REMIGIJUS LAPINSKAS

Computer Labs

PRACTICAL ECONOMETRICS. I. REGRESSION MODELS

PRAKTINĖ EKONOMETRIJA. I. REGRESINIAI MODELIAI

KOMPUTERINĖS PRATYBOS

remigijus.lapinskas@mif.vu.lt

Vilnius 2013.12

Contents

Introduction

These computer labs are designed to accompany the Lecture notes „R. Lapinskas, Practical econometrics.I. Regression models“ http://uosis.mif.vu.lt/~rlapinskas/. We shall interchangeable use two free software programs, GRETL and R.

GRETL is an econometrics package, including a shared library, a command-line client program and a graphical user interface. User-friendly GRETL offers an intuitive user interface; it is very easy to get up and running with econometric analysis. Thanks to its association with the econometrics textbooks by Ramu Ramanathan, Jeffrey Wooldridge, and James Stock and Mark Watson, the package offers many practice data files and command scripts. These are well annotated and accessible. Two other useful resources for GRETL users are the available documentation and the GRETL-users mailing list.

We assume that the reader has some knowledge about GRETL http://GRETL.sourceforge.net/win32/. For the newbies we recommend the author‘s Lecture Notes *A Very Short Introduction to Statistics with GRETL* http://uosis.mif.vu.lt/~rlapinskas/ShortStatGRETL/ or Adkins‘ Using GRETL for Principles of Econometrics, 3rd Edition Version 1.313 (see http://www.LearnEconometrics.com/GRETL.html), or T.Kufel‘s Ekonometria. Rozwiązywanie problemow z wykorzystaniem programu GRETL, Warszawa: Wydawnictwo Naukowe PWN, 2011 (can be found in the MIF library stock). Very useful is also GRETL‘s Users Guide which is dowloaded together with GRETL. GRETL allows to perform analysis from the pull-down menus or using proper commands that can be executed in the console or as a script using words only. More complex series of commands may require you to use the GRETL script facilities which basically allow you to write simple programs in their entirety, store them in a file, and then execute all of the commands in a single batch.

We also assume that the reader knows some basic fact about R. There are many introductory books on R, including the author‘s *Įvadas į statistiką su R* (see uosis.mif.vu.lt/~rlapinskas/). In R, you run your procedures interactively, entering commands at the command prompt and seeing the results of each statement as it is processed. Occasionally, you may want to run an R program in a repeated, standard, and possibly unattended fashion. For example, you may need to generate the same report once a month. You can also write your program in R and run it in batch mode.

1. First Steps

These Computer Labs are assumed to be performed with GRETL or R.

GRETL is an open-source statistical package, mainly for econometrics (econometrics is a part of statistics dealing mainly with economic models and/or economic data). The name is an acronym for *G*nu *R*egression, *E*conometrics and *T*ime-series *L*ibrary. The product can be freely downloaded from http://gretl.sourceforge.net/.

R is an open source programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians for developing statistical software and data analysis. R is an interpreted language typically used through a command line interpreter. The capabilities of R are extended through user-created *packages*, which allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, etc. R uses a command line interface; however, several graphical user interfaces are available for use with R. To download R, please choose your preferred CRAN mirror.

1.1. GRETL Basics

There are several different ways to work in GRETL. Until you learn to use GRETL's rather simple and intuitive language syntax, the easiest way to use the program is through its built in graphical user interface (GUI). The graphical interface should be familiar to most of you. Basically, you use your computer's mouse to open dialog boxes. Fill in the desired options and execute the commands by clicking on the OK button. Those of you who grew up using MS Windows will find this way of working quite easy. GRETL is using your input from the dialogs, delivered by mouse clicks and a few keystrokes, to generate computer code that is executed in the background.

GRETL offers a command line interface as well. In this mode you type in valid GRETL commands either singly from the console or in batches using scripts. Once you learn the commands, this is surely the easiest way to work. If you forget the commands, then return to the dialogs and let the graphical interface generate them for you.

GRETL is a very user-friendly program. However, in case of trouble, search for help. For example, if you want to perform *weighted regression*, it could be not quite clear where to find the necessary menu section or command. An approximate sequence of steps could be as follows:

- use the *Find text* button to search for *weighted regression* in the *Lecture notes* (LN) of this course
- use the *Find text* button and go through the *Computer Labs*
- open GRETL and search through the Help section
- search through the GRETL User's Guide
- open the Lee C.Adkins text *Using gretl for Principles of Econometrics, 3rd ed.* and search it through
- use Google and search for *gretl weighted regression*

Most of the procedures can be performed from the dialog boxes or from the script window. If you do not remember necessary commands, try to perform the procedure with the help of GUI, and, say, in the Model| Other linear models| Weighted Least Squares… box you will find Help button. Also, whichever is your session, performed with GUI or through the sript or scripts, when closing GRETL you can save the command variant of your work as an *.inp file.

1.2. Examples of Regression Models

This section accompanies Ch. 2 from the Lecture Notes *Practical Econometrics.I. Regression Models* (LN).

1.1 example. GRETL This example parallels 2.1 example from LN (but uses another data set). We shall use the pull-down menus first: open GRETL and go to File|Open data|Sample file...|Gretl and double-click on mroz87 (this is a data set on women‘s labor force participation and pay). Select two variables, `HE` and `HW`, where

`HE`	husband‘s educational attainment, in years
`HW`	husband‘s wage, in 1975 dollars

right-click on the selection and choose XY scatterplot; select `HE` variable as X-axis variable| OK – the scatter diagram obtained also contains the estimated regression line $\hat{E}(HW \mid HE) = \hat{\beta}_0 + \hat{\beta}_1 HE = 0.578 + 0.553\ HE$.

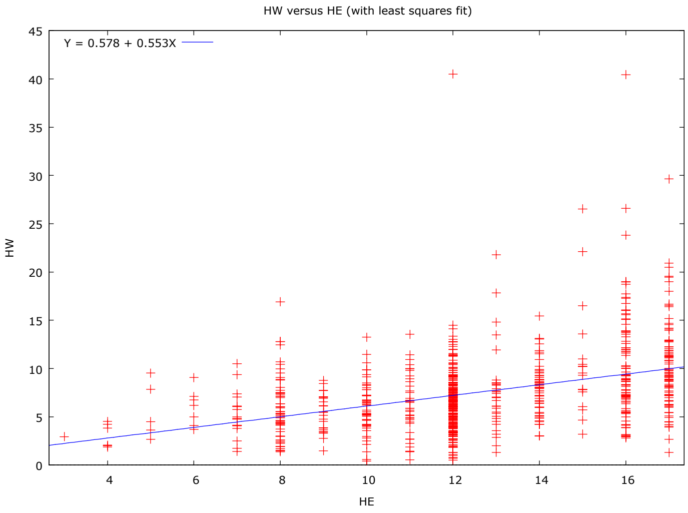
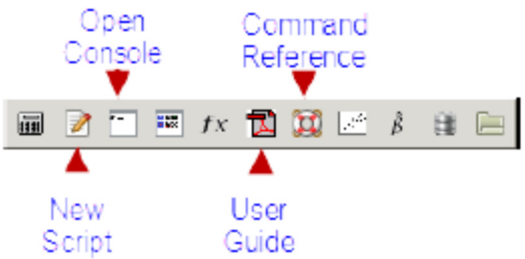


Figure 1.1. `HE` – `HW` scatter diagram with regression line (better education implies higher wages)

Another variant to obtain the graph is as follows: choose Model|Ordinary Least Squares...| select `HW` and choose it as Dependent variable, select `HE` and choose it as Independent variables, click OK; in the model window, choose Graphs|Fitted, actual plot|Against `HE`.

