, ..., \hat{p}_n)$ daugiklis yra $\hat{p}_i = Y_i = 1$; jei $Y_i = 0$, tai tikėtinumo funkcijos i-asis daugiklis yra $1-p_i$, taigi i-asis $L_0(\hat{p}_1, ..., \hat{p}_n)$ daugiklis yra $1-\hat{p}_i = 1-Y_i = 1$. Gavome, kad tikėtinumo funkcijos maksimumas

$$L_0(\hat{\boldsymbol{p}}) = 1.$$

2) Nagrinėkime (siauresnį) logistinės regresijos modelį (5.1.2), kai yra $m+1<n$ nežinomų parametrų $\beta_0,...,\beta_m$. Didžiausiojo tikėtinumo funkcija

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi(\boldsymbol{x}^{(i)})^{Y_i}(1-\pi(\boldsymbol{x}^{(i)}))^{1-Y_i}$$

maksimizuojama taške $\hat{\boldsymbol{\beta}}$ ir jos maksimumas yra

$$L(\hat{\boldsymbol{\beta}})=\prod_{i=1}^{n}\hat{Y}_i^{Y_i}(1-\hat{Y}_i)^{1-Y_i}\le L_0(\hat{\boldsymbol{p}}).$$

3) Dar siauresnis modelis gaunamas, kai apskritai nėra regresijos. Šiuo atveju

$$\beta_1=...=\beta_m=0\quad\text{ir}\quad\pi(\boldsymbol{x}^{(i)})=\frac{e^{\beta_0}}{1+e^{\beta_0}}=\pi=const.$$

Turime vieną nežinomą parametrą π.

Tikėtinumo funkcija

$$L_1(\pi)=\prod_{i=1}^{n}\pi^{Y_i}(1-\pi)^{1-Y_i}$$

maksimizuojama taške $\hat{\pi}=\bar{Y}=\frac{1}{n}\sum_i Y_i$ ir

$$L_1(\hat{\pi})=\prod_{i=1}^{n}\bar{Y}^{Y_i}(1-\bar{Y})^{1-Y_i}\le L(\hat{\boldsymbol{\beta}})\le L_0(\hat{\boldsymbol{p}}).$$

Nagrinėkime vieną į kitą įdėtų modelių tikėtinumų santykius. Imkime pirmąjį ir antrąjį modelius, turinčius n ir $m+1$ nežinomų parametrų. Jei n didelis ir galioja logistinės regresijos modelis, tai (žr. I d. 4.5.4 skyrelį) a. d.

$$D_E=-2\ln\frac{L(\hat{\boldsymbol{\beta}})}{L_0(\hat{\boldsymbol{p}})}=-2\ln L(\hat{\boldsymbol{\beta}})$$

skirstinys aproksimuojamas chi kvadrato skirstiniu su $n-m-1$ laisvės laipsnių. D_E yra kvadratų sumos SS_E, nagrinėjamos tiesinėje regresijoje, analogas logistinėje regresijoje. Jo skirstinys aproksimuojamas tuo pačiu dėsniu, kaip ir kvadratų suma SS_E.

Imame pirmąjį ir trečiąjį modelius, turinčius n ir 1 nežinomus parametrus. Atsitiktinio dydžio

$$D_T=-2\ln\frac{L_1(\hat{\pi})}{L_0(\hat{\boldsymbol{p}})}=-2\ln L_1(\hat{\pi})$$

dėsnis artimas chi kvadrato dėsniui su $n-1$ laisvės laipsnių, jei teisingas trečias modelis (t. y. nėra regresijos: $\beta_1=...=\beta_m=0$) ir kai n didelis.

Taigi kvadratų sumos SS_T ekvivalentas logistinėje regresijoje yra D_T.

Imame antrąjį ir trečiąjį modelius, turinčius $m+1$ ir 1 nežinomus parametrus. Atsitiktinio dydžio

$$D_R = -2 \ln \frac{L_1(\hat{\pi})}{L(\hat{\boldsymbol{\beta}})}$$

skirstinys artimas chi kvadrato skirstiniui su m laisvės laipsnių, jei $\beta_1 = ... = \beta_m = 0$ ir n didelis.

Atsitiktinis dydis D_R yra kvadratų sumos SS_R ekvivalentas.

Reikia pažymėti, kad teisinga lygybė

$$D_T = D_E + D_R. \tag{5.3.16}$$

Determinacijos koeficientu vadinamas a. d.

$$R^2 = 1 - \frac{D_E}{D_T} = \frac{D_R}{D_T}. \tag{5.3.17}$$

Jei regresinis modelis prognozuoja idealiai, t. y. $\hat{Y}_i = Y_i$, tai $D_E = 0$ ir $R^2 = 1$. Iš tikrųjų, kai $Y_i = 0$, tai $L(\hat{\boldsymbol{\beta}})$ išraiškoje i-asis daugiklis yra $1 - \hat{Y}_i$, o kai $Y_i = 1$, tai šioje išraiškoje i-asis daugiklis yra $\hat{Y}_i$. Taigi, kai $\hat{Y}_i = Y_i$, tai $L(\hat{\boldsymbol{\beta}})$ išraiškoje visi daugikliai lygūs 1 ir todėl $D_E = -2 \ln L(\hat{\boldsymbol{\beta}}) = 0$.

Jei prognozė vienoda su visais i: $\hat{Y}_i = \bar{Y}$, tai logistinės regresijos modelis prognozei netinkamas. Tuo atveju $L(\hat{\boldsymbol{\beta}}) = L_1(\hat{\pi})$, todėl $D_R = 0$ ir $R^2 = 0$.

Taigi analogiškai kaip ir tiesinėje regresijoje, determinacijos koeficientas yra *prognozės kokybės matas.*

5.5.1 pastaba. SAS priocedūra LOGISTIC apskaičiuoja kitaip apibrėžto determinacijos koeficiento

$$\tilde{R}^2 = [1 - (L(\hat{\pi})/L(\hat{\boldsymbol{\beta}}))^{2/n}]/[1 - (L(\hat{\pi}))^{2/n}]$$

reikšmę. Šis koeficientas taip pat lygus 1, kai prognozė tiksli ($L(\hat{\boldsymbol{\beta}}) = 1$), ir lygus 0, kai $\hat{Y}_i = \bar{Y}$ ($L(\hat{\boldsymbol{\beta}}) = L(\hat{\pi})$).

5.3.6. Regresijos parametrų lygybės nuliui hipotezių tikrinimas

Nagrinėkime hipotezę

$$H_0 : \beta_1 = ... = \beta_m = 0.$$

Ši hipotezė reiškia, kad regresijos nėra ir žinant $\boldsymbol{x}$ tikimybės $\pi(\boldsymbol{x})$ prognozė nepagerėja. Hipotezė H_0 gali būti užrašyta ekvivalenčia forma $H_0 : \pi(\boldsymbol{x}) = \pi = const.$

Kai teisinga hipotezė H_0, a. d. D_R skirstinys aproksimuojamas chi kvadrato skirstiniu su m laisvės laipsnių. Hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, jei

$$D_R > \chi^2_\alpha(m).$$

Nagrinėkime hipotezę

$$H_0 : \beta_{j_1} = ... = \beta_{j_l} = 0 \quad (1 \le j_1 < ... < j_l \le m,\ l < m).$$

Pažymėkime $D_R^{(m)}$ ir $D_R^{(m-l)}$ statistiką D_R atitinkamai modeliui (5.1.2) su visais $\beta_0, ..., \beta_m$ ir be $\beta_{j_1}, ..., \beta_{j_l}$. Kai H_0 teisinga, atsitiktinio dydžio $D_R^{(m)} - D_R^{(m-k)}$ skirstinys aproksimuojamae chi kvadrato skirstiniu su $k = m - (m-k)$ laisvės laipsnių (žr. I d. 4.5.4 skyrelį).

Hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, jei

$$D_R^{(m)} - D_R^{(m-k)} > \chi_\alpha^2(k).$$

Šis kriterijus gali būti naudojamas kovariančių sąveikos nebuvimo hipotezėms tikrinti. Pavyzdžiui, (5.2.5) modelyje ši hipotezė ekvivalenti hipotezei $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. Kriterijaus statistikos $D_R^{(7)} - D_R^{(3)}$ skirstinys aproksimuojamas chi kvadrato dėsniu su $k = 4$ laisvės laipsniais.

Hipotezė

$$H_j : \beta_j = 0 \ (j = 1, ..., m)$$

taip pat gali būti tikrinama naudojant Fišerio informacinės matricos įvertinį.

Jei n yra didelis, tai a. d. $\sqrt{n}(\hat{\beta}_j - \beta_j)$ dėsnis aproksimuojamas normaliuoju dėsniu $N(0, \sigma_{jj}(\boldsymbol{\beta}))$.

Statistikos

$$W_j = \sqrt{n}\frac{\hat{\beta}_j}{\sigma_{jj}(\hat{\boldsymbol{\beta}})}$$

skirstinys aproksimuojamas $N(0,1)$ dėsniu, kai n didelis. Hipotezė $H_0 : \beta_j = 0$ atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, jei $|W_j| > z_{\alpha/2}$.

5.6.1 pavyzdys (**5.3.1** pavyzdžio tęsinys). 5.3.1 pratimo sąlygomis patikrinsime hipotezę $H_0 : \beta_1 = ... = \beta_m = 0$ dėl a. d. Y prognozavimo remiantis kovariančių vektoriumi $\mathbf{X}$ prasmingumo, taip pat atskirų regresijos parametrų reikšmingumo hipotezes. Apskaičiuosime determinacijos koeficientų R^2 ir $\tilde{R}^2$ reikšmes.

Naudodamiesi 5.3.1 pavyzdyje gautu įverčiu $\hat{\boldsymbol{\beta}}$, randame

$$-\ln(L(\hat{\boldsymbol{\beta}})) = -\sum_{i=1}^{n}[Y_i \ln\frac{\hat{Y}_i}{1-\hat{Y}_i} + \ln(1-\hat{Y}_i)] = 24,3678,$$

$$-\ln(L(\hat{\pi})) = -n[\bar{Y}\ln\frac{\bar{Y}}{1-\bar{Y}} + \ln(1-\bar{Y})] = 40,7516.$$

Tada statistikos D_R realizacija yra

$$D_R = 2(\ln L(\hat{\boldsymbol{\beta}}) - \ln(L(\hat{\pi}))) = 32,7675,$$

o asimptotinė P-reikšmė $pv_a = \mathbf{P}\{\chi_5^2 > 32,7675\} = 4,2 \cdot 10^{-6}$. Hipotezė H_0 atmetama su aukštu reikšmingumo lygmeniu. Darome išvadą, kad kintamojo Y prognozavimas remiantis kovariančių vektoriumi $\mathbf{X}$ yra prasmingas.

Prognozės kokybės matai, t. y. determinacijos koeficientai, įgijo reikšmes $R^2 = 0,4020$, $\tilde{R}^2 = 0,5664$.

Tikrinsime gydymo, t. y. kovariantės X_1 reikšmingumo, hipotezę. Tuo tikslu reikia patikrinti hipotezę $H_{11,12} : \beta_{11} = \beta_{12} = 0$. Be $D_R = D_R^{(6)}$, papildomai apskaičiuojame D_R analogą D_R^{6-2} modelyje, kuriame praleista kovariantė X_1, t. y. $\beta_{11} = \beta_{12} = 0$. Gauname

statistikos $D_R^6 - D_R^{6-2}$ realizaciją 12,5310 ir asimptotinę P-reikšmę $pv_a = \mathbf{P}\{\chi_2^2 > 12,5310\} = 0,0019$. Darome išvadą, kad gydymas yra reikšminga kovariantė prognozuojant kintamąjį Y.

Tegu β_j yra kuris nors regresijos parametras. Parametro β_j reikšmingumo hipotezę $H_j : \beta_j = 0$ tikriname remdamiesi statistikos W_j aproksimacija normaliuoju skirstiniu, arba statistikos W_j^2 aproksimacija χ^2 skirstiniu su vienu laisvės laipsniu. Tarkime, w_j^2 yra statistikos W_j^2 realizacija. Tada hipotezė H_j atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai $pv_a = \mathbf{P}\{\chi_1^2 > w_j^2\} < \alpha$. Atliekant analizę SAS programa LOGISTIC automatiškai patikrinamos visų regresijos parametrų lygybės nuliui, atskirų kovariančių reikšmingumo ir atskirų regresijos parametrų reikšmingumo hipotezės. Pateikiamos statistikų realizacijos ir joms atitinkančios asimptotinės P-reikšmės. Pavyzdžiui, 5.3.2 lentelės stulpelyje „χ^2 statistika“ pateiktos realizacijos w_j^2, o priešpaskutiniame stulpelyje šias realizacijas atitinkančios asimptotinės P-reikšmės. Matome, kad visi parametrai statistiškai reikšmingi, išskyrus parametrą β_4. Tai galima aiškinti tuo, kad kovariantė X_4 yra gana subjektyvi ir mažai informatyvi (žr. kovariantės X_4 realizacijų sklaidą 5.3.1 lentelėje).