The same result can be obtained from the script window: open the script window (to do this, click on the second from the left icon on the bottom of the GRETL window),



then copy and paste there the text that follows:

```
open mroz87
gnuplot HW HE --output=display
```

or

```
open mroz87
ols HW 0 HE           # create a regression model
series yh = $yhat     # save predicted values
gnuplot HW yh HE --with-lines=yh --output=display
```

R The figure in LN, p.2-1, was drawn with the following R script:

```
mroz87 = read.table(file.choose(),header=TRUE)
attach(mroz87)
educ=HE
wage=HW
plot(jitter(educ),wage,col=2)
points(3:17,tapply(wage,educ,mean),pch=15)
abline(lm(wage~educ))
```

To import mroz87 data, when using `file.choose()`, navigate to the ...\PEdata directory.◀

1.2 example. GRETL Open GRETL and import the file shampoo.txt: File| Open data| Import| text/CSV...| OK| PEdata | select shampoo.txt| give the data a time series interpretation – it is a monthly time series beginning at 1980:01 (click yes)| etc. To get an impression of the series, select shampoo in GRETL window, right-click, and choose Time series plot:

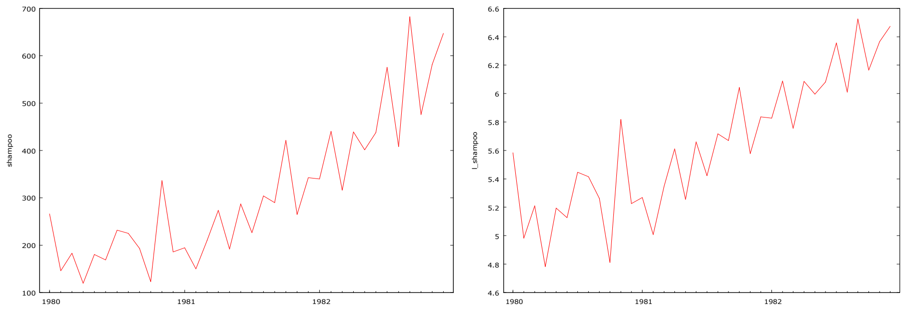


Figure 1.2. The sales (left) on average follow a parabola and we want to find its parameters; it seems that the log(sales) (right) may well be described as a straight line

1. We start with analysis in the script window

```
# Fig. 2.2 in LN, left
gnuplot shampoo --time-series --with-lines --output=display
genr time
series sq_time = time * time
ols shampoo 0 time           # create linear model
series sh_lin = $yhat
ols shampoo 0 time sq_time   # create quadratic model
series sh_sq = $yhat
```

It is easy to fit in a similar manner a linear trend to $\log(shampoo)$: `l_shampoo` $=$ $\log(bb) + cc*$`time`$+\varepsilon$, but it is more problematic to fit a non-linear exponential trend to `shampoo` itself: `shampoo`$=$ $bb*exp(cc*$`time`$)+\varepsilon$ (the usual problem is to initialize the *bb* and *cc* parameters in the iterative procedure of nonlinear regression). We may take the starting values from a similar linear model.

```
# create nonlinear exponential model
logs shampoo # creates log(shampoo)
ols l_shampoo 0 time               # create a linear model for l_shampoo

genr bb = exp($coeff(0))           # zeroth iteration of bb
genr cc = $coeff(time)             # zeroth iteration of cc
genr aa = 100 # we generalize the initial model by adding the intercept aa
nls shampoo = aa + bb*exp(cc*time) # create nonlinear model
params aa bb cc
end nls
series sh_exp = $yhat
gnuplot   shampoo   sh_lin   sh_sq   sh_exp   --time-series   --with-lines   --
output=display
```

We shall forecast `shampoo` 12 month ahead with the exponential model:

```
# Fig. 2.2 in LN, right
addobs 12
smpl 1980:1 1983:12
fcast 1980:1 1983:12 sh_f_exp
gnuplot shampoo sh_f_exp --time-series --with-lines --output=display
```

2. Having exposed the ideas, we shall demonstrate how to implement the above procedures through the pull-down menus. Import shampoo.txt anew and go to Add| Time trend, then select `time` and go to Add| Squares of selected variables (we are preparing the ground for the linear and quadratic models). To create these models, go to Model| Ordinary Least Squares...| choose `shampoo` as Dependent variable and `time` as Independent variable|OK. In the model window, go to Save| Fitted values| OK. Then repeat the procedure, adding `sq_time` to the Independent variable box| OK and saving the fitted value as `yhat2`. We skip the nonlinear exponential model and, in the main GRETL window, select `shampoo`, `yhat1` and `yhat2`, right-click on the selection and choose Time series plot (see Fig. 1.3, left).

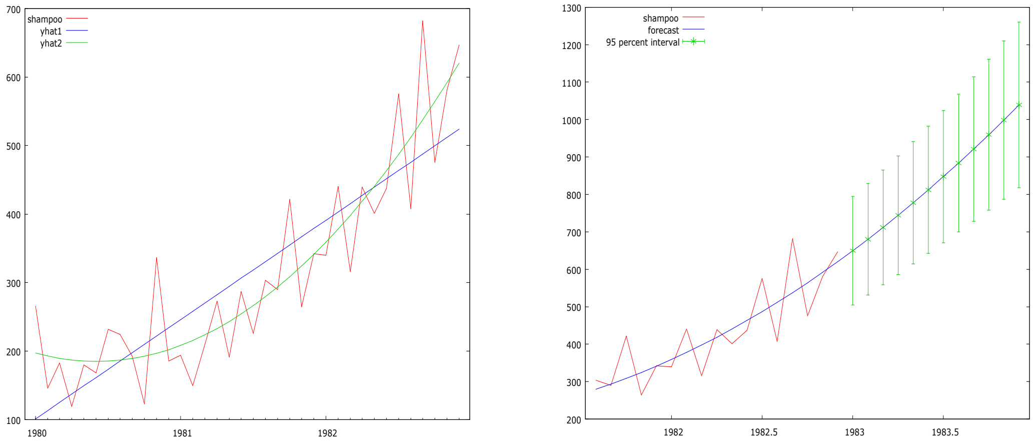


Figure 1.3. Linear and quadratic models (left); the 12 months forecast of the quadratic model (right)

To predict the quadratic model 12 months ahead, go to Data|Add observations...|12|OK, then in the model 2 window choose Analysis|Forecasts...|OK (see Fig. 1.3, right).

R

Now we shall do the same with R. Copy and paste the following into R's File| New script window (study each line!)

```
shamp = ts(scan(file.choose(),skip=1),freq=12,start=1980)
par(mfrow=c(1,2))
plot(shamp,main="Three regression models")
tt=seq(1980,1983-1/12,by=1/12)
tt.new=seq(1980,1984-1/12,by=1/12)
lines(tt,predict(lm(shamp~tt)),col=2)
shamp.sq=lm(shamp~tt+I(tt^2))
lines(tt,predict(shamp.sq),col=3)

## nonlinear model
ttt=1:36
ttt.new=1:48
shamp.exp=nls(shamp~aa + bb*exp(cc*ttt),start=list(aa=100,bb=10,cc=0.07))
summary(shamp.exp)
lines(tt,predict(shamp.exp),col=4)
legend(1980,650,c("lin","sq","exp"),lty=1,col=2:4)

## 12-months-ahead forecast
plot(shamp,xlim=c(1980,1984),ylim=c(100,1200),main="Two forecasts")
lines(tt.new,predict(shamp.sq,newdata=data.frame(tt=tt.new)),col=3)
lines(tt.new,predict(shamp.exp,newdata=data.frame(ttt=ttt.new)),col=4)
legend(1980,1150,c("sq","exp"),lty=1,col=3:4)
```

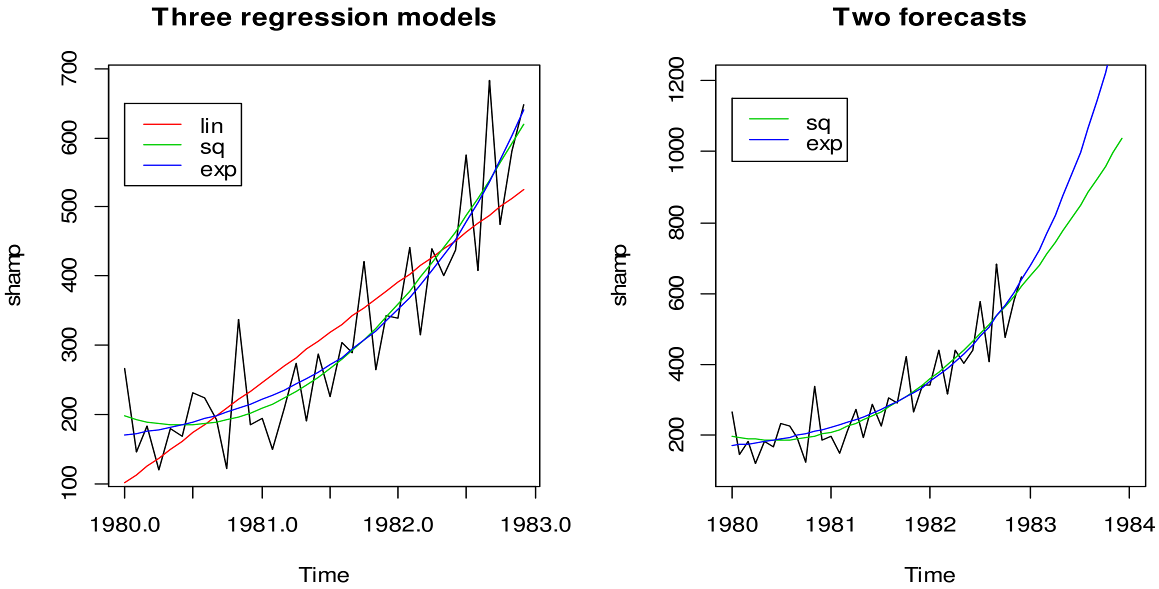


Figure 1.4. Three `shampoo` regression models (linear, square and exponential) (left) and two 12-months-ahead forecasts (using square and exponential models) (right)

1.3 example. R The data set …\PEdata\whh.txt contains 1428 observations on weight (in pounds), height (in inches) and sex (`male`=1 for males):

```
weight height male
147     68      0
195     74      1
165     66      1
..................
```

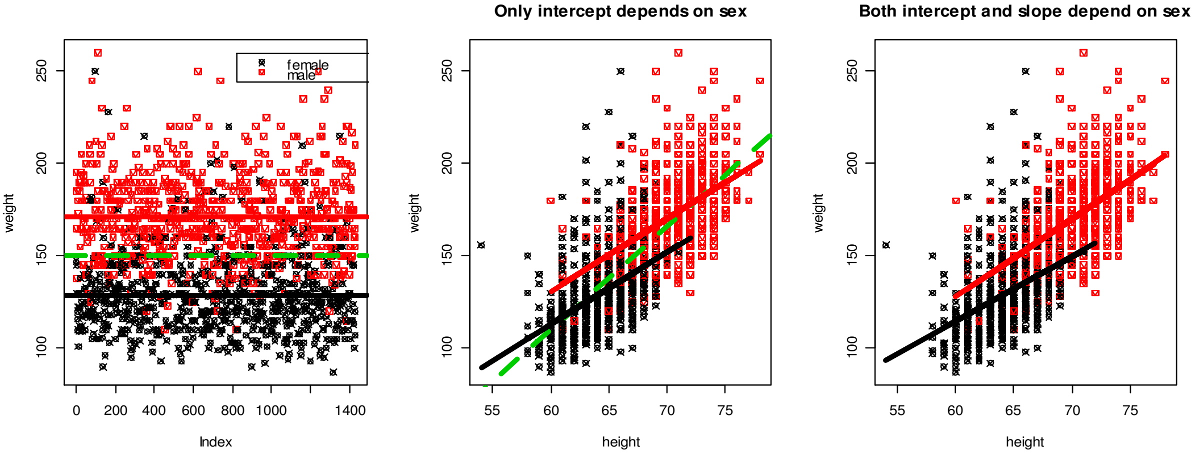


Figure 1.5. The green dashed line stands for the sample mean in the whole sample of `weight` while black and red are for the female and male subsamples, respectively (left); the green dashed line is a regression line in the whole `height-weight` scatter diagram whereas black and red in female and male subsamples, respectively (center); the most accurate models allow both intercept and slope to depend on sex (right)

We want to create a model for `weight`. The simplest model is $weight = \beta_0 + \varepsilon$ where the OLS estimation of β_0 is just the sample mean of `weight` (the green dashed line in Fig.1.5, left). Clearly, sex gives some information on weight, thus we can improve the model by adding `male` into it and considering $weight = \beta_0 + \beta_1 male + \varepsilon$:

$$weight = \begin{cases} 128.98 & \text{if } male = 0 \\ 171.24 & \text{if } male = 1 \end{cases}$$

(black and red lines in Fig. 1.5, left). We get still better model if we start controlling `height`: $weight = \beta_0 + \beta_1 height + \varepsilon$ or $weight = \beta_0 + \beta_1 male + \beta_2 height + \varepsilon$ or even $weight = \beta_0 + \beta_1 male + (\beta_2 + \beta_3 male) \cdot height + \varepsilon$ (in the second model only the intercept differs for males and females and in the third both intercept and slope differ for males and females). It can be shown (using the AIC coefficient which will be discussed later) that the last model, namely

```
lm(formula = weight ~ male + height + I(male * height))

Coefficients:
                 Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)      -95.8489  18.1114     -5.292    1.4e-07 ***
male             -31.4831  25.2537     -1.247    0.2127
height           3.5081    0.2824      12.423    < 2e-16 ***
I(male * height) 0.7420    0.3774      1.966     0.0494 *
```

or

$$weight = \begin{cases} -95.85 + 3.51\, height & \text{if } male = 0 \\ -127.33 + 4.25\, height & \text{if } male = 1 \end{cases}$$

is the most accurate (see Fig.1.5, right). The model can be interpreted as follows: if the female‘s height increases by 1 inch, her weight on average increases by 3.51 pound and if the male‘s height increases by 1 inch, his weight on average increases by 4.25 pound.

To perform the above mentioned calculation, we use the following script (the script will become more transparent if you read and copy+paste it in three portions):

```
whh=read.table(file.choose(),header=T) # go to whh.txt in ...\PEdata
head(whh)
attach(whh)
par(mfrow=c(1,3))
plot(weight,pch=male+13,col=male+1)
abline(mean(weight),0,lwd=4,col=3,lty=2)
legend(820,260,c("female","male"),pch=male+13,col=male+1)
wh.mod1=lm(weight~male)
summary(wh.mod1)
abline(128.98,0,lwd=4)
abline(128.98+42.26,0,lwd=4,col=2)

plot(height,weight,pch=male+13,col=male+1,main="Only intercept depends on
sex")
wh.mod2=lm(weight~height)
summary(wh.mod2)
abline(wh.mod2,lwd=4,col=3,lty=2)
```

```
wh.mod3=lm(weight~height+male)
summary(wh.mod3)
lines(height[male==0],predict(wh.mod3)[male==0],col=1,lwd=4)
lines(height[male==1],predict(wh.mod3)[male==1],col=2,lwd=4)

plot(height,weight,pch=male+13,col=male+1,main="Both intercept and slope
depend on sex")
wh.mod4=lm(weight~male+height+I(male*height))
summary(wh.mod4)
lines(height[male==0],predict(wh.mod4)[male==0],col=1,lwd=4)
lines(height[male==1],predict(wh.mod4)[male==1],col=2,lwd=4)
```

1.1 exercise. [redacted] The file br2-dat.txt contains data on 1080 houses sold in Baton Rouge, Louisiana, during mid-2005. The data include sale price, the house size in square feet, its age, whether it has a pool or fireplace or is on the waterfront. Also included is an indicator variable `trad` indicating whether the house style is traditional or not. Variable descriptions are in the file br2-def.txt. Do the exercise in both GRETL and R.

(a) Plot house price against house size for houses with traditional style.

(b) For the traditional-style houses estimate the linear regression model $\texttt{price} = \beta_0 + \beta_1\,\texttt{sqft} + \varepsilon$. Add the regresion line to the (a) graph.

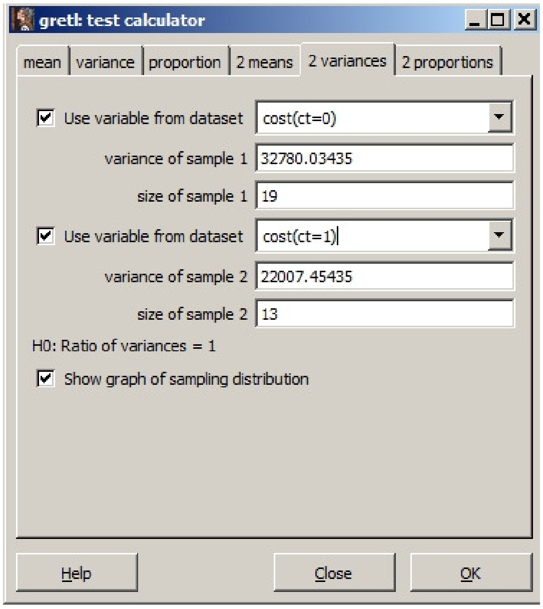
(c) For the traditional-style houses estimate the quadratic regression models $\texttt{price} = \beta_0 + \beta_1\,\texttt{sqft}^2 + \varepsilon$ and $\texttt{price} = \beta_0 + \beta_1\,\texttt{sqft} + \beta_2\,\texttt{sqft}^2 + \varepsilon$. Add the fitted curves `price2` and `price3` to the scatter diagram. Which of the three models seems to be the best?