Tokioje situacijoje natūralu kovariantę X_4 praleisti ir nagrinėti logistinės regresijos modelį be šios kovariantės. Analizės rezultatai iš esmės nesiskiria. Vietoje 5.3.2 lentelės gauname tokią lentelę

5.6.1 lentelė. Parametrų įverčiai

Parametras	Įvertis	Paklaida	χ^2 statistika	P-reikšmė	Šansų santykis
β_0	15,8669	6,4056	6,1357	0,0132	
β_{11}	3,1790	1,0135	9,8375	0,0017	24,022
β_{12}	3,7284	1,1339	10,8006	0,0010	41,528
β_{21}	1,8235	0,7920	5,3013	0,0213	6,194
β_3	-0,2650	0,0959	7,6314	0,0057	0,767

Matome, kad šiame modelyje visi regresijos koeficientai statistiškai reikšmingai skiriasi nuo nulio. Determinacijos koeficientų įverčiai $R^2 = 0,4016, \tilde{R}^2 = 0,5660$. Fišerio atvirkštinės matricos $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ įvertis šiame modelyje

$$\frac{\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})}{n} = \begin{pmatrix} 41,0316 & 2,0248 & 3,7419 & 0,4158 & -0,6095 \\ 2,0248 & 1,0273 & 0,7228 & 0,3176 & -0,0387 \\ 3,7419 & 0,7228 & 1,2857 & 0,3226 & -0,0643 \\ 0,4158 & 0,3176 & 0,3226 & 0,6272 & -0,0128 \\ -0,6095 & -0,0387 & -0,0643 & -0,0128 & 0,0092 \end{pmatrix} = \left[\frac{\sigma_{ls}(\hat{\boldsymbol{\beta}})}{n}\right]_{5\times 5}.$$

5.3.7. Įvykio A tikimybės ir regresinių parametrų pasikliovimo intervalai

Remiantis 5.4 skyrelyje pateiktomis aproksimacijomis, tiesinio regresijos koeficientų darinio $\boldsymbol{c}^T\boldsymbol{\beta}$ asimptotinis pasikliovimo intervalas, kai pasikliovimo lygmuo $Q = 1 - 2\alpha$, yra

$$(\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - z_\alpha\sqrt{\boldsymbol{c}^T\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}/n};\ \ \boldsymbol{c}^T\hat{\boldsymbol{\beta}} + z_\alpha\sqrt{\boldsymbol{c}^T\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}/n}). \tag{5.3.18}$$

Taigi funkcijos

$$logit\boldsymbol{x} = \mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m = \boldsymbol{x}^T\boldsymbol{\beta},$$

kai kovariančių vektorius $\mathbf{X}$ įgijo reikšmę $\boldsymbol{x}$, asimptotinis pasikliovimo intervalas yra

$$(\underline{\mu}(\boldsymbol{x});\ \bar{\mu}(\boldsymbol{x})) = \left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}} - z_\alpha\sqrt{\boldsymbol{x}^T\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}/n};\ \ \boldsymbol{x}^T\hat{\boldsymbol{\beta}} + z_\alpha\sqrt{\boldsymbol{x}^T\boldsymbol{i}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}/n}\right). \tag{5.3.19}$$

Tarę, kad $\boldsymbol{c} = (0, ..., 0, 1, 0, ..., 0)^T$, kai 1 yra j-oje pozicijoje, gauname regresijos parametro β_j asimptotinį pasikliovimo intervalą

$$(\underline{\beta}_j;\ \bar{\beta}_j) = \left(\hat{\beta}_j - z_\alpha\sqrt{\sigma_{jj}(\hat{\boldsymbol{\beta}})/n};\ \ \hat{\beta}_j + z_\alpha\sqrt{\sigma_{jj}(\hat{\boldsymbol{\beta}})/n}\right). \tag{5.3.20}$$

Koeficientų β_j ir šansų santykių ryšys suteikia galimybę sudaryti pasikliovimo intervalus šansų santykiams. Pasikliovimo intervalų rėžiai šansų santykiams (5.2.1), (5.2.3) ir (5.2.6) yra atitinkamai

$$\exp\{\hat{\beta}_j \pm z_\alpha\sqrt{\sigma_{jj}(\hat{(}\boldsymbol{\beta}))}/\sqrt{n}\}, \quad \exp\{\hat{\beta}_{jl} \pm z_\alpha\sqrt{\sigma_{jl}(\hat{\boldsymbol{\beta}})}/\sqrt{n}\}$$

ir

$$\exp\{\hat{\beta}_1 + \hat{\beta}_3 x_2 \pm \quad z_\alpha\sqrt{\sigma_{11}(\hat{\boldsymbol{\beta}}) + 2x_2\sigma_{13}(\hat{\boldsymbol{\beta}}) + x_2^2\sigma_{33}(\hat{\boldsymbol{\beta}})}/\sqrt{n}\}.$$

Tikimybė $\mathbf{P}\{Y = 1|\boldsymbol{x}\} = \pi(\boldsymbol{x})$ yra monotoniškai didėjanti $\mu(\boldsymbol{x})$ funkcija. Todėl tikimybės pasikliovimo intervalas gaunamas tiesiai iš intervalo (5.7.2):

$$(\underline{\pi}(\boldsymbol{x});\ \bar{\pi}(\boldsymbol{x})) = \left(\frac{e^{\underline{\mu}(\boldsymbol{x})}}{1 + e^{\underline{\mu}(\boldsymbol{x})}};\ \ \frac{e^{\bar{\mu}(\boldsymbol{x})}}{1 + e^{\bar{\mu}(\boldsymbol{x})}}\right). \tag{5.3.21}$$

5.7.1 pavyzdys (**5.3.1** pavyzdžio tęsinys). 5.3.1 pavyzdžio sąlygomis rasime regresijos parametrų, šansų santykių ir tikimybės $\pi(\boldsymbol{x})$, kai $\mathbf{X} = \boldsymbol{x} = (1; (1,\ 0); 1; 70)^T$, asimptotinius pasikliovimo lygmens $Q = 0,95$ pasikliovimo intervalus (naudojame modelį be kovariantės X_4).

Regresijos parametrų ir šansų santykių pasikliovimo intervalai pateikti 5.7.1 lentelėje.

5.7.1 lentelė. Pasikliovimo intervalai

Parametras	Pasikliovimo	intervalai	Šansų	santykiai
β_0	3,3122	28,4216		
β_{11}	1,1924	5,1655	3,295	175,121
β_{12}	1,5040	5,9487	4,500	383,262
β_{21}	0,2713	3,3758	1,312	29,248
β_3	-0,4529	-0,0770	0,636	0,926

Remiantis (5.7.4) tikimybės $\pi(\boldsymbol{x})$ asimptotinis pasikliovimo intervalas yra

$$(\underline{\pi}(\boldsymbol{x});\ \bar{\pi}(\boldsymbol{x})) = (0,6506;\ \ 0,9824).$$

5.3.8. Klasifikavimo uždaviniai

Objektus klasifikuoti į tam tikras grupes reikia daugelyje praktinių situacijų. Pavyzdžiui, gamintojas, remdamasis gaminio, jo mazgų ir technologinio proceso charakteristikų matavimais, stengiasi suklasifikuoti gaminius į tokius, kurie bus reklamuoti garantiniu laikotarpiu, ir kurie nebus reklamuoti. Chirurgas, remdamasis savo patirtimi ir paciento simptomų matavimais, bando suklasifikuoti pacientus į tokius, kuriems chirurginis gydymas bus sėkmingas, ir tokius, kuriems jis bus nesėkmingas. Automobilių pardavėjas bando atskirti tokius

automobilius, kurie bus parduoti per metus priklausomai nuo gamybos šalies, gamybos metų, kilometražo, pardavimo kainos ir pan.

Detaliau klasifikavimo uždaviniai nagrinėjami 4 vadovėlio dalyje. Kadangi klasifikavimo taisyklė dažnai kuriama remiantis logistine regresija, trumpai tokie uždaviniai aptariami šiame skyrelyje.

Tarkime, kad objektas priklauso pirmai grupei, jeigu jį atitinkantis a. d. Y įgijo reikšmę 1, ir objektas priklauso nulinei grupei, jeigu a. d. Y įgijo reikšmę 0. Tegu a. d. η įgyja reikšmę 1, jeigu, remiantis klasifikavimo taisykle, objektą priskiriame pirmai grupei, ir reikšmę 0, jeigu objektą priskiriame nulinei grupei. Klasifikavimo taisyklės (klasifikatoriaus) tikslumą galima apibūdinti tikimybėmis

$$\alpha_{ij} = \mathbf{P}\{\eta = i | Y = j\}, \quad i, j = 0,\ 1, \tag{5.3.22}$$

jas surašykime į lentelę.

5.8.1 lentelė. Klasifikavimo tikslumo tikimybės

	$\eta = 1$	$\eta = 0$	$\sum$
$Y = 1$	α_{11}	α_{01}	1
$Y = 0$	α_{10}	α_{00}	1

Šioje lentelėje α_{11} ir α_{00} yra teisingų sprendimų tikimybės, o α_{10} ir α_{01} – klaidingų. Klasifikavimo taisyklė tuo geresnė, kuo didesnės tikimybės α_{11} ir α_{00} ir kuo mažesnės tikimybės α_{10} ir α_{01}.

Praktiniu požiūriu kur kas svarbesnės yra *aposteriorinės* klasifikavimo tikslumo tikimybės

$$\beta_{ji} = \mathbf{P}\{Y = j | \eta = i\} = \frac{\alpha_{ij}\omega_j}{\alpha_{i1}\omega_1 + \alpha_{i0}\omega_0}, \quad i, j = 0,\ 1; \tag{5.3.23}$$

čia $\omega_1 = \mathbf{P}\{Y = 1\}, \omega_2 = \mathbf{P}\{Y = 0\}$, $\omega_1 + \omega_0 = 1$, yra apriorinės klasių tikimybės. Tikimybės (5.3.23) apibūdina gautų po klasifikavimo grupių užterštumą kitos grupės elementais.

Pavyzdžiui, tegu gaminys yra geras, jei $Y = 1$, ir defektinis, jei $Y = 0$. Atliekama išleidžiamoji produkcijos kontrolė, po kurios vartotojui siunčiami tik gerais pripažinti gaminiai ($\eta = 1$). Vartotoją visų pirma domina, kokią jo gautos produkcijos dalį sudaro defektiniai gaminiai (tikimybė $\beta_{01} = \mathbf{P}\{Y = 0 | \eta = 1\}$), o ne kaip dažnai kontrolės metu defektinis gaminys pripažįstamas geru (tikimybė $\alpha_{10} = \mathbf{P}\{\eta = 1 | Y = 0\}$). Pavyzdžiui, jeigu $\omega_0 = 1$, tai kad ir kokia maža būtų tikimybė α_{10}, vartotojui pateks vien defektiniai gaminiai (kitokių ir nebuvo). Taigi tikimybės β_{01} mažumą nelemia vien α_{10} mažumas, bet tai priklauso ir nuo tikimybių ω_1, ω_0.

Tarkime, kad teisingas logistinės regresijos modelis

$$\mathbf{P}\{Y = 1 | \boldsymbol{x}\} = \pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}. \tag{5.3.24}$$

Tada objektą, kurio kovariančių vektorius $\mathbf{X}$ įgijo reikšmę $\boldsymbol{x}$, natūralu priskirti

pirmai grupei ($Y = 1$), jeigu įvykio $\{Y = 1\}$ šansas

$$\gamma(\boldsymbol{x}) = \frac{\mathbf{P}\{Y = 1|\boldsymbol{x}\}}{\mathbf{P}\{Y = 0|\boldsymbol{x}\}} = e^{\boldsymbol{\beta}^T\boldsymbol{x}} \geq 1.$$

Nelygybė $\gamma(\boldsymbol{x}) \geq 1$ ekvivalenti nelygybėms

$$\pi(\boldsymbol{x}) \geq \frac{1}{2}, \quad e^{\boldsymbol{\beta}^T\boldsymbol{x}} \geq 1, \quad \boldsymbol{\beta}^T\boldsymbol{x} \geq \ln 1 = 0.$$

Klasifikavimo taisyklę galima suformuluoti taip:
– objektas, kurio $\mathbf{X} = \boldsymbol{x}$, priskiriamas pirmai grupei, jei $\pi(\boldsymbol{x}) \geq 1/2$;
– objektas, kurio $\mathbf{X} = \boldsymbol{x}$, priskiriamas nulinei grupei, jei $\pi(\boldsymbol{x}) < 1/2$.

5.8.1 pastaba. Atsižvelgiant į praktinius poreikius, klasifikavimo taisyklės slenkstį galima keisti. Pavyzdžiui, gali būti pageidaujama, kad į pirmą grupę patektų objektai, kurių įvykio $\{Y = 1\}$ šansas ne mažesnis už 4, t. y. $\pi(\boldsymbol{x}) \geq 0,8$. Tada bendresniu atveju klasifikavimo taisyklę galima suformuluoti taip:
– objektas, kurio $\mathbf{X} = \boldsymbol{x}$, priskiriamas pirmai grupei, jei $\pi(\boldsymbol{x}) \geq z$;
– objektas, kurio $\mathbf{X} = \boldsymbol{x}$, priskiriamas nulinei grupei, jei $\pi(\boldsymbol{x}) < z$.
Suprantama, tokiu atveju 5.8.1 lentelės tikimybės priklausys nuo slenksčio z, $0 \leq z \leq 1$.

Tarkime, kad logistinės regresijos modelio (5.8.2) parametras $\boldsymbol{\beta}$ nežinomas ir buvo įvertintas naudojantis imtimi (5.3.1). Tada, remdamiesi parametro įverčiu $\hat{\boldsymbol{\beta}}$, galime suformuluoti įvertintą klasifikavimo taisyklę:
– objektas, kurio $\mathbf{X} = \boldsymbol{x}$, priskiriamas pirmai grupei, jei $\hat{\pi}(\boldsymbol{x}) \geq z$;
– objektas, kurio $\mathbf{X} = \boldsymbol{x}$, priskiriamas nulinei grupei, jei $\hat{\pi}(\boldsymbol{x}) < z$.
Pritaikykime šią taisyklę imties (5.3.1) elementams klasifikuoti. Tegu n_1 yra imties elementų, kurių $Y = 1$, skaičius (pažymėkime C_1 šių elementų indeksų aibę), o n_0 yra imties elementų, kurių $Y = 0$, skaičius (pažymėkime C_0 šių elementų indeksų aibę); $n_1 + n_0 = n$. Naudodami įverčius $\hat{\pi}(\boldsymbol{x}^{(i)})$ gausime

$$V_{11}(z) = \sum_{i\in C_1} \mathbf{1}_{[z,1]}(\hat{\pi}(\boldsymbol{x}^{(i)})),\ V_{01}(z) = \sum_{i\in C_1} \mathbf{1}_{[0,z)}(\hat{\pi}(\boldsymbol{x}^{(i)})),\ V_{11}(z)+V_{01}(z) = n_1,$$

$$V_{10}(z) = \sum_{i\in C_2} \mathbf{1}_{[z,1]}(\hat{\pi}(\boldsymbol{x}^{(i)})),\ V_{00}(z) = \sum_{i\in C_2} \mathbf{1}_{[0,z)}(\hat{\pi}(\boldsymbol{x}^{(i)})),\ V_{10}(z)+V_{00}(z) = n_0,$$

imties elementų suskaidymą į keturias aibes.