(d) For the traditional-style houses estimate the log-linear regression model $\log\texttt{price} = \beta_0 + \beta_1\texttt{sqft} + \varepsilon$. Create a new variable `price4` equal to the exponent of the fitted values of this model. Draw a scatter diagram together with `price3` and `price4`.

1.2 exercise. The file table2.5.txt consists of 11 columns and 32 rows, it contains data characterizing the construction of atomic power plants in the USA:

`cost`	cost of the construction
`date`	date when the construction permission was issued
`t1`	...
`t2`	...
`cap`	the power plant capacity
`pr`	...
`ne`	...
`ct`	1 if the cooling tower is present
`bw`	...
`n`	...
`pt`	1 if it the plant is the partial turn key one

Use GRETL to estimate the sample means, variances and the coefficient of correlation of `cost` ir `cap`. Do variances and means differ in the two groups corresponding to `ct`=0 or =1? – go to Tools| Test statistic calculator| 2 variances and fill the window as shown:



gretl: test calculator

☑ Use variable from dataset: cost(ct=0)

variance of sample 1: 32780.03435

size of sample 1: 19

☑ Use variable from dataset: cost(ct=1)

variance of sample 2: 22007.45435

size of sample 2: 13

H0: Ratio of variances = 1

☑ Show graph of sampling distribution

Is the distribution of `cost` normal in both groups? – extract the observation group with `ct`=0 through Sample| Restrict, based on criterion... | `ct`=0| select `cost` and right-click on it| select Frequency distribution and check Test against normal distribution. Create two regression models with all the data: $\texttt{cost} = \beta_0 + \beta_1 \texttt{cap} + \varepsilon$ and $\texttt{cost} = \beta_0 + \beta_1 \texttt{cap} + \beta_2 \texttt{ct} + \varepsilon$. How do you interpret the second model? Plot necessary graphs. Do the same with R.

2. UNIVARIATE REGRESSION MODELS

Later in this chapter, we shall study the data set CPS1985.txt:

```
ID wage education experience age ethnicity region gender occupation        sector union married
1   4.95         9          42  57      cauc  other female      worker manufacturing    no     yes
2   6.67        12           1  19      cauc  other   male      worker manufacturing    no      no
3   4.00        12           4  22      cauc  other   male      worker         other    no      no
4   7.50        12          17  35      cauc  other   male      worker         other    no     yes
5 13.07         13           9  28      cauc  other   male      worker         other   yes      no
.....................................................................................
```

where

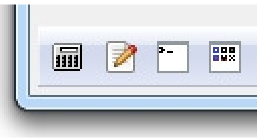
`wage` wage (in dollars per hour)
`education` number of years of education
`experience` number of years of potential work experience (age - education - 6)
`age` age in years
`ethnicity` "cauc" (→1), "hispanic"(→3), or "other"(→2)
`region` does the individual live in the South? (South → 2, Other → 1)
`gender` gender (Female → 1, Male → 2)
`occupation` factor with levels "worker" (tradesperson or assembly line worker), „technical" (technical or professional worker), "services" (service worker), "office" (office and clerical worker), "sales" (sales worker), "management" (management and administration)
`sector` "manufacturing" (manufacturing or mining), "construction", "other"
`union` does the individual work on a union job?
`married` is the individual married?

Note that the string (or nominal) variables when imported to GRETL are recoded to numbers (for example, ethnicity takes on values `cauc`, `hispanic`, `other`, therefore they will be recoded to numbers 1, 3, 2):

```
One or more non-numeric variables were found.
Gretl cannot handle such variables directly, so they
have been given numeric codes as follows.

String code table for variable 6 (ethnicity):
  1 = 'cauc'
  2 = 'other'
  3 = 'hispanic'
```

etc. We want to understand how the variable `wage` relates to other variables and also to get some numerical characteristics of the goodness-of-fit of a model.

There are several different ways to work in gretl: 1) through its built in graphical user interface (GUI) and 2) command line interface (in this mode you type in valid gretl commands either singly from the console or in batches using scripts (the second icon from the left is the script[1] window, the third one is console[2]).



[1] A "script" is a file containing a sequence of gretl commands.

[2] Type commands and execute them one by one (by pressing the Enter key) interactively.

This chapter deals with a *univariate regression* case – we analyse the dependence of `wage` on only one variable, say, `experience` (that is, `education`, `age` and other variables will be included into the error term ε):

$$\texttt{wage} = f(\texttt{experience}) + \varepsilon .$$

The most simple, though not always the most appropriate candidate for a regression curve is a straight line: $y = f_1(x) = \beta_0 + \beta_1 x$ (the coefficient β_0 is called an *intercept* while β_1 the *slope* of the regression line); a bit more complicated model is given by the quadratic curve or parabola: $y = f_2(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. In what follows, we use the so-called ordinary least squares (OLS) method to find the *estimates* of these coefficients. To find the estimates in GRETL, after importing CPS1985.txt, we start with a linear model and go to Model|Ordinary Least Squares...| move `wage` to Dependent variable box and `experience` to Independent variables box|OK. We get the following Model 1:

```
Model 1: OLS, using observations 1-533
Dependent variable: wage

               coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const         8,38474      0,389135      21,55     1,83e-074 ***
  experience    0,0362978    0,0179370      2,024    0,0435     **

Mean dependent var   9,031426   S.D. dependent var   5,141105
Sum squared resid    13953,66   S.E. of regression   5,126215
R-squared            0,007653   Adjusted R-squared   0,005784
F(1, 531)            4,095074   P-value(F)           0,043509
Log-likelihood      -1626,410   Akaike criterion     3256,821
Schwarz criterion    3265,378   Hannan-Quinn         3260,169

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 223,98
  with p-value = 2,30886e-049
```

In the table, the OLS estimate of the linear regression model $\widehat{\texttt{wage}} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{experience}$ = 8.385 + 0.036·`experience` is presented (the numbers $\hat{\beta}_0$ and $\hat{\beta}_1$[3] are the *estimates* of unknown coefficients β_0 and β_1.) Other important numbers in the table are 1) the p - value 0.0435 which informs us that the hypothesis $H_0 : \beta_1 = 0$ must be rejected[4], i.e., the term `experience` is *significant* in our model, i.e., we cannot remove `experience` from the model, 2) the *coefficient of determination* `R-squared` which is always between 0 and 1 (the more the better), now it indicates that `experience` explains only 0.765% of `wage` variation (this means that 99.235% of this variation remains unexplained – it is the first indication that our model is not satisfactory, it lacks some important ingredients.)

[3] What is the meaning of $\hat{\beta}_1$ (=0.036)? Take any two strata (layers, slices) of workers; if the experience in the first strata is 1 (year) higher, then the wage in this stratum will be on average 0.036 (dollars per hour) higher.

[4] Because it is less than 0.05.

The estimates and p- values are estimated correctly only if a model satisfies certain conditions. The most important are:

1) The spread (variance) of errors must be constant, i.e., $\operatorname{var}\varepsilon_i \equiv \sigma^2$
2) The errors must be uncorrelated, i.e., $cor(\varepsilon_i, \varepsilon_j) \equiv 0$, and
3) The errors must have a normal distribution, i.e., $\varepsilon_i \sim N(0, \sigma^2)$

Prior to testing these hypotheses, in order to get some intuition about the model, we shall plot a scatter plot with a regression line. In the Model 1 window, go to Graphs * Fitted, actual plot * Against experience:

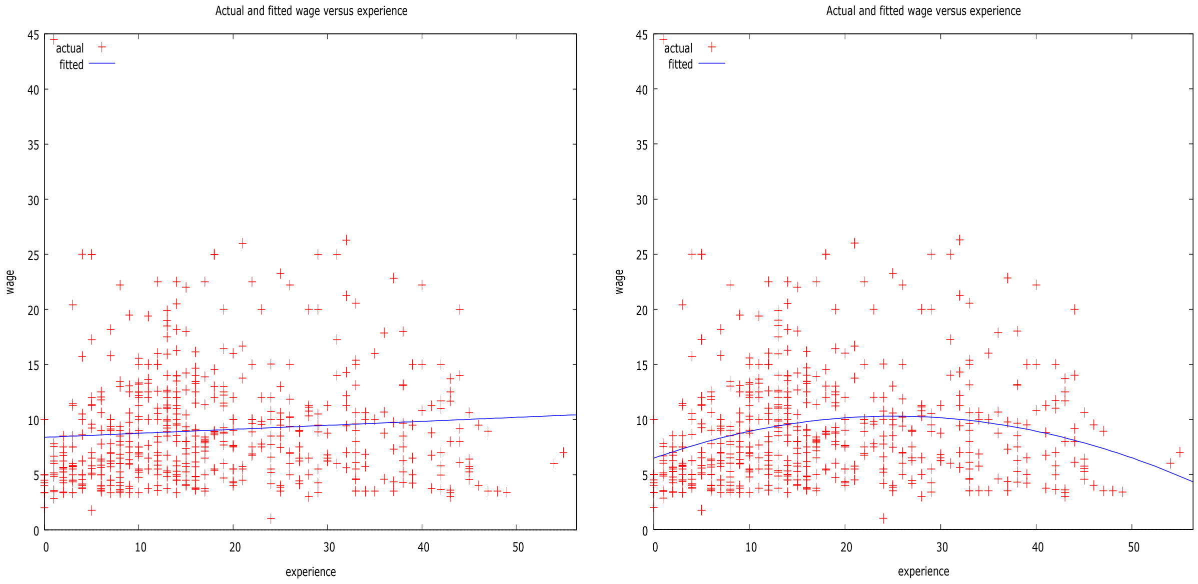


Figure 2.1. Linear (left) and quadratic (right) models. When a worker gets older (and his experience increases), his wage finally begins to decrease (thus, economically speaking, the parabolic model is more appropriate)

In Fig. 2.1, we can see that `wage` hardly depends on `experience` alone (the model claims that when `experience` changes from 0 to 50 `wage` increases only from roughly 8 to 10.) The distribution of residuals is not symmetric (why?), i.e., nonnormal[5]. Thus, we must either look for another functional form[6] of the dependence or include new variables into the model. The latter variant leads to multivariate regression and will be discussed later, now we replace linear dependence by the parabolic one. In order to be able to compare graphs in the future, go to Save|Fitted values|yhat1|OK.

To create a quadratic model, we have to append the list of our variables with a square of `experience`: in GRETL's window select `experience` and go to Add|Squares of selected variables|OK (a new variable, `sq_experience`, will appear in GRETL's window.) Now go

[5] The bottom line of Model 1 table shows that the p- value of the hypothesis H_0 : *errors are normal* equals 2,30886e-049 (<0.05), that is we (reject or accept?) (which hypothesis?).

[6] Suitable candidate is a parabola.

to Model|Ordinary least squares…|fill Dependent variable box with `wage` and append Independent variables box with `sq_experience`|OK.

Table 2.1

```
Model 2: OLS, using observations 1-533
Dependent variable: wage

                  coefficient    std. error    t-ratio     p-value
  ---------------------------------------------------------------
  const            6,46409       0,568640      11,37      6,02e-027 ***
  experience       0,304855      0,0614651      4,960     9,52e-07  ***
  sq_experience   -0,00608515    0,00133432    -4,560     6,35e-06  ***

Mean dependent var   9,031426   S.D. dependent var   5,141105
Sum squared resid    13426,77   S.E. of regression   5,033243
R-squared            0,045124   Adjusted R-squared   0,041520
F(2, 530)            12,52287   P-value(F)           4,85e-06
Log-likelihood      -1616,152   Akaike criterion     3238,305
Schwarz criterion    3251,140   Hannan-Quinn         3243,328

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 219,189
  with p-value = 2,53379e-048
```

Both variables (`experience` and `sq_experience`) in Model 2 are significant[7]. To choose between two or more models with the same left-hand side variable, use the Akaike and/or Schwarz criteria (choose the model with the smallest value of the criterion, thus, in our case select Model 2.) Note that its `R-squared` is still very low, residuals are nonnormal, thus the model is still unsatisfactory. We will improve it in the next chapter by adding new variables.

2.1 exercise. Regress `experience` on `age` and analyse the model. How do you interpret the coefficient $\hat{\beta}_1$? Add `sq_age` to the model and analyze it. Which of the two models is better?

2.1. The output of the `ols` procedure

We have already not once seen the output of the gretl's `ols` (ordinary least squares) procedure (c.f. Table 2.1). A detailed exposition of all the concepts is presented in the LN, here is a short explanation of the most important parameters and their meaning.

- **The method of (ordinary) least squares (OLS)**

Let us assume that the data generating process (DGP) is described by a k – variate (in our case $k=1$) regression, i.e., the observations $(Y_i, X_{1i}, ..., X_{ki}), k = 1, ..., N$, are defined by the system of equations

[7] Note the rule – if the square term is significant, do not remove linear term.

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1} + \varepsilon_1 \\ .... \\ Y_N = \beta_0 + \beta_1 X_{1N} + ... + \beta_k X_{kN} + \varepsilon_N \end{cases}$$

or, in a more compact form, $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$ where

$$\vec{Y} = \begin{pmatrix} Y_1 \\ ... \\ Y_N \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1\, X_{11} \; ... \; X_{k1} \\ .... \\ 1\, X_{1N} ... \; X_{kN} \end{pmatrix}, \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ .... \\ \beta_k \end{pmatrix}, \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_N \end{pmatrix}.$$

The values of $b_m, m = 1,...,k,$ which minimize the residual sum of squares $RSS(b_0, b_1, ..., b_k) = \sum_{i=1}^{N} (Y_i - (b_0 + b_1 X_{1i} + ... + b_k X_{ki}))^2$ are called the ols estimator of unknown parameters β_m and denoted by $\hat{\beta}_m$ or $\hat{\beta}_m^{OLS}$; the differences $Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki}\right) = \hat{\varepsilon}_i$ are called the residuals of the model and the expression $RSS = RSS(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k) = \sum (Y_i - \hat{Y}_i)^2$ the `Sum squared resid`. It can be shown that $\hat{\vec{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{Y}$ [8], or, in a univariate case,

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \dfrac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (Y_i - \bar{Y}) X_i}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \dfrac{\widehat{\text{cov}}(X,Y)}{\widehat{\text{var}} X} \end{cases}$$

The meaning of β_1 can be explained as follows: take two stratas of our population with $X = x$ and $X = x + 1$; then the mean value of Y in the second strata will differ from the first one by β_1 (these words are often replaced by „if X increases by 1, Y will change by β_1“). Note that is does not mean that if, for example, someone studies one more year, his/her wage will automatically increase by β_1; it only means that he/she will get to another strata where on average the wage is higher by β_1 (in fact, we only know the approximation of β_1, namely $\hat{\beta}_1$).

- **Standard errors**

The (unknown) variance of the error terms $\text{var}\,\varepsilon_i \equiv \sigma^2$ is estimated by $\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = RSS/(N-k) = \sum \hat{\varepsilon}_i^2 / (N-k)$ [9] (the square root of this number is called[10] `S.E. of regression`). The variance-covariance matrix of the random estimator $\hat{\vec{\beta}}$ equals

[8] The coefficients in Table 2.1 were estimated using this formula.
[9] In „good“ models it ought to be a „small“ number.
[10] Standard Error of regression.

$\widehat{\text{cov}}\hat{\beta} = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$; the square roots of the numbers on the diagonal of this matrix are termed as `std.error` in Table 2.1 and stand for the standard errors of the estimators of respective coefficients, i.e., $\sqrt{\widehat{\text{var}}\hat{\beta}_m}$. The 95% confidence interval for β_m is $(\hat{\beta}_m - 2\sqrt{\widehat{\text{var}}\hat{\beta}_m},\ \hat{\beta}_m + 2\sqrt{\widehat{\text{var}}\hat{\beta}_m})$, thus if it covers, for example, 1, we do not reject the hypothesis $H_0 : \beta_m = 1$ (with 5% confidence level). Recall that in univariate case $\widehat{\text{var}}\hat{\beta}_1 = s_\varepsilon^2 / \sum (X_i - \bar{X})^2$.

- **`t-ratio`**

`t-ratio` statististics (or `t value` in R) is used for testing the hypothesis $H_0 : \beta_m = 0$ against the alternative $H_1 : \beta_m \neq 0$ (in other words, H_0 tests whether X_m is significant). The value of this statistics is calculated as $t_m = \hat{\beta}_m / \sqrt{\widehat{\text{var}}\hat{\beta}_m} \left(= 0{,}304855 / 0{,}0614651 = 4{,}960\right)$. If the modulus of this number is greater than the $(1-\alpha/2)$ - quantile of the Student r.v. T_{N-k}, we reject H_0 at the α significance level. Note that this quantile is approximately equal to 2, therefore if $|t_m| > 2$, we reject H_0 (X_m is significant). Similar, but more accurate way to test H_0, is to use the `p-value` of this test.

- **`p-value`**

The last column of the output table contains the `p-value` (or `Pr(>|t|)` in R). The number is calculated using the formula `p-value` = $P(|T_{N-k}| > |t_m|)$. If the `p-value` $< \alpha$ (usually, $\alpha = 0.01, 0.05$ or, sometimes, 0.1), then respective variable is significant and we should not remove it from the model of DGP.

- **The coefficient of determination (`R-squared`)**

It can be easily shown that the total sample variation or the total sum of squares $TSS = \sum (Y_i - \bar{Y})^2$ can be decomposed into the sum of RSS and explained sum of squares $ESS = \sum (\hat{Y}_i - \bar{Y})^2$: $TSS = RSS + ESS$. The coefficient of determination or R^2 is the the proportion of variation in Y explained by X within the regression model: $R^2 = ESS / TSS = 1 - RSS / TSS$. It is easy to verify that $R^2 = \max_{b_0, b_1, \ldots, b_k} cor^2(Y, b_0 + b_1 X_1 + \ldots + b_k X_k) = cor^2(Y, \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k)$, i.e., OLS finds a linear combination of *X*es which correlates with Y best. For „good“ models R^2 is close to 1 (for example, if $R^2 = 0.51$, we say that the model explains 51% of response variability). If one has two different models for Y with the same number of explanatory variables, the model with higher R^2 has better predictive properties. However, R^2 always increases when we augment the model with new variables, therefore to compare nested models by their coefficients of determination is incorrect (in such a situation, one has to use the Akaike or similar information criteria, see below).

The above formula for the coefficient of determination applies only for the models containing the intercept β_0. However, if the DGP is described by $Y_i = \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, then the traditional definition of R^2 is inadequate (R^2 can take negative values). In such a case, the coefficient of determination should be estimated through[11] the formula $R_{-}^2 = \sum \hat{Y}_i^2 / \sum Y_i^2$ (generally, R_{-}^2 is higher then the standard R^2). However, to compare these two coefficients in order to opt for the „right" model is not correct (these coefficients are not comparable). The usual recommendation is as follows: keep the intercept in the model if 1) it is significant, and 2) the `S.E. of regression` $\hat{\sigma}^2$ (of the model with the intercept) is less than $\hat{\sigma}_{-}^2$.

- **`Adjusted R-squared`**

This coefficient is estimated by the formula

$$\overline{R^2} = 1 - \frac{RSS/(N-k-1)}{TSS/(N-1)} = 1 - \left(1 - R^2\right)\frac{N-1}{N-k-1}$$

which penalizes R^2 for the inclusion of additional parameters, other things equal . More popular are informational criteria (see below).

- **`Akaike`** `(information)` **`criterion`**

One of the variants to define the criterion is

$$AIC = \log\frac{RSS}{N} + 2k - N$$

Although the Akaike criterion is designed to favor parsimony, arguably it does not go far enough in that direction. For instance, if we have two nested models with $k-1$ and k parameters respectively, and if the null hypothesis that parameter k equals 0 is true, in large samples the AIC will nonetheless tend to select the less parsimonious model about 16 percent of the time. Whatever the problems are, if you have two models with the same Y on the lhs, choose the one with smaller AIC.

- **`Schwarz`** `(Bayesian information)` **`criterion`**

An alternative to the AIC which avoids the problem of "too small penalty" is the Schwarz Bayesian information criterion (BIC). The BIC can be written as

$$BIC = \log\frac{RSS}{N} + \frac{k}{N}\log N\ .$$

Now the penalty for adding extra parameters grows with the sample size. This ensures that, asymptotically, one will not select a larger model over a correctly specified parsimonious model. Again, choose the model with smaller BIC.

[11] R uses the relevant formula to estimate the coefficients.

2.2 exercise. You have the results of a simple linear regression based on N=52 observations (states of the USA).

(a) The estimated error variance $\hat{\sigma}_\varepsilon^2 = 2.04672$. What is the RSS?

(b) The estimated variance of $\hat{\beta}_1$ is 0.00098. What is the standard error of $\hat{\beta}_1$? What is the value of $\sum(X_i - X)^2$?

(c) Suppose the dependent variable Y_i = the state's mean income (in thousands of dollars) of males who are 18 years of age or older and X_i the percentage of males 18 years or older who are high school graduates. If $\hat{\beta}_1$ = 0.18, interpret this result.

(d) Suppose $\overline{X} = 69.139$ and $\overline{Y} = 15.187$, what is the estimate of the intercept parameter?

(e) Given the results in (b) and (d), what is $\sum X_i^2$?

(f) For the state of Arkansas the value of Y_i=12.274 and the value of X_i=58.3. Compute the least squares residual for Arkansas (hint: use the information in parts (c) and (d)).

2.3 exercise. The file stockton4.dat.txt contains data on 1500 houses sold in Stockton, CA during 1996–1998. Variable descriptions are in the file stockton4.def.txt.

(a) Plot house selling price against house living area for all houses in the sample.

(b) Estimate the regression model $sprice = \beta_0 + \beta_1\, livarea + \varepsilon$ for all the houses in the sample. Interpret the estimates. Draw the fitted line.

(c) Estimate two quadratic models $sprice = \beta_0 + \beta_1\, livarea^2 + \varepsilon$ and $sprice = \beta_0 + \beta_1\, livarea + \beta_2\, livarea^2 + \varepsilon$ for all the houses in the sample. Which of the three models do you prefer (use AIC and BIC)? What is the marginal effect (i.e., $d\,\widehat{sprice} / d\, livarea$) of an additional 100 square feet of living area for a home with 1500 square feet of living area for all of the three models?

(d) In the same graph, plot the fitted lines from the linear and chosen quadratic models. Which seems to fit the data better? Compare the sum of squared residuals (RSS) for the two models. Which RSS is smaller? Which AIC is smaller?

(e) Estimate the regression model in (c) using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Interpret the estimates. How do the estimates compare?

(f) Plot house selling price against age using only houses that are on large lots. Estimate the linear model $sprice = \beta_0 + \beta_1\, age + \varepsilon$. Interpret the estimated coefficients. Repeat this exercise using the log-linear model $\log(sprice) = \beta_0 + \beta_1\, age + \varepsilon$. Based on the plots and visual fit of the estimated regression lines, which of these two models would you prefer? Explain.

(g) Estimate a linear regression $sprice = \beta_0 + \beta_1\, lglot + \varepsilon$ where the indicator $lglot$ identifies houses on larger lots. Interpret these results.

2.4 exercise. With R How much does education affect wage rates? We shall analyze the CPS1985.txt file.

(a) Obtain the summary statistics and histograms for the variables *wage* and *educ*. Discuss the data characteristics.

(b) Estimate the linear regression $wage = \beta_0 + \beta_1\, educ + \varepsilon$ and discuss the results.

(c) Calculate the least squares residuals and plot them against *educ*. Are any patterns evident? If assumptions U1-U3 hold, should any patterns be evident in the least squares residuals?
(d) Estimate separate regressions for males, females, and three ethnic groups. Compare the results.

(e) Estimate the quadratic regression $wage = \beta_0 + \beta_1 educ + \beta_2 educ^2 + \varepsilon$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education, and for a person with 14 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).
(f) Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *wage* and *educ*. Which model appears to fit the data better?
(g) Construct a histogram of $\log(wage)$. Compare the shape of this histogram to that for *wage* from part (a). Which appears more symmetric and bell-shaped?
(h) Estimate the log-linear regression $\log(wage) = \beta_0 + \beta_1 educ + \varepsilon$. Estimate the marginal effect of another year of education on wage, i.e., $d\,\widehat{wage} / d\,educ = d(\exp(\hat{\beta}_0 + \hat{\beta}_1 educ)) /$ $d\,educ$ for a person with 12 years of education, and for a person with 14 years of education. Compare these values to the estimated marginal effects of education from the linear regression in part (b) and the quadratic equation in part (e).

2.2. Choosing a Functional Form

For a curvilinear relationship like that in Fig. 2.2, the marginal effect of a change in the explanatory variable is measured by the slope of the tangent to the curve at a particular point. The marginal effect of a change in X is greater at the point (X_1, Y_1) than it is at the point (X_2, Y_2). As X increases, the value of Y increases, but the slope is becoming smaller. This is the meaning of ‘‘increasing at a decreasing rate.’’ In the economic context of the food expenditure model, the marginal propensity to spend on food is greater at lower incomes, and as income increases the marginal propensity to spend on food declines.