5.8.2 lentelė. Imties elementų klasifikacija

	$\eta = 1$	$\eta = 0$	$\sum$
$Y = 1$	$V_{11}(z)$	$V_{01}(z)$	n_1
$Y = 0$	$V_{10}(z)$	$V_{00}(z)$	n_0
$\sum$	$V_{11}(z) + V_{10}(z)$	$V_{01}(z) + V_{00}(z)$	n

Čia $V_{11}(z), V_{00}(z)$ skaičiai objektų, kurie buvo suskirstyti į grupes teisingai, o $V_{10}(z), V_{01}(z)$ – neteisingai. Remdamiesi šiais skaičiais gauname 5.8.1 lentelės

tikimybių $\alpha_{ij}(z)$ įverčius

$$\hat{\alpha}_{ij}(z) = \frac{V_{ij}(z)}{n_j}, \quad i,j = 0,\ 1. \tag{5.3.25}$$

Jeigu objektai į imtį parenkami nepriklausomai vienas nuo kito ir atsitiktinai iš visos tokių objektų populiacijos, tai apriorines klasių tikimybes galima įvertinti santykiniais dažniais

$$\hat{\omega}_0 = \frac{n_0}{n}, \quad \hat{\omega}_1 = \frac{n_1}{n}.$$

Tada aposteriorinių klasifikavimo tikslumo tikimybių įverčiai

$$\hat{\beta}_{ji}(z) = \frac{\hat{\alpha}_{ij}(z)\hat{\omega}_j}{\hat{\alpha}_{i1}(z)\hat{\omega}_1 + \hat{\alpha}_{i0}(z)\hat{\omega}_0} = \frac{\frac{V_{ij}(z)}{n_j}\frac{n_j}{n}}{\frac{V_{i1}(z)}{n_1}\frac{n_1}{n} + \frac{V_{i0}(z)}{n_0}\frac{n_0}{n}}$$

$$= \frac{V_{ij}(z)}{V_{i1}(z) + V_{i0}(z)}, \quad i,j = 0,\ 1. \tag{5.3.26}$$

5.8.2 pastaba. Klasifikavimo tikslumo vertinimas remiantis imtimi (5.3.1) nėra visai korektiškas. Klasifikavimo taisyklei rasti naudojami tie patys stebėjimai, kurie naudojami ir klasifikavimo tikslumui vertinti. Jeigu imtis pakankamai didelė, rekomenduojama suskaidyti ją į dvi dalis. Vieną iš jų naudojame klasifikatoriui kurti (t. y. parametrui $\boldsymbol{\beta}$ vertinti), o kitą (testinė aibė) jo tikslumui vertinti.

Atliekant duomenų analizę SAS procedūra LOGISTIC ir nurodžius norimą slenksčių z rinkinį, yra pateikiama klasifikavimo lentelė. Joje duodami skaičiai $V_{11}(z)$, $V_{00}(z)$, $V_{10}(z)$, $V_{01}(z)$; dalis teisingai suklasifikuotų imties elementų $(V_{11}(z) + V_{00}(z))/n$ (procentais); įvertis $\hat{\alpha}_{11}(z)$ (stulpelio pavadinimas *Sensitivity*); įvertis $\hat{\alpha}_{00}(z)$ (*Specificity*); (5.8.5) įverčiai $\hat{\beta}_{10}(z)$ (*False POS*), $\hat{\beta}_{01}(z)$ (*False NEG*). Pageidaujant nubraižoma vadinamoji *ROC* kreivė, kurioje pavaizduojama $\hat{\alpha}_{11}(z)$ priklausomybė nuo $1 - \hat{\alpha}_{00}(z)$, kai z perbėga intervalą [0, 1]. Klasifikavimas tuo geresnis, kuo staigiau kreivė auga nuo 0 iki 1.

5.8.1 pavyzdys. (**5.3.1** pavyzdžio tęsinys). 5.3.1 pavyzdžio sąlygomis atliksime imties (5.3.1) klasifikavimą į dvi grupes remdamiesi logistinės regresijos modeliu be kovariantės X_4.

Naudodami SAS procedūrą LOGISTIC gauname klasifikavimo lentelę. Pateikiame dalį lentelės, kai slenkstis $z = 0,4;\ 0,5;\ 0,6$.

5.8.3 lentelė. Klasifikavimo lentelė.

z	$V_{11}(z)$	$V_{00}(z)$	$V_{10}(z)$	$V_{01}(z)$	$\hat{\alpha}_{11}(z)$	$\hat{\alpha}_{00}(z)$	$\hat{\beta}_{10}(z)$	$\hat{\beta}_{01}(z)$
0,4	32	16	9	3	91,4	64,0	22,0	15,8
0,5	30	19	6	5	85,7	76,0	16,7	20,8
0,6	27	20	5	8	77,1	80,0	15,6	28,6

5.4. Pratimai

5.1. Firmoje užregistruoti klientų telefono skambučių skaičiai per kiekvieną iš 7 darbo valandų (kovariantė X_1) kiekvienai iš 5 savaitės darbo dienų (kovariantė X_2). Toks pat eksperimentas pakartotas kitą savaitę. Skambučių skaičiaus Y_{ijk}, $i = 1, ..., 7$, $j = 1, ..., 5$, $k = 1, 2$ stebiniai pateikti lentelėje.

Z_{1i}	Z_{2j}										$\sum$
	-2		-1		0		1		-2		
-3	30	44	30	36	26	30	31	31	18	43	325
-2	29	34	31	36	22	35	18	30	25	31	291
-1	28	41	22	24	23	26	21	29	26	28	268
0	23	24	19	24	23	31	20	25	21	31	241
1	30	30	32	40	26	33	26	34	26	36	313
2	30	38	28	40	36	37	23	25	20	24	301
3	34	39	24	41	25	34	21	26	25	41	310
$\sum$	454		427		407		366		395		2049

Lentelėje pateiktos centruotos kovariančių reikšmės $Z_{1i} = X_{1i} - 4, i = 1, ..., 7$; $Z_{2j} = X_{2j} - 3, j = 1, ..., 5$.

Tarkime, kad skambučių srautas yra puasoninis su pastoviu intensyvumu valandos bėgyje, t.y. $Y_{ijk} \sim \mathcal{P}(\lambda_{ij})$.

Atlikite imties

$$(Y_{111}, Z_{11}, Z_{21}), (Y_{112}, Z_{11}, Z_{21}), ..., (Y_{771}, Z_{17}, Z_{27}), (Y_{772}, Z_{17}, Z_{27})$$

puasoninę regresinę analizę. Iš paskutinio lentelės stulpelio matyti, kad skambučių skaičiaus kitimas dienos bėgyje nėra tiesinis, todėl Y_{ijk} priklausomybei nuo kovariančių apibūdinti naudokite tokį modelį (žr. (5.2.2)):

$$\mu_{ij} = \mathbf{E}Y_{ijk} = \dot{B}(\boldsymbol{\beta}^T \boldsymbol{Z}_{ij}) = e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ij}} = e^{\beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{1i}^2 + \beta_3 Z_{2j}},$$

$$i = 1, ..., 7, \quad j = 1, ..., 5, \quad k = 1, 2.$$

a) Raskite parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ DT įvertį $\hat{\boldsymbol{\beta}}$ ir kovariacinės matricos $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$ įvertį $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})$.

b) Patikrinkite hipotezes $H_j : \beta_j = 0, j = 1, 2, 3$.

c) Raskite trečios savaitės darbo dienos pirmos ir ketvirtos valandos vidutinio skambučių skaičiaus taškinius ir intervalinius ($Q = 0,95$) įverčius.

5.2 (**5.1** pratimo tęsinys). Atlikite 5.1 pratimo užduotis a), b), c) tardami, kad teisingas normaliosios teorijos tiesinis regresijos modelis:

$$Y_{ijk} = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{1i}^2 + \alpha_3 Z_{2i} + e_{ijk},$$

čia e_{ijk} nepriklausomi normalieji a. d. $e_{ijk} \sim N(0, \sigma^2)$. Palyginkite 5.1 ir 5.2 pratimų atsakymus.

5.3 (**5.1** pratimo tęsinys). Atlikite 5.1 pratimo stebinių dispersiją stabilizuojančią transformaciją $U_{ijk} = \sqrt{4Y_{ijk} - 1}$. Kai λ_{ij} didelis, a. d. U_{ijk} skirstinys aproksimuojamas normaliuoju su vienetine dispersija. Atlikite 5.1 pratimo užduotis a), b), c) tardami, kad teisingas tiesinis regresijos modelis:

$$U_{ijk} = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{1i}^2 + \gamma_3 Z_{2i} + e_{ijk},$$

čia e_{ijk} nepriklausomi normalieji a. d. $e_{ijk} \sim N(0,\ 1)$. Palyginkite 5.1, 5.2 ir 5.3 pratimų atsakymus.

5.4. Tiriant gaminių patikimumą iš 4 įmonių (kovariantė X) pagamintos produkcijos atsitiktinai atrinkta ir išbandyta po 20 gaminių. Gaminių darbo laiko iki gedimo Y_{ij} stebiniai pateikti lentelėje.

Įmonė	Y_{ij}									
I	65,10	74,80	25,11	69,89	28,73	13,27	49,60	26,96	30,03	16,46
II	7,98	31,27	29,81	51,75	18,61	7,07	13,60	27,30	5,56	17,50
III	27,17	13,64	10,66	15,59	26,56	17,76	26,11	13,32	14,40	20,80
IV	15,55	30,47	14,16	24,02	9,09	13,82	18,68	16,73	19,69	21,54

Įmonė	Y_{ij}									
I	38,72	59,69	80,03	86,27	19,77	37,36	36,30	37,98	60,31	51,45
II	19,99	53,88	21,94	32,27	32,34	7,28	17,87	7,42	17,17	11,66
III	20,82	30,93	29,28	15,57	35,78	33,10	32,44	15,84	17,11	16,04
IV	27,46	42,90	11,62	13,66	11,54	16,33	18,75	13,14	36,04	24,92

Tarkime, kad gaminio darbo laikas Y_{ij} turi gama skirstinį $Y_{ij} \sim G(\lambda_j,\ 3)$. Kadangi kovariantė X nominali, tai ją koduokime keisdami vektoriumi $\boldsymbol{Z} = (Z_1, Z_2, Z_3)^T$, kuris įgyja reikšmes $(1,\ 0,\ 0)^T$, $(0,\ 1,\ 0)^T$, $(0,\ 0,\ 1)^T$ ir $(0,\ 0,\ 0)^T$ atitinkamai pirmajai, antrajai, trečiajai ir ketvirtajai įmonei.

Atlikite imties

$$(Y_{11}, \boldsymbol{Z}_{11}), ..., (Y_{20,4}, Z_{20,4})$$

gama regresinę analizę. Priklausomybei nuo kovariantės X apibūdinti naudokite tokį modelį:

$$\mu_j = \mathbf{E}Y_{ij} = 3e^{\beta_0+\beta_j}, j = 1, 2, 3; \quad \mu_4 = \mathbf{E}Y_{i4} = 3e^{\beta_0}, \quad i = 1, ..., 20.$$

a) Raskite parametro $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ DT įvertį $\hat{\boldsymbol{\beta}}$ ir kovariacinės matricos $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$ įvertį $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})$.

b) Patikrinkite hipotezes $H_j : \beta_j = 0,\ j = 1, 2, 3$ ir hipotezę $H_{23} : \beta_2 = \beta_3$.

c) Raskite trečios įmonės pagaminto gaminio darbo laiko vidurkio taškinį ir intervalinį $(Q = 0,95)$ įverčius.

5.5. Turime 24 studentų įskaitos duomenis. Priklausomas kintamasis $Y = 1$, jeigu studentas gavo įskaitą, ir $Y = 0$ – jeigu negavo. Kiek valandų studentas dirbo pratybų metu, rodo kintamasis X_2. Ar studentas prieš pat sesiją ko nors klausė dėstytojo, rodo kintamasis X_1 ($X_1 = 1$, jeigu klausė, ir $X_1 = 0$, jeigu neklausė). Duomenys pateikti lentelėje ([4],II dalis).

Y	0	0	0	0	0	0	0	0	0	0	0	1
X_1	1	1	1	1	1	0	0	0	0	0	0	1
X_2	19	17	13	15	19	21	17	18	23	15	13	26
Y	1	1	1	1	1	1	1	1	1	1	1	1
X_1	1	1	0	0	0	0	0	0	0	1	1	1
X_2	30	19	22	21	24	28	30	27	21	24	20	16

a) Įvertinkite logistinės regresijos parametrus.

b) Įvertinkite kovariančių X_1 ir X_2 šansų santykius.

c) Įvertinkite tikimybę gauti įskaitą, kai $X_2 = 20$, $X_1 = 0$ ir kai $X_2 = 20$, $X_1 = 1$.

d) Sudarykite klasifikacinę lentelę, kai objektas priskiriamas tai klasei, kurios tikimybės įvertis didesnis.

5.6. Gimdymo namuose surinkti duomenys apie gimdyvės amžių X_1, rūkymą ($X_2 = 1$ – rūko, $X_2 = 0$ – nerūko), hipertoniją X_3 ($X_3 = 1$ – serga, $X_3 = 0$ – neserga), moters svorį X_4 ir naujagimio svorį Z. Naujagimis sveria nepakankamai, jeigu jo svoris nesiekia 2500 g ([4], II dalis).