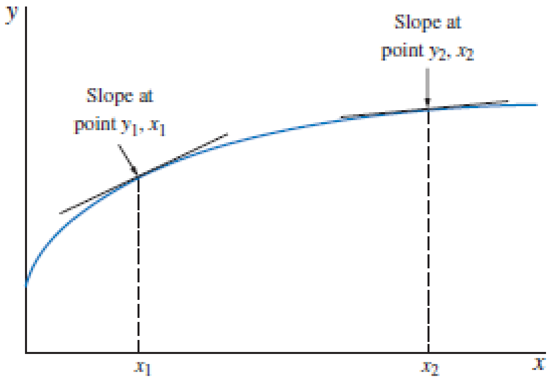


Figure 2.2. A nonlinear relationship between food expenditure and income.

By transforming the variables Y and X we can represent many curved, nonlinear relatiohips and still use the linear regression model.

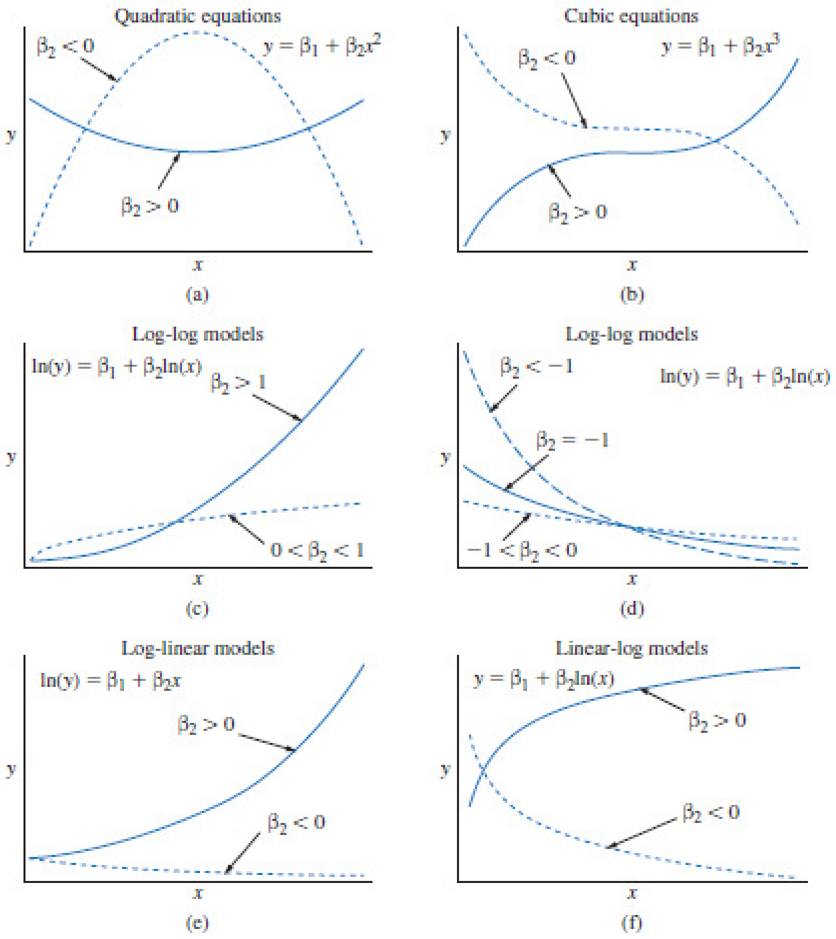
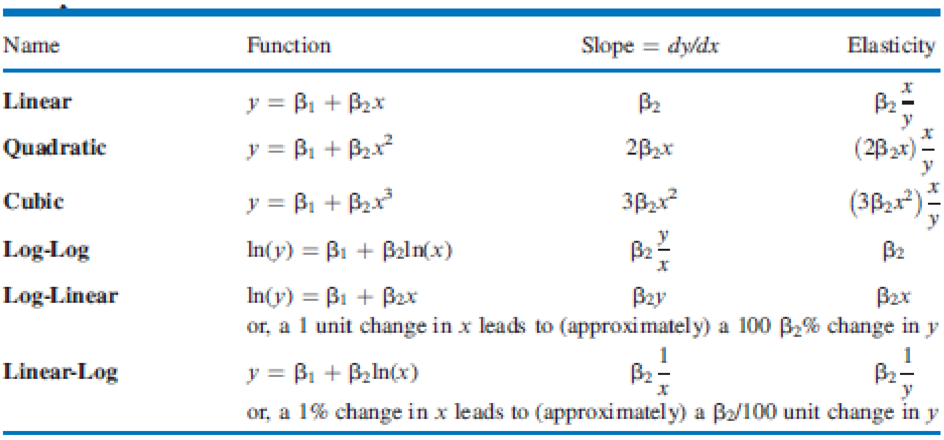


Figure 2.3. Alternative “linear” regression curves

Note that, say, log-log model $\log Y = \beta_0 + \beta_1 \log X$ is linear in $(logX, \log Y)$ coordinate system (and, therefore, we can apply OLS) whereas in (X, Y) coordinate system it represents a nonlinear relationship (see Fig. 2.3). This model is a constant elasticity[12] model while the linear one is the constant slope model.



Name	Function	Slope $= dy/dx$	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-Linear	$\ln(y) = \beta_1 + \beta_2 x$ or, a 1 unit change in x leads to (approximately) a 100 β_2% change in y	$\beta_2 y$	$\beta_2 x$
Linear-Log	$y = \beta_1 + \beta_2 \ln(x)$ or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$

[12] Recall that the elasticity of Y with respect to X is defined as $dY/dX \cdot X/Y$ (percentage response in Y to a 1% change in X). Thus, whatever is X, its 1% increase now leads to always the same $\beta_2\%$ increase in Y.

2.1 example. (i) Suppose that the estimated constant elasticity demand curve is given by $\log Q = \log 200 - 0.5 \log P$ or, what is the same, by $\hat{Q} = 200P^{-0.5}$. What is the price elasticity of demand? *Answer*. It equals -0.5 everywhere along the demand curve (draw the curve). (ii) Suppose an estimated linear demand curve is given by the formula $Q = 400 - 10P$. What is the price elasticity of demand at $P = 30$? At $P = 10$? *Answer*. $E\hat{Q}(P) = (-10) \cdot P / (400 - 10P) |_{P=30} = -3$ (|-3|>1, thus demand is elastic) but at $P = 10$ it is … Plot elasticity between 0 and 30.

2.2 example. Consider an annual time series Y_t evolving over time so that it grows annually at *rate* g: $Y_t = (1+g)Y_{t-1}$ (this might roughly describe the growth of a country's population, GDP, or price level). The definition implies that $Y_t = (1+g)^t Y_0 = Y_0 e^{\log(1+g) \cdot t}$ and this is called a constant growth model. Note that this model is equivalent to the log-linear model $\log(Y_t) =$ $\log(Y_0) + \log(1+g)t$ ($\approx \log(Y_0) + gt$ when g is „small"). In practical situations, we add a disturbance term ε_t and consider a regression model $\log(Y_t) = \log(Y_0) + gt + \varepsilon_t$ with the objective of estimating the growth rate g.

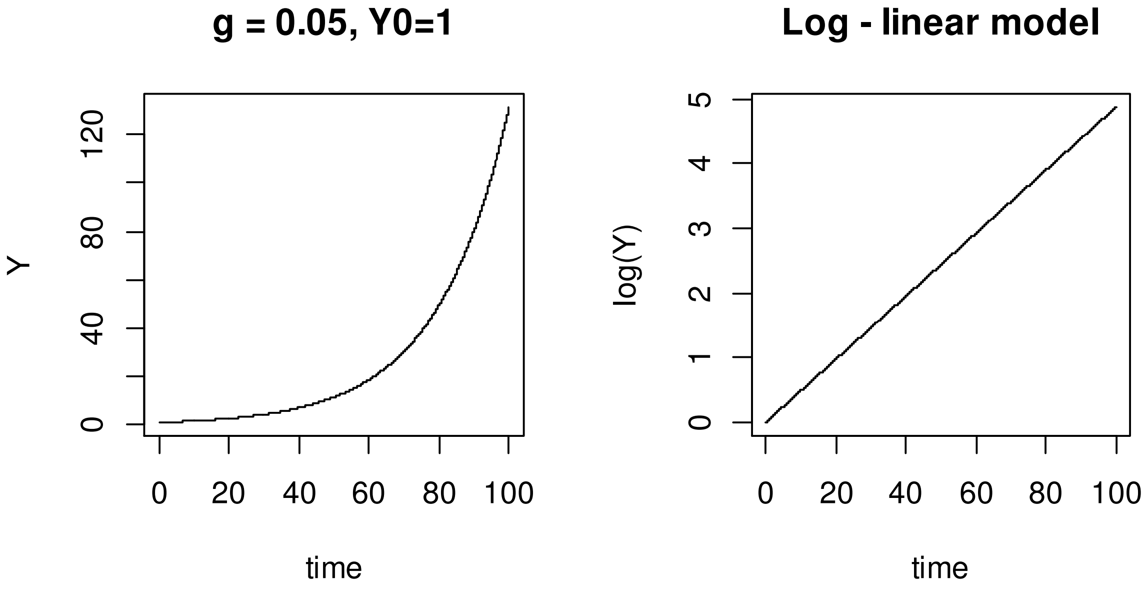


Figure 2.4. Constant growth model

2.5 exercise. With GRETL The data file jones.txt contains yearly U.S. GDP, 1880-1987, named GDP. Using different models, estimate the constant growth rate g.