X_1	X_2	X_3	X_4	Z	X_1	X_2	X_3	X_4	Z
24	0	0	64,0	1703	29	0	0	75,0	2922
21	1	1	82,5	1792	26	1	0	84,0	2922
21	0	0	100,0	1930	17	0	0	56,5	2922
19	0	0	51,0	2084	35	1	0	60,5	2950
24	0	0	69,0	2102	33	1	0	54,5	3035
17	1	0	55,0	2227	21	1	0	92,5	3044
18	0	0	74,0	2284	19	0	0	94,5	3064
15	0	0	57,5	2383	21	0	0	80,0	3064
17	0	0	60,0	2440	19	0	0	57,5	3177
20	0	0	52,5	2452	28	0	0	70,0	3236
14	1	0	50,5	2468	16	1	0	67,5	3376
14	0	0	50,0	2497	22	0	0	65,5	3462
21	1	1	65,0	2497	32	0	0	85,0	3475
33	0	0	77,5	2553	19	0	0	52,5	3574
32	0	0	60,5	2837	24	0	0	55,0	3730
28	0	0	83,5	2879	25	0	1	60,0	3985

a) Tarkime $Y = 1$, jeigu naujagimio svoris mažesnis už 2500 g, ir $Y = 0$ – priešingu atveju. Įvertinkite logistinės regresijos parametrus prognozuodami kintamąjį Y pagal X_1, X_2, X_3, X_4.

b) Patikrinkite hipotezes dėl regresijos parametrų reikšmingumo.

5.7. Ar galima pagal pajamas (kintamasis X_1) ir darbo prestižiškumo indeksą (kintamasis X_2) atpažinti, kad respondentas turi aukštąjį išsilavinimą ($Y = 1$ – jeigu turi ir $Y = 0$ – jeigu neturi)? Duomenys pateikti lentelėje ([4], II dalis).

X_1	3670	1923	3067	3811	3494	2012	1637	1265	2722
X_2	60	65	70	105	70	55	55	35	105
Y	1	0	1	1	1	0	0	0	0
X_1	4050	1501	3340	3193	3125	4050	3458	2219	3781
X_2	135	50	65	60	95	115	95	95	90
Y	1	0	1	1	0	1	0	0	1
X_1	2736	2568	3408	3298	3043	3536	3780	3798	
X_2	85	135	110	60	95	80	94	78	
Y	0	0	0	1	1	1	1	1	

Atlikite duomenų logistinę regresiją.

5.8. Lentelėje pateikti dviejų futbolo lygų (kintamasis Z) duomenys: atstumas iki vartų (X), bandymų įmušti įvartį skaičius (N), sėkmių (įvykių) skaičius (M) [2].

Lyga Z	Atstumas X	Sėkmių skaičius M	Bandymų skaičius N
0	14,5	68	77
0	24,5	74	95
0	34,5	61	113
0	44,5	38	138
0	52,0	2	38
1	14,5	62	67
1	24,5	49	70
1	34,5	43	79
1	44,5	25	82
1	52,0	7	24

a) Nagrinėdami lygas atskirai, parinkite logistinės regresijos modelį, kuriame kovariantė yra atstumas.

b) Parinkite logistinės regresijos modelį, kuriame kovariantės yra atstumas ir lyga.

5.5. Atsakymai ir nurodymai

5.1. a) $\hat\beta_0 = 3,2975$, $\hat\beta_1 = 0,0023$, $\hat\beta_2 = 0,0188$, $\hat\beta_3 = -0,0437$;

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = [c_{ij}(\hat{\boldsymbol{\beta}})]_{4\times 4} = 10^{-3}\begin{pmatrix} 1,198 & 0 & -0,168 & 0,021 \\ 0 & 0,115 & 0 & 0 \\ -0,168 & 0 & 0,040 & 0 \\ 0,021 & 0 & 0 & 0,245 \end{pmatrix}.$$

b) Tikrinant hipotezes $H_j : \beta_j = 0, j = 1,\ 2,\ 3$, statistikų (5.1.12) kvadratai įgijo reikšmes 0,05, 8,89, 7,81; atitinkamos P reikšmės 0,8299, 0,0029, 0,0052. **c)** $\hat\mu_{13} = e^{\hat\beta_0 - 3\hat\beta_1 + 9\hat\beta_2} = 31,80$; $\hat\mu_{43} = e^{\hat\beta_0} = 27,04$; $(\underline{\mu}_{13};\ \bar\mu_{13}) = (28,86;\ 35,04)$; $(\underline{\mu}_{43};\ \bar\mu_{43}) = (25,27;\ 28,94)$. **Nurodymas.** Taškinis įvertis $\hat\mu_{ij} = e^{\hat\theta_{ij}}, \hat\theta_{ij} = \hat{\boldsymbol{\beta}}^T\boldsymbol{Z}_{ij}$. Tiesinio darinio $\theta_{ij} = \boldsymbol{\beta}^T\boldsymbol{Z}_{ij}$ pasikliovimo lygmens Q pasikliovimo intervalą gauname naudodami statistikos $(\hat\theta_{ij} - \theta_{ij})/\sqrt{\hat{\boldsymbol{V}}(\hat\theta_{ij})}$, $\hat{\boldsymbol{V}}(\hat\theta_{ij}) = \boldsymbol{Z}_{ij}^T\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})\boldsymbol{Z}_{ij}$, aproksimaciją standartiniu normaliuoju skirstiniu:

$$(\underline{\theta}_{ij}, \bar\theta_{ij}) = (\hat\theta - z_P\sqrt{\hat{\boldsymbol{V}}(\hat\theta_{ij})}, \hat\theta + z_P\sqrt{\hat{\boldsymbol{V}}(\hat\theta_{ij})}), \quad P = (1-Q)/2.$$

Tada $\underline{\mu}_{ij} = e^{\underline{\theta}_{ij}}$, $\bar\mu_{ij} = e^{\bar\theta_{ij}}$. **5.2. a)** $\hat\alpha_0 = 27,043$, $\hat\alpha_1 = 0,0714$, $\hat\alpha_2 = 0,5571$, $\hat\alpha_3 = -1,279$, $\hat\sigma^2 = s^2 = SS_E/(n-4) = 37,891$;

$$\hat{\boldsymbol{V}}(\hat\alpha) = \begin{pmatrix} 1,2630 & 0 & -0,1804 & 0 \\ 0 & 0,1353 & 0 & 0 \\ -0,1804 & 0 & 0,0451 & 0 \\ 0 & 0 & 0 & 0,2707 \end{pmatrix}.$$

b) Tikrinant hipotezes $H_j : \alpha_j = 0, j = 1,\ 2,\ 3$, statistikos (3.3.16) įgijo reikšmes 0,19, 2,62, -2,46; atitinkamos P reikšmės 0,8466, 0,0108, 0,0166. **c)** $\hat\mu_{13} = \hat\alpha_0 - 3\hat\alpha_1 + 9\hat\alpha_2 = 31,84$; $\hat\mu_{43} = \hat\alpha_0 = 27,04$; $(\underline{\mu}_{13};\ \bar\mu_{13}) = (28,45;\ 35,24)$; $(\underline{\mu}_{43};\ \bar\mu_{43}) = (24,80;\ 29,29)$. **5.3. a)** $\hat\gamma_0 = 10,31$, $\hat\gamma_1 = 0,0141$, $\hat\gamma_2 = 0,0998$, $\hat\gamma_3 = -0,2409$, $\hat\sigma^2 = s^2 = SS_E/(n-4) = 1,295$;

$$\hat{\boldsymbol{V}}(\hat\gamma) = \begin{pmatrix} 0,04317 & 0 & -0,00617 & 0 \\ 0 & 0,00463 & 0 & 0 \\ -0,00617 & 0 & 0,00154 & 0 \\ 0 & 0 & 0 & 0,00925 \end{pmatrix}.$$

b) Tikrinant hipotezes $H_j : \gamma_j = 0, j = 1,\ 2,\ 3$, statistikos (3.3.16) įgijo reikšmes 0,21, 2,54, -2,50; atitinkamos P reikšmės 0,8362, 0,0134, 0,0147. **c)** $\hat\mu_{13} = 31,14$; $\hat\mu_{43} = 26,57$; $(\underline{\mu}_{13};\ \bar\mu_{13}) = (27,77;\ 34,75)$; $(\underline{\mu}_{43};\ \bar\mu_{43}) = (24,45;\ 28,73)$. **5.4. a)** $\hat\beta_0 = 1,897$, $\hat\beta_1 = 0,8193$, $\hat\beta_2 = 0,0773$, $\hat\beta_3 = 0,0788$; $\boldsymbol{V}(\hat\beta_0) = 1/60$, $\boldsymbol{V}(\hat\beta_j) = 1/30$, $j = 1,\ 2,\ 3$; $\mathbf{Cov}\,(\hat\beta_0, \hat\beta_j) = -1/60$, $j = 1,\ 2,\ 3$, $\mathbf{Cov}\,(\hat\beta_j, \hat\beta_l) = 1/60$, $j \neq l = 1,\ 2,\ 3$. **b)** Tikrinant hipotezę $H : \beta_1 = \beta_2 = \beta_3 = 0$, tikėtinumų santykio statistika

$$D_R = -120\,(8\ln 2 + \ln(T_1\,T_2\,T_3\,T_4/T^4)),$$

čia $T_j = \sum_{i=1}^{20} Y_{ij}$, $j = 1, 2, 3, 4$; $T = T_1 + T_2 + T_3 + T_4 = 907,83 + 432,27 + 432,92 + 400,11 = 2173,13$, įgijo rekšmę 29,77; P reikšmė $pv = \mathbf{P}\{\chi_3^2 > 8,53\} = 1,54\times 10^{-6}$. Tikrinant hipotezes $H_j : \beta_j = 0$ statistikos $30\hat\beta_j^2$ įgijo reikšmes 20,14, 0,179, 0,186; atitinkamos P reikšmės $7,20\times 10^{-6}$, 0,6720, 0,6660. Tikrinant hipotezę $H_{23} : \beta_2 = \beta_3$ parametro $\theta = \beta_2 - \beta_3$ įvertinio $\hat\theta$ dispersija $\boldsymbol{V}\hat\theta = \boldsymbol{V}(\hat\beta_2 - \hat\beta_3) = 1/30$; statistika $30\hat\theta^2 = 6,8\times 10^{-5}$; P reikšmė 0,9934. **c)** $\hat\mu_3 = T_3/20 = 21,65$; $(\underline{\mu_3};\ \overline{\mu_3}) = (6T_3/\chi^2_{0,025}(120);\ 6T_3/\chi^2_{0,975}(120)) = (17,07;\ 28,37)$. **5.5. a)** $\hat\beta_0 = -12,371$, $\hat\beta_1 = 1,459$, $\hat\beta_2 = 0,590$; Tikrinant hipotezes $H_j : \beta_j = 0, j = 1,\ 2$, statistikų W_j kvadratai įgijo reikšmes 1,125, 4,978; atitinkamos P reikšmės 0,2889, 0,0257. **b)** $\exp(\hat\beta_1) = 4,302$, $\exp(\hat\beta_2) = 1,803$. **c)** $\pi(X_2 = 20, X_1 = 0) = 0,3593$, $\pi(X_2 = 20, X_1 = 1) = 0,7070$. **d)** $V_{11}(0,5) = 9$, $V_{01}(0,5) = V_{10}(0,5) = 4$, $V_{00}(0,5) = 7$. **5.6. a)** $\hat\beta_0 = 6,549$, $\hat\beta_1 = -0,303$, $\hat\beta_2 = -0,327$; $\hat\beta_3 = 1,741$; $\hat\beta_4 = -0,009$. **b)** Tikrinant hipotezes

$H_j : \beta_j = 0, j = 1, 2, 3, 4$, statistikų W_j kvadratai įgijo reikšmes 4,987, 0,082, 1,372, 0,071; atitinkamos P reikšmės 0,0255, 0,7742, 0,2415, 0,7899. **5.7.** $\hat{\beta}_0 = -1,721$, $\hat{\beta}_1 = 0,00117$, $\hat{\beta}_2 = -0,0199$. Tikrinant hipotezes $H_j : \beta_j = 0, j = 1, 2$, statistikų W_j kvadratai įgijo reikšmes 3,211, 0,921; atitinkamos P reikšmės 0,0731, 0,3372. **5.8. a)** Kai $Z = 0$, gauta $\hat{\beta}_0 = 3,95$, $\hat{\beta}_1 = -0,112$; $exp(\hat{\beta}_1) = 0,894$; tikrinant hipotezę $H_1 : \beta_1 = 0$, statistikos W_1 kvadratas įgijo reikšmę 102,6, todėl hipotezė atmetama kriterijumi su gana aukštu reikšmingumo. Kai $Z = 0$, gauta $\hat{\beta}_0 = 3,33$, $\hat{\beta}_1 = -0,091$; $exp(\hat{\beta}_1) = 0,913$; tikrinant hipotezę $H_1 : \beta_1 = 0$, statistikos W_1 kvadratas įgijo reikšmę 58,8, todėl hipotezė atmetama kriterijumi su gana aukštu reikšmingumo. **b)** $\hat{\beta}_0 = 3,63$, $\hat{\beta}_1 = -0,103$ (atstumas), $\hat{\beta}_2 = 0,1036$ (lyga); Tikrinant hipotezę $H_1 : \beta_1 = 0$, statistikos W_1 kvadratas įgijo reikšmę 161,6, todėl hipotezė atmetama kriterijumi su gana aukštu reikšmingumo. Tikrinant hipotezę $H_2 : \beta_2 = 0$, statistikos W_2 kvadratas įgijo reikšmę 0,372, atitinkama P reikšmė 0,5419.

6 skyrius

1 Priedas. Tiesinės algebros elementai

Nagrinėjant tiesinius modelius yra patogu naudoti vektorinius ir matricinius žymenis. Tai gerokai supaprastina formules ir padeda geriau suvokti dėstomą medžiagą. Reikalingos matematinėje statistikoje tiesinės algebros žinios yra susistemintos knygoje [12], joje yra ir kompaktiški pateikiamų faktų įrodymai. Šiame priede pateikiama tiesinės algebros faktų, kuriais remiamasi šioje knygoje.

6.1. Vektoriai

1P.1 apibrėžimas. Dimensijos k vektoriumi (vektoriumi stulpeliu) $\boldsymbol{x}$ vadinsime stulpelį, kurio elementai $x_1, ..., x_k$ yra realūs skaičiai, t.y. vektorių traktuosime kaip k-matės Euklido erdvės $\mathbf{R}^k$ elementą. Transponuotą vektorių (vektorių eilutę) žymėsime $\boldsymbol{x}^T = (x_1, ..., x_k)$.

1. *Vektorių sudėtis.* Vienodos dimensijos vektorių $\boldsymbol{x}^T = (x_1, ..., x_k)$ ir $\boldsymbol{y}^T = (y_1, ..., y_k)$ suma yra vektorius $(\boldsymbol{x} + \boldsymbol{y})^T = (x_1 + y_1, ..., x_k + y_k)$, kurio elementai gaunami sudedant atitinkamus dėmenų elementus. Sumos operacija tenkina komutatyvumo

$$\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x} \tag{6.1.1}$$

ir distributyvumo

$$(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) \tag{6.1.2}$$

dėsnius. Egzistuoja nulinis $\mathbf{0}^T = (0, ..., 0)$ ir atvirkštinis $(-\boldsymbol{x})^T = (-x_1, ..., -x_k)$ vektoriai, kad

$$\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}, \quad \boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}. \tag{6.1.3}$$

2. *Daugyba iš skaliaro.* Vektoriaus $\boldsymbol{x}^T = (x_1, ..., x_k)$ sandauga iš skaliaro $c \in \mathbf{R}$ suprantamas vektorius $(c\boldsymbol{x})^T = (cx_1, ..., cx_k)$, kuris gaunamas padauginant iš c visas vektoriaus $\boldsymbol{x}$ koordinates. Šis veiksmas tenkina distributyvumo dėsnius

$$c(\boldsymbol{x} + \boldsymbol{y}) = c\boldsymbol{x} + c\boldsymbol{y}, \quad (c_1 + c_2)\boldsymbol{x} = c_1\boldsymbol{x} + c_2\boldsymbol{x}, \tag{6.1.4}$$

ir asociatyvumo dėsnį

$$c_1(c_2\boldsymbol{x}) = (c_1c_2)\boldsymbol{x}. \tag{6.1.5}$$

1P.2 apibrėžimas. Dimensijos k vektorių, kuriems apibrėžtos sudėties ir daugybos iš skaliaro operacijos, visuma vadinama *tiesine k-mate Euklido erdve* $\mathbf{R}^k$.

1P.3 apibrėžimas. Erdvės $\mathbf{R}^k$ poaibis $\mathcal{M}$, uždaras sudėties ir daugybos iš skaliaro operacijų atžvilgiu, t.y. jei $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, tai $c\boldsymbol{x} + d\boldsymbol{y} \in \mathcal{M}$ su visais $c, d \in \mathbf{R}$, vadinamas tiesiniu erdvės $\mathbf{R}^k$ *poerdviu.* Suprantama, kad kiekvienas tiesinis poerdvis yra tiesinė vektorinė erdvė.

Pavyzdžiui, aibė, susidedanti iš nulinio vektoriaus $\mathbf{0}$, arba visų vektorių aibė yra tiesiniai poerdviai. Imdami visus galimus aibės $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ vektorių tiesinius darinius

$$c_1\boldsymbol{x}_1 + ... + c_m\boldsymbol{x}_m, \quad c_1, ..., c_m \in \mathbf{R},$$

gausime tiesinį poerdvį $\mathcal{M}(S)$, generuotą vektorių aibės S.

1P.4 apibrėžimas. Tariame, kad vektoriai $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ yra *tiesiškai priklausomi*, jei egzistuoja skaliarai $c_1, ..., c_m \in \mathbf{R}$, ne visi lygūs 0, kad būtų patenkinta sąlyga

$$c_1\boldsymbol{x}_1 + ... + c_m\boldsymbol{x}_m = \mathbf{0}. \tag{6.1.6}$$

Jeigu tokių skaliarų neegzistuoja, sakoma, kad vektoriai $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ yra *tiesiškai nepriklausomi*. Iš šio apibrėžimo išplaukia tokios išvados: 1) Nulinis vektorius sudaro aibę tiesiškai priklausomų vektorių.

2) Kiekviena vektorių aibė, kuriai priklauso nulinis vektorius $\mathbf{0}$, yra tiesiškai priklausomų vektorių aibė.

3) Aibė nenulinių vektorių $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ yra tiesiškai priklausoma tada ir tik tada, kai kuris nors vektorius yra tiesinis kitų vektorių darinys.

1P.5 apibrėžimas. Tiesinės vektorinės erdvės $\mathcal{M}$ poaibis, kuris generuoja tiesinę erdvę $\mathcal{M}$, vadinamas tiesinės vektorinės erdvės $\mathcal{M}$ *baze*.

1) Kiekviena vektorinė erdvė turi bazę.

2) Jeigu $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ ir $\boldsymbol{b}_1, ..., \boldsymbol{b}_r$ yra dvi tos pačios erdvės bazės, tai $m = r$.

3) Kiekvienas erdvės $\mathcal{M}$ vektorius vienintelių būdu išreiškiamas jos baze. Pavyzdžiui, k-matės erdvės $\mathbf{R}^k$ elementai $\boldsymbol{e}_1 = (1,\ 0, ..., 0)^T$, $\boldsymbol{e}_2 = (0,\ 1, ..., 0)^T$, ..., $\boldsymbol{e}_k = (0,\ 0, ..., 1)^T$, sudaro bazę, nes jie yra tiesiškai nepriklausomi ir kiekvienas vektorius $\boldsymbol{x} = (x_1, ..., x_k)^T \in \mathbf{R}^k$ išreiškiamas šiais vektoriais:

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + ... + x_k\boldsymbol{e}_k.$$

Vektorių skaliarinė sandauga. Dviejų vienodos dimensijos vektorių $\boldsymbol{x} = (x_1, ..., x_k)^T$ ir $\boldsymbol{y} = (y_1, ..., y_k)^T$ *skaliarine sandauga* vadinama suma

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x} = \sum_{j=1}^{k} x_j y_j. \tag{6.1.7}$$

Skaliarinė sandauga tenkina tokias sąlygas. 1) $\boldsymbol{x}^T\boldsymbol{x} = 0$ tada ir tik tada, kai $\boldsymbol{x} = \mathbf{0}$.

2) Patenkintos sąlygos

$$c(\boldsymbol{x}^T\boldsymbol{y}) = (c\boldsymbol{x}^T)\boldsymbol{y}, \quad (\boldsymbol{x} + \boldsymbol{y})^T\boldsymbol{z} = \boldsymbol{x}^T\boldsymbol{z} + \boldsymbol{y}^T\boldsymbol{z}. \tag{6.1.8}$$

3) Tenkinama Koši ir Švarco nelygybė

$$(\boldsymbol{x}^T\boldsymbol{y})^2 \le (\boldsymbol{x}^T\boldsymbol{x})(\boldsymbol{y}^T\boldsymbol{y}). \tag{6.1.9}$$

Teigiama kvadratinė šaknis iš sandaugos $\boldsymbol{x}^T\boldsymbol{x}$ vadinama vektoriaus $\boldsymbol{x}$ *norma* arba *ilgiu*

$$||\boldsymbol{x}|| = +\sqrt{\boldsymbol{x}^T\boldsymbol{x}}. \tag{6.1.10}$$

Ji tenkina trikampio nelygybę

$$||\boldsymbol{x} - \boldsymbol{y}|| + ||\boldsymbol{y} - \boldsymbol{z}|| \ge ||\boldsymbol{x} - \boldsymbol{z}||. \tag{6.1.11}$$

Vektoriai $\boldsymbol{x}$ ir $\boldsymbol{y}$ yra *ortogonalūs*, jeigu skaliarinė sandauga

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x} = 0. \tag{6.1.12}$$

Kampas tarp dviejų nenulinių vektorių θ apibrėžiamas lygybe

$$\cos\theta = \frac{\boldsymbol{x}^T\boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}.$$

Nenuliniai poromis ortogonalūs vektoriai $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ yra tiesiškai nepriklausomi. Ortogonalūs vektoriai, generuojantys tiesinę erdvę $\mathcal{M}$, vadinami *ortogonalia* erdvės $\mathcal{M}$ *baze*, o jei jie turi vienetinius ilgius, – *ortonormuota baze*. Tiesinės erdvės ortonormuota bazė visada egzistuoja. Pavyzdžiui, vektoriai $\boldsymbol{e}_1 = (1,\ 0, ..., 0)^T$, $\boldsymbol{e}_2 = (0,\ 1, ..., 0)^T$, ..., $\boldsymbol{e}_k = (0,\ 0, ..., 1)^T$ sudaro erdvės $\mathbf{R}^k$ ortonormuotą bazę.

6.2. Matricos ir determinantai

Matrica $\boldsymbol{A}$ yra stačiakampė lentelė užpildyta realiais skaičiais. Kiekvieną matricos elementą a_{ij} numeruosime dviem indeksais: indeksas i nurodo eilutės, o indeksas j stulpelio, kurių sankirtoje yra elementas a_{ij}, numerius. Žymėsime $\boldsymbol{A} = [a_{ij}]_{m\times n}$, čia m yra eilučių skaičius, o n – stulpelių skaičius. Matricos $\boldsymbol{A}$ stulpeliai yra dimensijos m vektoriai, o eilutės – dimensijos n transponuoti vektoriai. *Matricų sudėtis.* Dviejų vienodos dimensijos $(m \times n)$ matricų $\boldsymbol{A} = [a_{ij}]_{m\times n}$ ir $\boldsymbol{B} = [b_{ij}]_{m\times n}$ suma yra matrica $\boldsymbol{A}+\boldsymbol{B} = [a_{ij}+b_{ij}]_{m\times n}$, kurios elementai gaunami sudedant atitinkamus matricų $\boldsymbol{A}$ ir $\boldsymbol{B}$ elementus. Šis veiksmas tenkina komutatyvumo ir distributyvumo savybes

$$\boldsymbol{A}+\boldsymbol{B} = \boldsymbol{B}+\boldsymbol{A}, \quad \boldsymbol{A}+(\boldsymbol{B}+\boldsymbol{C}) = (\boldsymbol{A}+\boldsymbol{B})+\boldsymbol{C}.$$

Daugyba iš skaliaro. Matricos $\boldsymbol{A} = [a_{ij}]_{m\times n}$ sandauga iš skaliaro $c \in \mathbf{R}$ suprantama kaip matrica, kurios kiekvienas elementas padaugintas iš skaliaro c:

$$c\boldsymbol{A} = [ca_{ij}]_{m\times n}.$$

Akivaizdu, kad tenkinamos sąlygos

$$(c+d)\boldsymbol{A} = c\boldsymbol{A}+d\boldsymbol{A}, \quad c(\boldsymbol{A}+\boldsymbol{B}) = c\boldsymbol{A}+c\boldsymbol{B}.$$

Matricų sandauga. Matricų $\boldsymbol{A} = [a_{ij}]_{m\times n}$ ir $\boldsymbol{B} = [b_{ij}]_{n\times r}$ sandauga $\boldsymbol{AB}$ apibrėžta tik tada, kai matricos $\boldsymbol{A}$ stulpelių skaičius sutampa su matricos $\boldsymbol{B}$ eilučių skaičiumi. Daugybos rezultatas yra matrica $\boldsymbol{AB} = [c_{ij}]_{m\times r}$, kurios elementas c_{ij} gaunamas imant matricos $\boldsymbol{A}$ i-osios eilutės ir matricos $\boldsymbol{B}$ j-ojo stulpelio skaliarinę sandaugą:

$$c_{ij} = \sum_{s=1}^{n} a_{is}b_{sj}, \quad i = 1,...,m, \quad j = 1,...,r. \tag{6.2.1}$$

Matricos $\boldsymbol{AB}$ eilučių skaičius sutampa su matricos $\boldsymbol{A}$ eilučių skaičiumi, o stulpelių skaičius lygus matricos $\boldsymbol{B}$ stulpelių skaičiui. Pavyzdžiui, sandauga $\boldsymbol{Ax}$, kai $\boldsymbol{A} = [a_{ij}]_{k\times n}$ yra matrica, o $\boldsymbol{x} = (x_1,...,x_n)^T$ – vektorius, yra k-matis vektorius. Kai $\boldsymbol{x} = (x_1,...,x_k)^T$ yra vektorius, tai $\boldsymbol{x}^T\boldsymbol{x}$ yra skaliaras (skaliarinė sandauga), o $\boldsymbol{C} = \boldsymbol{x}\boldsymbol{x}^T$ yra matrica $\boldsymbol{C} = [c_{ij}]_{k\times k}$, kai $c_{ij} = x_ix_j$.

Matricų daugyba netenkina komutatyvumo dėsnio, nes, pavyzdžiui, sandauga $\boldsymbol{AB}$ gali būti apibrėžta, o sandauga $\boldsymbol{BA}$ – neapibrėžta.

Asociatyvumo ir distributyvumo dėsniai yra tenkinami

$$\boldsymbol{ABC} = (\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}), \quad \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{BC},$$

$$(\boldsymbol{A}+\boldsymbol{B})(\boldsymbol{C}+\boldsymbol{D}) = \boldsymbol{A}(\boldsymbol{C}+\boldsymbol{D})+\boldsymbol{B}(\boldsymbol{C}+\boldsymbol{D}),$$

jeigu matricų dimensijos yra tokios, kad visi daugybos veiksmai yra apibrėžti.

Nulinė ir vienetinė matricos. Nulinė matrica $\mathbf{0}$ yra tokia, kurios visi elementai lygūs 0. Dimensijos $(m \times m)$ kvadratinė matrica vadinama *vienetine*, jeigu jos visi diagonaliniai elementai lygūs 1, o visi kiti elementai lygūs 0. Žymėsime $\boldsymbol{I} = \boldsymbol{I}_m$. Akivaizdu, kad

$$\boldsymbol{A}+\mathbf{0} = \boldsymbol{A}, \quad \boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A}. \tag{6.2.2}$$

Transponuota matrica. Matrica $\boldsymbol{A}^T$, kuri gaunama iš matricos $\boldsymbol{A}$ pakeitus jos eilutes stulpeliais ir, atvirkščiai, vadiname transponuota matrica, t.y.

$$\boldsymbol{A} = [a_{ij}]_{m\times n}, \quad \boldsymbol{A}^T = [a_{ji}]_{n\times m}.$$

Transponavimo operacija tenkina sąlygas

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T, \quad (\boldsymbol{ABC}) = \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T, ... \tag{6.2.3}$$

Matricos pėdsakas. Kvadratinės matricos $\boldsymbol{A} = [a_{ij}]_{m\times m}$ diagonalinių elementų suma vadinama matricos pėdsaku. Žymėsime

$$Tr(\boldsymbol{A}) = \sum_{j=1}^{m} a_{jj}.$$

Pėdsakas tenkina sąlygas

$$Tr(\boldsymbol{A}+\boldsymbol{B}) = Tr(\boldsymbol{A}) + Tr(\boldsymbol{B}), \quad Tr(\boldsymbol{AB}) = Tr(\boldsymbol{BA}). \tag{6.2.4}$$

Blokinės matricos. Kartais matricą $\boldsymbol{A}$ yra patogu suskaidyti į blokus

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{D} & \mathbf{E} \end{pmatrix}.$$

Transponavimo operaciją galima atlikti taip:

$$\boldsymbol{A}^T = \begin{pmatrix} \boldsymbol{B}^T & \boldsymbol{D}^T \\ \boldsymbol{C}^T & \mathbf{E}^T \end{pmatrix}.$$

Blokinių matricų daugyba atliekama pagal įprastines matricų daugybos taisykles

$$\begin{pmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{D} & \mathbf{E} \end{pmatrix} \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix} = \begin{pmatrix} \boldsymbol{BG}+\boldsymbol{CH} \\ \boldsymbol{DG}+\mathbf{E}\boldsymbol{H} \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{D} & \mathbf{E} \end{pmatrix} \begin{pmatrix} \boldsymbol{G} & \boldsymbol{J} \\ \boldsymbol{H} & \boldsymbol{L} \end{pmatrix} = \begin{pmatrix} \boldsymbol{BG}+\boldsymbol{CH} & \boldsymbol{BJ}+\boldsymbol{CL} \\ \boldsymbol{DG}+\mathbf{E}\boldsymbol{H} & \boldsymbol{DJ}+\mathbf{E}\boldsymbol{L} \end{pmatrix},$$

jeigu tik visi daugybos veiksmai yra apibrėžti. *Matricos rangas.* Matricą $\boldsymbol{A} = [a_{ij}]_{m\times n}$ sudaro n dimensijos m vektorių (matricos stulpeliai), arba m dimensijos n transponuotų vektorių (matricos eilutės). Tiesiškai nepriklausomų stulpelių (arba eilučių) skaičius vadinamas *matricos rangu.* Žymėsime $Rang(\boldsymbol{A})$. Matricos rangas turi tokias savybes

$$Rang(\boldsymbol{A}) = Rang(\boldsymbol{A}^T\boldsymbol{A}), \quad Rang(\boldsymbol{A}) \le \min(m,\ n) \tag{6.2.5}$$

$$Rang(\boldsymbol{AB}) \le \min(Rang(\boldsymbol{A}), Rang(\boldsymbol{B})), \quad Rang(\boldsymbol{A}+\boldsymbol{B}) \le Rang(\boldsymbol{A}) + Rang(\boldsymbol{B}).$$

Jeigu matrica $\boldsymbol{A} = [a_{ij}]_{n\times n}$ yra *idempotentinė*, t. y. $\boldsymbol{AA} = \boldsymbol{A}$, tai

$$Rang(\boldsymbol{A}) + Rang(\boldsymbol{I}-\boldsymbol{A}) = n. \tag{6.2.6}$$

Atvirkštinė matrica. Kvadratinė matrica $\boldsymbol{A} = [a_{ij}]_{n\times n}$ vadinama *neišsigimusia*, jeigu jos rangas lygus n. Tokiu atveju egzistuoja vienintelė matrica $\boldsymbol{A}^{-1} = [a^{ij}]_{n\times n}$, vadinama *atvirkštine* matricai $\boldsymbol{A}$, kad patenkintos sąlygos

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}. \tag{6.2.7}$$

Jeigu $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ vienodos dimensijos neišsigimusios matricos, tai

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}, \quad (\boldsymbol{ABC})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}, \quad (\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T. \tag{6.2.8}$$

Kvadratinė matrica $\boldsymbol{A}$ vadinama *ortogonalia*, jeigu $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$. Tokiu atveju atvirkštinė matrica sutampa su transponuota $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$ ir

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}. \tag{6.2.9}$$

Determinantas. Kvadratinės matricos $\boldsymbol{A} = [a_{ij}]_{m\times m}$ *determinantas* yra jos elementų a_{ij} skaliarinė funkcija:

$$|\boldsymbol{A}| = \sum \pm(a_{1i_1}a_{2i_2}...a_{mi_m});$$

sumavimas atliekamas pagal visus skirtingus skaičių $(1, 2, ..., m)$ kėlinius $(i_1, i_2, ..., i_m)$; sandauga imama su teigiamu ženklu, jei kėlinys gaunamas perstatant aibės $(1, 2, ..., m)$ elementus lyginį skaičių kartų ir su neigiamu ženklu, jei – nelyginį skaičių kartų.

Pažymėkime $\boldsymbol{A}_{ij}$ elemento a_{ij} algebrinį papildinį. Jis lygus sandaugai $(-1)^{i+j}$ iš determinanto matricos, gaunamos išbraukus i-ąją eilutę ir j-ąjį stulpelį. Tada

$$|\boldsymbol{A}| = \sum_{i=1}^{m} a_{ri}\boldsymbol{A}_{ri}, \quad \text{kiekvienam} \quad r = 1, ..., m; \tag{6.2.10}$$

$$|\boldsymbol{A}| = \sum_{r=1}^{m} a_{ri}\boldsymbol{A}_{ri}, \quad \text{kiekvienam} \quad i = 1, ..., m. \tag{6.2.11}$$

Pateikiame dar keletą savybių.

1) $|\boldsymbol{A}| = 0$ tada ir tik tada, kai $Rang(\boldsymbol{A}) \neq m$.
2) Jeigu $|\boldsymbol{A}| \neq 0$, tai $\boldsymbol{A}^{-1} = (\boldsymbol{A}_{ij}/|\boldsymbol{A}|)^T$.
3) Jei $\boldsymbol{A}$ diagonalinė (nenuliniai elementai gali būti tik ant pagrindinės diagonalės) ar trikampė (matricos elementai į kairę arba į dešinę nuo pagrindinės diagonalės lygūs nuliui) matrica, tai $|\boldsymbol{A}|$ lygus diagonalinių elementų sandaugai.
4) Jei $\boldsymbol{A}$ ir $\boldsymbol{B}$ vienodos dimensijos kvadratinės matricos, tai

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|. \tag{6.2.12}$$

5) Jeigu blokinėje matricoje

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{D} & \mathbf{E} \end{pmatrix},$$

determinantas $|\boldsymbol{B}| \neq 0$, tai

$$|\boldsymbol{A}| = |\boldsymbol{B}||\mathbf{E} - \boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C}|. \tag{6.2.13}$$

Tiesinės lygčių sistemos. Tiesinę m lygčių sistemą kintamųjų $x_1, ..., x_m$ atžvilgiu matricine forma galima užrašyti taip:

$$\boldsymbol{Ax} = \boldsymbol{b}, \tag{6.2.14}$$

čia $\boldsymbol{A} = [a_{ij}]_{m\times m}$ žinomų koeficientų matrica, $\boldsymbol{b}^T = (b_1, ..., b_m)$ laisvasis narys, o $\boldsymbol{x}^T = (x_1, ..., x_m)$ nežinomas vektorius. Arba ekvivalenčia forma

$$x_1\boldsymbol{a}_1 + ... + x_m\boldsymbol{a}_m = \boldsymbol{b}, \tag{6.2.15}$$

čia $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ yra matricos $\boldsymbol{A}$ stulpeliai. 1) Homogeninė lygčių sistema ($\boldsymbol{b} = \mathbf{0}$) turi nenulinį sprendinį tada ir tik tada, kai vektoriai $\boldsymbol{a}_1, ...\boldsymbol{a}_m$ yra tiesiškai priklausomi.
2) Nehomogeninė lygčių sistema ($\boldsymbol{b} \neq \mathbf{0}$) turi sprendinį tada ir tik tada, kai vektorius $\boldsymbol{b}$ gali būti išreikštas tiesiniu vektorių $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ dariniu. Bendras šios sistemos sprendinys yra kurio nors jos sprendinio ir bendrojo homogeninės sistemos sprendinio suma.
3) Nehomogeninė lygčių sistema turi vienintelį sprendinį, kai vektoriai $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ yra tiesiškai nepriklausomi, t.y. kai $Rang(\boldsymbol{A}) = m$. Sprendinys turi tokį pavidalą

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}. \tag{6.2.16}$$

Šiuo atveju homogeninė lygčių sistema turi tik trivialų sprendinį $\boldsymbol{x} = \mathbf{0}$.

7 skyrius

2 priedas. Atsitiktiniai vektoriai

7.1. Atsitiktinio vektoriaus skirstinys

Tarkime turime tikimybinę erdvę $(\Omega, \mathcal{F}, P)$.
2P.1 apibrėžimas. Realią vinareikšmę k-matę funkciją $\mathbf{X} = \mathbf{X}(\omega) = (X_1(\omega), ..., X_k(\omega))^T$, apibrėžtą aibėje Ω ir tokią, kad su visais $\boldsymbol{x} = (x_1, ..., x_k)^T \in \mathbf{R}^k$

$$\{\omega : X_1(\omega) \le x_1, ..., X_k(\omega) \le x_k\} \in \mathcal{F}$$

vadiname k-mačiu *atsitiktiniu vektoriumi* (a.v.). 1) Bet koks a.v. vienareikšmiškai nusakomas jo pasiskirstymo funkcijos

$$F(\boldsymbol{x}) = F(x_1, ..., x_k) = \mathbf{P}\{X_1 \le x_1, ..., X_k \le x_k\}, \tag{7.1.1}$$

apibrėžtos su visais $\boldsymbol{x} = (x_1, ..., x_k)^T \in \mathbf{R}^k$. 2) Jeigu a.v. $\mathbf{X}$ igyjamų reikšmių aibė yra baigtinė arba skaiti, tai tokio a.v. skirstinys vadinamas *diskrečiuoju*. Jo skirstinys visiškai nusakomas išvardijant galimas reikšmes $\boldsymbol{x}_1, \boldsymbol{x}_2, ... \in \mathbf{R}^k$ ir jų įgijimo tikimybes

$$p_i = \{\mathbf{X} = \boldsymbol{x}_i\}, \quad i = 1, 2, ...; \quad \sum_i p_i = 1. \tag{7.1.2}$$

3) Absoliučiai tolydaus a.v. $\mathbf{X}$ skirstinys visiškai nusakomas jo *tankio funkcija*

$$f(\boldsymbol{x}) = f(x_1, ..., x_k) = \frac{\partial^k}{\partial x_1 ... \partial x_k} F(x_1, ..., x_k). \tag{7.1.3}$$

4) Bet kokio a.v. $\mathbf{X}$ skirstinį vienareikšmiškai nusako jo *charakteristinė funkcija*

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(t_1, ..., t_k) = \mathbf{E}e^{i\boldsymbol{t}^T\boldsymbol{X}} = \mathbf{E}e^{i(t_1X_1 + ... + t_kX_k)}, \tag{7.1.4}$$

nusakyta su visais $\boldsymbol{t} = (t_1, ..., t_k)^T \in \mathbf{R}^k$. Pateiksime keletą charakteristinės funkcijos savybių. a) A.v. $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\mathbf{X}$ ($\boldsymbol{a}$ fiksuotas dimensijos k vektorius, $\boldsymbol{B}$ fiksuota dimensijos k kvadratinė matrica) charakteristinė funkcija

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{a}}\varphi_{\boldsymbol{X}}(\boldsymbol{t}^T\boldsymbol{B}_1, ..., \boldsymbol{t}^T\boldsymbol{B}_k), \tag{7.1.5}$$

čia $\boldsymbol{B}_1, ..., \boldsymbol{B}_k$ yra matricos $\boldsymbol{B}$ stulpeliai.

b) Nepriklausomų a.v. $\mathbf{X}_1, ..., \mathbf{X}_n$ sumos $\boldsymbol{S}_n = \mathbf{X}_1 + ... + \mathbf{X}_n$ charakteristinė funkcija lygi dėmenų charakteristinių funkcijų sandaugai

$$\varphi_{\boldsymbol{S}_n}(\boldsymbol{t}) = \prod_{j=1}^{n} \varphi_{\boldsymbol{X}_j}(\boldsymbol{t}). \tag{7.1.6}$$

c) Remiantis charakteristinės funkcijos apibrėžimu įrodoma Kramero ir Voldo teorema. A.v. $\mathbf{X}$ tikimybinis skirstinys nusakytas tada ir tik tada, kai nusakyti skirstiniai vienmačių a.d.

$$Y_{\boldsymbol{L}} = \boldsymbol{L}^T\mathbf{X} = L_1X_1 + ... + L_kX_k$$

su visais vektoriais $\boldsymbol{L} = (L_1, ..., L_k)^T \in \mathbf{R}^k$.

7.2. Marginalieji ir sąlyginiai skirstiniai

Nagrinėsime $(k+s)$-matį absoliučiai tolydųjį a.v. $\mathbf{X} = (\boldsymbol{Y}^T, \boldsymbol{Z}^T)^T = (Y_1, ..., Y_k, Z_1, ..., Z_s)^T$. Vektorių $\mathbf{X}, \boldsymbol{Y}$ ir $\boldsymbol{Z}$ pasiskirstymo funkcijas žymėsime $F(\boldsymbol{y}, \boldsymbol{z}) = F(y_1, ..., y_k, z_1, ..., z_s)$, $G(\boldsymbol{y}) = G(y_1, ..., y_k)$ ir $H(\boldsymbol{z}) = H(z_1, ..., z_s)$, o tankių funkcijas $f(\boldsymbol{y}, \boldsymbol{z}), g(\boldsymbol{y})$ ir $h(\boldsymbol{z})$.

Marginalieji skirstiniai. Vektoriaus, sudaryto iš bet kurių pradinio vektoriaus koordinačių, skirstinys vadinamas *marginaliuoju* skirstiniu pradinio vektoriaus atžvilgiu. Pavyzdžiui, a.v. $\boldsymbol{Y}$ skirstinys vadinamas marginaliuoju jungtinio vektoriaus $\mathbf{X} = (\boldsymbol{Y}^T, \boldsymbol{Z}^T)^T$ atžvilgiu.

1) Vektoriaus $\boldsymbol{Y}$ pasiskirstymo funkcija gaunama iš a.v. $\mathbf{X}$ pasiskirstymo funkcijos įrašant $+\infty$ vietoje argumentų, atitinkančių likusias pradinio vektoriaus koordinates

$$G(\boldsymbol{y}) = G(y_1, ..., y_k) = F(y_1, ..., y_k, +\infty, ..., +\infty). \tag{7.2.1}$$

2) Marginaliojo skirstinio tankis gaunamas integruojant:

$$g(\boldsymbol{y}) = \int_{\mathbf{R}^s} f(\boldsymbol{y}, \boldsymbol{z})d\boldsymbol{z}. \tag{7.2.2}$$

3) A.v. $\boldsymbol{Y}$ charakteristinė funkcija

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = \varphi_{\boldsymbol{Y}}(t_1, ..., t_k) = \mathbf{E}e^{i\boldsymbol{t}^T\boldsymbol{Y}} = \mathbf{E}e^{i(t_1Y_1 + ... + t_kY_k)}$$

gaunama iš jungtinio vektoriaus $\mathbf{X} = (\boldsymbol{Y}^T, \boldsymbol{Z}^T)^T$ charakteristinės funkcijos

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}, \boldsymbol{\theta}) = \varphi_{\boldsymbol{X}}(t_1, ..., t_k, \theta_1, ..., \theta_s) = \mathbf{E}e^{i(\boldsymbol{t}^T\boldsymbol{Y} + \boldsymbol{\theta}^T\boldsymbol{Z})}$$

įrašant vietoje $\boldsymbol{\theta}$ nulinį vektorių $\mathbf{0}$:

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(\boldsymbol{t}, \mathbf{0}) = \varphi_{\boldsymbol{X}}(t_1, ...t_k, 0, ..., 0). \tag{7.2.3}$$

Sąlyginiai skirstiniai. Tarkime,kad a.v. $\boldsymbol{Z} = \boldsymbol{z}$ yra fiksuotas. Jeigu $h(\boldsymbol{z}) \neq 0$, tai a.v. $\boldsymbol{Y}$ sąlyginio skirstinio tankis

$$g(\boldsymbol{y}|\boldsymbol{z}) = \frac{f(\boldsymbol{y}, \boldsymbol{z})}{h(\boldsymbol{z})}. \tag{7.2.4}$$

Turėdami a.v. $\boldsymbol{Y}$ sąlyginį tankį, kai $\boldsymbol{Z} = \boldsymbol{z}$ fiksuotas, ir besąlyginį a.v. $\boldsymbol{Z}$ tankį, galima atstatyti a.v. $\boldsymbol{Y}$ besąlyginį tankį:

$$g(\boldsymbol{y}) = \int_{\mathbf{R}^s} g(\boldsymbol{y}|\boldsymbol{z})h(\boldsymbol{z})d\boldsymbol{z}. \tag{7.2.5}$$

Ši formulė apibendrina pilnosios tikimybės formulę. Analogiškai apibendrinamos ir Bejeso formulės

$$h(\boldsymbol{z}|\boldsymbol{y}) = \frac{g(\boldsymbol{y}|\boldsymbol{z})h(\boldsymbol{z})}{\int_{\mathbf{R}^s} g(\boldsymbol{y}|\boldsymbol{z})h(\boldsymbol{z})d\boldsymbol{z}}. \tag{7.2.6}$$

Nepriklausomi vektoriai. Jeigu kiekvieno iš dviejų vektorių sąlyginiai skirstiniai esant fiksuotoms kito vektoriaus reikšmėms, nepriklauso nuo fiksuotųjų reikšmių ir sutampa su besąlyginiais skirstiniais, tai tokie vektoriai vadinami nepriklausomais.

Suformuluosime keletą nepriklausomumo kriterijų. Atsitiktiniai vektoriai $\boldsymbol{Y}$ ir $\boldsymbol{Z}$ nepriklausomi tada ir tik tada, kai

1) pasiskirstymo funkcija $F(\boldsymbol{y}, \boldsymbol{z})$ yra lygi marginaliųjų pasiskirstymo funkcijų sandaugai

$$F(\boldsymbol{y}, \boldsymbol{z}) = G(\boldsymbol{y})H(\boldsymbol{z}), \quad \boldsymbol{y} \in \mathbf{R}^k, \quad \boldsymbol{z} \in \mathbf{R}^s; \tag{7.2.7}$$

2) tankio funkcija $f(\boldsymbol{y},\boldsymbol{z})$ yra lygi marginaliųjų tankio funkcijų sandaugai

$$f(\boldsymbol{y},\boldsymbol{z}) = g(\boldsymbol{y})h(\boldsymbol{z}), \quad \boldsymbol{y}\in\mathbf{R}^k, \quad \boldsymbol{z}\in\mathbf{R}^s; \tag{7.2.8}$$

3) charakteristinė funkcija $\varphi_{\boldsymbol{X}}(\boldsymbol{t},\boldsymbol{\theta})$ yra lygi marginaliųjų charakteristinių funkcijų sandaugai

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t},\boldsymbol{\theta}) = \varphi_{\boldsymbol{Y}}(\boldsymbol{t})\varphi_{\boldsymbol{Z}}(\boldsymbol{\theta}), \quad \boldsymbol{t}\in\mathbf{R}^k, \quad \boldsymbol{\theta}\in\mathbf{R}^s. \tag{7.2.9}$$

Analogiškai formuluojami ir didesnio skaičiaus atsitiktinių vektorių nepriklausomumo kriterijai.

7.3. Atsitiktinio vektoriaus momentai

.

1) Atsitiktinio vektoriaus $\mathbf{X} = (X_1,...,X_k)^T$ *vidurkis* yra vektorius, sudarytas iš jo koordinačių vidurkių

$$\mathbf{E}(\mathbf{X}) = (\mathbf{E}X_1,...,\mathbf{E}X_k)^T = (\mu_1,...,\mu_k)^T = \boldsymbol{\mu}. \tag{7.3.1}$$

2) Atsitiktinio vektoriaus $\boldsymbol{Y}$ sąlyginis vidurkis, kai kitas vektorius $\boldsymbol{Z}=\boldsymbol{z}$ yra fiksuotas, t.y. vidurkis apibrėžtas remiantis sąlyginiu tankiu (7.2.4)

$$\boldsymbol{\mu}(\boldsymbol{z}) = \mathbf{E}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}) = \int_{\mathbf{R}^k} \boldsymbol{y}g(\boldsymbol{y}|\boldsymbol{z})d\boldsymbol{y},$$

vadinamas a.v. $\boldsymbol{Y}$ *regresija* a.v. $\boldsymbol{Z}$ atžvilgiu. 3) Atsitiktinio vektoriaus $\mathbf{X} = (X_1,...,X_k)^T$ antrųjų mišriųjų centrinių momentų matrica vadinama *kovariacijų matrica*

$$\boldsymbol{V}(\mathbf{X}) = \boldsymbol{\Sigma} = [\sigma_{ij}]_{k\times k}, \tag{7.3.2}$$

čia

$$\sigma_{ij} = \mathbf{Cov}\,(X_i,X_j) = \mathbf{E}[(X_i-\mathbf{E}X_i)(X_j-\mathbf{E}X_j)] = \mathbf{E}(X_iX_j) - \mathbf{E}X_i\mathbf{E}X_j,$$
$$\sigma_{ii} = \mathbf{Cov}\,(X_i,X_i) = \boldsymbol{V}X_i, \quad i,j=1,...,k,$$

arba, trumpiau,

$$\boldsymbol{V}(\mathbf{X}) = \mathbf{E}[(\mathbf{X}-\mathbf{E}(\mathbf{X}))(\mathbf{X}-\mathbf{E}(\mathbf{X}))^T] = \mathbf{E}(\mathbf{X}\mathbf{X}^T) - \mathbf{E}(\mathbf{X})(\mathbf{E}(\mathbf{X}))^T.$$

Naudojant kovariacijas sudaroma *koreliacijos koeficientų* matrica

$$\boldsymbol{R} = [\rho_{ij}]_{k\times k}, \quad \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \quad i,j=1,...,k. \tag{7.3.3}$$

3) Tarkime $\boldsymbol{A} = [a_{ij}]_{m\times k}$ yra matrica su pastoviais koeficientais, o $\mathbf{X} = (X_1,...,X_k)^T$ k-matis vektorius. Tada $\boldsymbol{Y} = \boldsymbol{A}\mathbf{X}$ yra m-matis atsitiktinis vektorius. Vektoriaus $\boldsymbol{Y}$ pirmieji momentai

$$\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\mu}, \quad \boldsymbol{V}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T = [\gamma_{ij}]_{m\times m}. \tag{7.3.4}$$

Jeigu $\boldsymbol{B} = [b_{ij}]_{n\times k}$ yra kita matrica su pastoviais koeficientais, tai

$$\mathbf{Cov}\,(\boldsymbol{A}\mathbf{X},\boldsymbol{B}\mathbf{X}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = [\delta_{ij}]_{m\times n}.$$

Kvadratinės formos

$$Q = \mathbf{X}^T\boldsymbol{A}\mathbf{X}$$

vidurkis

$$\mathbf{E}Q = Tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}. \tag{7.3.5}$$

4) Tarkime $\mathbf{X}_1,...,\mathbf{X}_n$ yra vienodai pasiskirstę nepriklausomi atsitiktiniai vektoriai, kurių $\mathbf{E}(\mathbf{X}_j) = \boldsymbol{\mu}$ ir $\boldsymbol{V}(\mathbf{X}_j) = \boldsymbol{\Sigma}$. Tada centruotos ir normuotos sumos

$$\bar{\boldsymbol{S}}_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})$$

pirmieji momentai

$$\mathbf{E}(\bar{\boldsymbol{S}}_n) = \mathbf{0}, \quad \boldsymbol{V}(\bar{\boldsymbol{S}}_n) = \boldsymbol{\Sigma}. \tag{7.3.6}$$

Remiantis charakteristinių funkcijų savybe (9.1.6), Kramero ir Voldo teorema ir vienmate centrine ribine teorema (CRT) gauname, pavyzdžiui, tokį paprasčiausią daugiamatės CRT variantą. Jeigu $|\boldsymbol{\Sigma}| > 0$ ir $n \to \infty$, tai

$$\bar{\boldsymbol{S}}_n \xrightarrow{d} \boldsymbol{Y} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma}), \quad \bar{\boldsymbol{S}}_n \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{S}}_n \xrightarrow{d} \chi^2_k. \tag{7.3.7}$$

Jeigu $\boldsymbol{\Sigma}$ yra rango $r \leq k$ idempotentinė matrica, tai

$$\bar{\boldsymbol{S}}_n \boldsymbol{\Sigma}^{-} \bar{\boldsymbol{S}}_n \xrightarrow{d} \chi^2_r. \tag{7.3.8}$$

7.4. Daugiamatis normalusis skirstinys

Atsitiktinio normaliojo vektoriaus $\mathbf{X} = (X_1, ..., X_n)^T$ skirstinys visiškai nusakomas jo vidurkių vektoriumi $\mathbf{E}(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ ir kovariacijų matrica $\boldsymbol{V}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = [\sigma_{ij}]_{n\times n}$. Sutrumpintai žymima $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Charakteristinė funkcija

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(t_1, ..., t_n) = e^{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}}. \tag{7.4.1}$$

Remiantis charakteristinės funkcijos išraiška (7.4.1) ir savybėmis (7.1.5), (7.1.6), (7.2.3), (7.2.4) gaunamos tokios išvados:

1) Jeigu $\boldsymbol{a}$ fiksuotas k-matis vektorius, o $\boldsymbol{B} = [b_{ij}]_{k\times n}$ fiksuota matrica, tai

$$\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\mathbf{X} \sim N_k(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T).$$

2) Tegu $\mathbf{X}_1, ..., \mathbf{X}_k$ yra nepriklausomi a.v. ir $\mathbf{X}_i \sim N_n(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, o $c_1, ..., c_k$ – konstantos. Tada

$$\boldsymbol{Y} = c_1\mathbf{X}_1 + ... + c_k\mathbf{X}_k \sim N_n\left(\sum_{j=1}^{k} c_j\boldsymbol{\mu}_j, \sum_{j=1}^{k} c_j^2\boldsymbol{\Sigma}_j\right). \tag{7.4.2}$$

3) Vektoriaus $\mathbf{X}'$, sudaryto iš k skirtingų vektoriaus $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ koordinačių, skirstinys yra k-matis normalusis.

4) Vektoriai $\mathbf{X}^{(1)}$ ir $\mathbf{X}^{(2)}$, sudaryti iš skirtingų a.v. $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ koordinačių, yra nepriklausomi tada ir tik tada, kai $\mathbf{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \mathbf{0}$. Vektoriaus $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ visos koordinatės nepriklausomos tarpusavyje tada ir tik tada, kai matrica $\boldsymbol{\Sigma}$ yra diagonalioji. Tarkime, kad kovariacinė matrica $\boldsymbol{\Sigma}$ neišsigimusi, t.y. $|\boldsymbol{\Sigma}| \neq 0$. Tada egzistuoja a.v. $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tankio funkcija

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi|\boldsymbol{\Sigma}|)^{-\frac{n}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}. \tag{7.4.3}$$

Kvadratinė forma

$$Q = (\mathbf{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(n) \tag{7.4.4}$$

turi χ^2 skirstinį su n laisvės laipsnių. Jeigu vietoje vidurkio $\boldsymbol{\mu}$ įrašysime kitą vektorių $\boldsymbol{\nu}$, tai

$$Q = (\mathbf{X} - \boldsymbol{\nu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\nu}) \sim \chi^2(n; \delta) \tag{7.4.5}$$

turi necentrinį χ^2 skirstinį su n laisvės laipsnių ir necentriškumo parametru

$$\delta = (\boldsymbol{\mu} - \boldsymbol{\nu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}). \tag{7.4.6}$$

Tegu $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, o $\mathbf{X}^{(1)}$ ir $\mathbf{X}^{(2)}$ yra k-matis ir $(n-k)$-matis vektoriai, sudaryti iš skirtingų vektoriaus $\mathbf{X}$ koordinačių. Pažymėkime $\boldsymbol{\mu}^{(1)} = \mathbf{E}(\mathbf{X}^{(1)})$, $\boldsymbol{\mu}^{(2)} = \mathbf{E}(\mathbf{X}^{(2)})$, $\boldsymbol{\Sigma}_{11} = \boldsymbol{V}(\mathbf{X}^{(1)})$, $\boldsymbol{\Sigma}_{22} = \boldsymbol{V}(\mathbf{X}^{(2)})$, $\boldsymbol{\Sigma}_{12} = \mathbf{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \boldsymbol{\Sigma}_{21}^T$. Tada sąlyginis $\mathbf{X}^{(2)}$ skirstinys, kai $\mathbf{X}^{(1)}$ fiksuotas yra $(n-k)$-matis normalusis su parametrais

$$\begin{aligned}\mathbf{E}(\mathbf{X}^{(2)}|\mathbf{X}^{(1)} = \boldsymbol{x}^{(1)}) &= \boldsymbol{\mu}^{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}^{(1)} - \boldsymbol{\mu}^{(1)}),\\ \boldsymbol{V}(\mathbf{X}^{(2)}|\mathbf{X}^{(1)} = \boldsymbol{x}^{(1)}) &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}.\end{aligned} \tag{7.4.7}$$

Taigi bet kurios normaliojo vektoriaus koordinatės regresija kitų koordinačių atžvilgiu yra tiesinė funkcija.

Literatūra

1. Afifi A. A., Azen S. P. Statistical Analysis A Computer Oriented Approach. Vertimas į rusų kalbą.– Maskva: „Mir“, 1982.

2. Chatterjee M. H., Hadi A. S. Regression Analysis by Example (fourth edition). Wiley, 2006.

3. Cochran W. G., Cox G. Experimental Designs (second edition). New York: John Wiley, 1957.

4. Čekanavičius V., Murauskas G. Statistika ir jos taikymai. Vilnius: TEV, I dalis – 2000; II dalis – 2002; III dalis – 2009.

5. Čekanavičius V., Murauskas G. Taikomoji regresinė analizė socialiniuose tyrimuose. Vilnius: VU leidykla, 2014 (http://www.statistika.mif.vu.lt/wp - content/uploads/2014/04/regresine - analize.pdf).

6. Dobson A. J. An Introduction to Generalized Linear Models. Chapman and Hall/CRC, 2002.

7. Hiks Ch. R. Fundamental Consepts in the Design of Experiments. Vertimas į rusų kalbą. – Maskva: „Mir“, 1967.

8. Hosmer J. D., Lemeshov S. Applied Logistic Regression (second edition). Wiley, 2000.

9. Kleinbaum D. G., Klein M. Logistic Regression A Self – Learning Text. Springer – Verlag, 2002.

10. Kutner M. H., Nachtsheim Ch. J., Neter J., Li W. Applied Linear Statistical Models (fifth edition). New York: Mc Grow – Hill, 2005.

11. Levulienė R. Statistikos taikymai naudojant SAS. Vilnius: VU leidykla, 2009.

12. Rao C. R. Linear Statistical Inference and its Applications. 2nd edn. Wiley, 2002 (yra vertimas į rusų kalbą).

13. SAS. Help and Documentation. Kompaktinė plokštelė (platinama su SAS sistema).

14. Scheffe H. The Analysis of Variance. Wiley, 1999 (yra vertimas į rusų kalbą).

15. Skorniakov V. Apibendrinti tiesiniai modeliai. 2015 (http://www.mif.vu.lt/ visk/).

16. Weisberg S. Applied Linear Regression. Wiley, 2005.

Dalykinė rodyklė

analizė
  dispersinė
    dvifaktorė, 52
  dispersinė, 11
    atsitiktinių blokų, 98
    atsitiktinių faktorių, 75
    atsitiktinio faktoriaus, 69
    blokuotųjų duomenų, 102
    daugiafaktorė, 85
    dvifaktorė, 41, 48
    grupuotųjų planų, 93
    hipotezių tikrinimas, 89, 97, 98, 104, 108, 114
    kvadratų sumos, 87
    lotyniškųjų kvadratų, 111
    mišriojo modelio, 79
    MK įvertiniai, 32, 43, 53
    nepilnų blokų, 105
    nepilni planai, 93
    parametrų įvertiniai, 72, 78, 85, 89, 95, 107, 113
  kontrastų
    S metodas, 48, 51, 63, 67, 109, 115
    T metodas, 48, 51
  kovariacinė, 164
    bendras atvejis, 169
    hipotezių tikrinimas, 168, 171
    parametrų įvertiniai, 166, 170
  regresinė, 11
    parametrų įvertiniai, 135, 148

bazė
  ortogonali, 227
  ortonormuota, 227
  vektorinės erdvės, 227

daugyba
  matricų, 228
determinantas, 229
dispersija
  liekamoji, 130

eksperimentas
  faktorinis
    koduotas planas, 182
    rototabilus, 184
    faktorinis 2^2, 181
    faktorinis 2^3, 182
    faktorinis 2^m, 179, 183
eksperimento planavimas, 90
erdvė
  tiesinė vektorinė, 227

formulė
  Bejeso
    apibendrintoji, 232
  pilnosios tikimybės
    apibendrintoji, 232
funkcija
  charakteristinė
    atsitiktinio vektoriaus, 231
  jungties, 196
    kanoninė, 196
  tankio
    daugiamačio normaliojo vektoriaus, 234

hipotezė
  dėl regresijos parametro, 137
  faktorių įtakos nebuvimo, 56
  modelio adityvumo, 49
  regresijos tiesiškumo, 139
  sąveikos nebuvimo, 56, 57
  sudėtinė, 18
  vidurkių lygybės, 33, 44

imties plotis
  stjudentizuotas, 39
interpretacija
  regresijos parametrų, 146
intervalas
  pasikliovimo
    dispersijos, 22
    parametrų funkcijos, 22
    regresijos parametrų, 137, 149
  prognozės, 137, 150
įvertinys
  mažiausiųjų kvadratų, 12
    tiesinės parametrų funkcijos, 15

klasifikacija
  hierarchinė, 92

kryžminė, 92
kodavimas
kovariančių
nominalių, 146
nominalios kovariantės, 206
koeficientas
dalinės koreliacijos
empirinis, 154
determinacijos, 152, 199, 212, 214
koreliacijos
dalinis, 134
dauginis, 131
dauginis empirinis, 152
kontrastas, 36
normuotas, 38
repliką generuojantis, 185
kovariantė
nominali, 146
kriterijai
dėl regresijos parametrų, 151
kriterijus
dėl daugiamačio parametro, 24
dėl dispersijos reikšmės, 22
dėl parametrų funkcijos, 22
faktorių įtakos nebuvimo, 63, 67, 84
faktoriaus įtakos nebuvimo, 71, 77, 100, 101
regresijos tiesiškumo, 139, 155
sąveikos nebuvimo, 60, 77, 84
Tjukio
modelio adityvumo, 49
vidurkių lygybės, 34, 45

lentelė
dispersinės analizės, 36, 46, 76, 83, 89, 96, 101, 103
atsitiktinis faktorius, 71
kovariacinės analizės, 167
liekana
standartizuota, 140

matrica, 228
atvirkštinė, 229
blokinė, 229
eksperimento plano, 10
idempotentinė, 229
koreliacinė, 233
kovariacinė, 233
tiesinės formos, 233
neišsigimusi, 229
nulinė, 228
ortogonalioji, 229
transponuota, 228
vienetinė, 228
metodas
evoliucinio planavimo, 180
modelis
adityvus, 48, 63
dispersinės analizės
adityvus, 43, 61
atsitiktiniai faktoriai, 73
atsitiktinis faktorius, 69
blokuotieji duomenys, 98
daugiafaktoris, 86
dvifaktorės, 41
hierarchinis, 94, 97
lotyniškųjų kvadratų, 111
mišrusis, 79
nepilnų blokų, 106
nesubalansuotas, 52
vienfaktorės, 31
kovariacinės analizės, 164
kelių kintamųjų, 165
vieno kintamojo, 165
neadityvus, 64
prisotintas, 196
regresijos
logistinės, 205
vieno kintamojo, 134
regresinės analizės
kelių kintamųjų, 145
tiesinis, 10
apibendrintas, 196
Gauso ir Markovo, 10
momentai
atsitiktinio vektoriaus, 233

nelygybė
Koši ir Švarco, 227
nepriklausomumas
atsitiktinių vektorių, 232
normaliųjų vektorių, 234

pėdsakas
matricos, 229
paklaida
prognozės, 129, 148
vidutinė kvadratinė, 129
planas
eksperimentų
blokuotųjų duomenų, 91
randomizuotas, 91
subalansuotas, 41
prognozė, 129
optimalioji, 129
tiesinė, 130
optimalioji, 130

rangas
matricos, 229
regresija, 129, 233
binominė, 204
gama, 201
logistinė, 204

hipotezių tikrinimas, 214
parametrų įvertiniai, 208
parametrų interpretavimas, 205
parametrų kodavimas, 206
pasikliovimo intervalai, 217
neigiamoji binominė, 202
netiesinė, 143
pažingsninė, 156
puasoninė, 200
replika
faktorinio eksperimento, 185
rinkinys
pasikliovimo intervalų, 25

S metodas
vidurkių palyginimo, 37
sąryšis
repliką generuojantis, 185
sąveika
faktorių, 43
sandauga
skaliarinė, 227
santykis
koreliacinis, 130
dalinis, 133
tikėtinumų, 213
savybės
įvertinių
mažiausiųjų kvadratų, 13, 16
sistema
lygčių
mažiausiųjų kvadratų, 12
svorių
dažnuminė, 56
vienodų, 56
tiesinių lygčių, 230
skirstiniai
įvertinių
mažiausiųjų kvadratų, 17
skirstinys
atsitiktinio vektoriaus, 231
diskretusis, 231
tolydusis, 231
eksponentinio tipo, 196
marginalusis, 232
normalusis
daugiamatis, 234
sąlyginis, 232
normaliojo vektoriaus, 234
skirtumai
likutiniai, 140
sprendinys
tiesinių lygčių sistemos, 230
sritis
pasikliovimo
daugiamačio parametro, 23
sudėtis
matricų, 228
vektorių, 226
suma
kvadratų
liekamoji, 13
šansų santykis, 206
šansas, 205, 206

T metodas
vidurkių palyginimo, 39
teorema
Šefės, 56
centrinė ribinė
daugiamatė, 234
Kramero ir Voldo, 232
Rao, 170
tikimybė
klasifikavimo tikslumo, 218
aposteriorinė, 218

uždavinys prognozavimo, 128

vektoriai
ortogonalūs, 227
tiesiškai nepriklausomi, 227
tiesiškai priklausomi, 227
vektorius, 226
atsitiktinis, 231
vidurkių
tiesinės formos, 233
vidurkis
vektoriaus kvadratinės formos, 233

Vilijandas Bagdonavičius, Julius Jonas Kruopis
Matematinė statistika: vadovėlis
Antra dalis. Tiesiniai modeliai. – Vilnius: Vilniaus universitetas, 2015. – 237 p.
ISBN 978–609–459–516–5

Antroje vadovėlio dalyje nagrinėjami tiesiniai modeliai, kuriuose imties elementų vidurkiai išreiškiami tiesinėmis nežinomų parametrų funkcijomis. Parametrai apibūdina imties skirstinio priklausomybę nuo tam tikrų kovariančių. Tiesiniai modeliai – viena iš svarbiausių matematinės statistikos statistikos dalių. Jie labai plačiai taikomi įvairiose mokslo ir praktikos srityse kylantiems uždaviniams spręsti. Vadovėlyje gana detaliai nagrinėjami tiesinių modelių – dispersinės analizės, regresinės analizės, kovariacinės analizės uždaviniai. Trumpai aptariami apibendrintieji tiesiniai modeliai ir logistinė regresija.

519.2(075.8)

Vlijandas Bagdonavičius, Julius Jonas Kruopis
Matematinė statistika. II dalis. Tiesiniai modeliai
Vadovėlis

Lietuvių kalbos redaktorė *Danutė Petrauskienė*
Maketuotoja *Rūta Levulienė*

Išleido *Vilniaus universiteto leidykla*