a) Plot three series, $GDP_t, \log(GDP)_t$, and $\Delta \log(GDP)_t = \log(GDP)_t - \log(GDP)_{t-1}$.

b) Since $\log(Y_t) = \log(Y_0) + gt$, estimate β_1 in $\log GDP_t = \beta_0 + \beta_1 t + \varepsilon_t$ (recall the meaning of β_1: it is the percentage change of GDP in one year)

c) Since $\Delta \log GDP_t = g$, estimate β_0 in the model $\Delta \log GDP_t = \beta_0 + \varepsilon_t$ (recall the meaning of β_0: it is the percentage change of GDP in one year).

d) Since $GDP_t = Y_0 e^{\log(1+g)\cdot t}$, estimate the nonlinear model $GDP_t = \beta_0 e^{\beta_1 t} + \varepsilon_t$ and set $\hat{g} = e^{\hat{\beta}_1} - 1$.

e) Compare all three estimates of g.

f) Plot and compare the residuals of the three models.

g) In 1929, the U.S. witnessed the Great depression. Ask the 1929 econometrist to forecast GDP till 1987 and compare the forecast with the real data. Use the log-linear[13] model. ◀◀

2.6 exercise. With R GDP per capita (GDP/pop) is often considered an indicator of a country's standard of living (for example, China is now (2011 estimate) 2nd in the world by its GDP, but only 91st by its GDP per capita). Create one of the model of the previous exercise for GDP per capita. ◀◀

XIX a. vokiečių statistikas Ernst'as Engel'is nustatė, kad "kuo skurdesnė (belgų darbininkų) šeima, tuo didesnė jos išlaidų dalis skiriama maistui". Aišku, kad namų ūkio išlaidos kokiai nors prekei priklauso ne tik nuo šeimos pajamų, bet ir nuo kainos (ne tik pačios prekės, bet ir jos pakaitalo – taigi tai daugelio kintamųjų regresijos uždavinys). Antra vertus, jei duomenys surinkti vienu metu, kainos visiems namų ūkiams bus tos pačios ir galėsime nagrinėti, pvz., tokį dviejų kintamųjų modelį: $q_i = \beta_0 + \beta_1 \log(m_i) + \varepsilon_i$ (logaritmą renkamės todėl, kad jis paprastai atitinka ne tik realius duomenis, bet ir Engel'io dėsnį – pajamoms (arba bendroms namų ūkio išlaidoms) m didėjant, tiriamos prekės vartojimas q auga vis lėčiau).

1955 m. Prais'as ir Houthakker'as tyrė britų namų ūkių duomenis ir sudarė tokį darbininkų šeimų mėsos paklausos modelį: $\hat{q}_i^{(1)} = -40.8 + 16.3\log(m_i)$ (taigi pajamoms padidėjus 1%, išlaidos mėsai padidėja 16.3/100 piniginiais vienetais). Deja, mes neturime pradinių duomenų, todėl modeliuosime modelio kreivę - sukurkime cross-sectional duomenų rinkinį su 150 stebinių:

```
nulldata 150
series m = index
series q1 = -40.8 + 16.3*log(m)
gnuplot q1 m --output=display
```

Šio modelio elastingumas priklauso nuo m ir lygus $\beta_1 / q1$. Jį apskaičiuosime trijuose taškuose: m = 40, 62.2 ir 100 (tai maždaug 1-asis, 2-asis ir 3-iasis kvartiliai (kodėl?) – prisiminkite kvartilių apibrėžimus). Komandiniame (script) lange surinkę

```
scalar q11 = 16.3/(-40.8+16.3*log(40))
scalar q12 = 16.3/(-40.8+16.3*log(62.2))
scalar q13 = 16.3/(-40.8+16.3*log(100))
```

pamatysite atsakymus 0.8433, 0.6145 ir 0.4757, taigi sutinkamai su Engel'io dėsniu mėsa yra pirmo būtinumo[14] prekė (beje, augant pajamoms elastingumas mažėja[15]).

Tyrimo autoriai pateikia dar vieną modelį, būtent $\hat{q}_i^{(2)} = 41.0 - 801\cdot(1/m_i)$.

[13] Once you found $\widehat{\log GDP_t} = \hat{\beta}_0 + \hat{\beta}_1 t$, to forecast GDP_t use the formula $\widehat{GDP_t} = \exp\left(\widehat{\log GDP_t} + \hat{\sigma}_\varepsilon^2 / 2\right)$.

[14] Nes visi elastingumai yra moduliu mažesni už 1.

[15] Originaliame darbe nurodoma, kad pastovaus elastingumo (t.y. log-log) modelyje elastingumas lygus 0.69.

2.7 exercise. Apskaičiuokite šio (atvirkštinio) modelio elastingumus tuose pačiuose trijuose taškuose.

2.8 exercise. Duomenų rinkinyje houthak1.txt yra dvi sekos: `consump` ($=q$) (namų ūkyje sunaudotos elektros energijos kiekis, kWh) ir `income` ($=m$) (namų ūkio pajamos per metus, GBP). Sudarykite tris modelius :

log-log	$\log(q_i) = \beta_0 + \beta_1 \log(m_i) + \varepsilon_i$
lin-log	$q_i = \beta_0 + \beta_1 \log(m_i) + \varepsilon_i$
atvirkštinį	$q_i = \beta_0 + \beta_1 \cdot (1/m_i) + \varepsilon_i$.

Išbrėžkite duomenų sklaidos diagramą ir visas tris regresijos kreives. Kuris iš modelių geriausias? Apskaičiuokite visų modelių elastingumus taškuose, artimuose trims m kvartiliams. *Nuoroda*. Kvartilius apskaičiuoti galima su tokiu GRETLo skriptu :

```
scalar q1 = round(quantile(income,.25))
scalar q2 = round(quantile(income,.50))
scalar q3 = round(quantile(income,.75))
```

2.9 exercise. In GRETL, import the data set data8-2 from File| Open data| Sample file...| Ramanathan – it contains

`exptrav`	travel expenditures in 51 U.S. state in billions of dollars
`income`	personal income in billion of dollars
`pop`	population in millions

Create four models

mod1	$exptrav = \beta_0 + \beta_1\, income + \varepsilon$
mod2	$exptrav = \beta_0 + \beta_1 \log(income) + \varepsilon$
mod3	$\log(exptrav) = \beta_0 + \beta_1\, income + \varepsilon$
mod4	$\log(exptrav) = \beta_0 + \beta_1 \log(income) + \varepsilon$

Draw an *income-exptravel* scatter diagram and append it with the four fitted regression curves. Calculate all four coefficients of determination and compare them. Are the residuals of the best model normal? Repeat the same calculation with R. *Hint*. In 3 and 4 cases, use the formula $R^2_{(3)} = corr\left(exptravel, \exp(yhat3 + sigma3\,^\wedge 2/2)\right)^2$.

2.3. Testing the hypothesis $H_0 : \beta_m = \beta_m^0$

The standard OLS estimation output reports a `t-ratio` for testing the null hypothesis that the true regression coefficient is zero: $H_0 : \beta_m = 0$ (see, for example, p.2-2):

```
Dependent variable: wage

                 coefficient   std. error   t-ratio    p-value
  -----------------------------------------------------------
  const           8,38474      0,389135     21,55     1,83e-074 ***
  experience      0,0362978    0,0179370     2,024    0,0435    **
```

where `t-ratio` $= t_{N-(k+1)} = (\hat{\beta}_m - 0)/\sqrt{\widehat{\mathrm{var}}\,\hat{\beta}_m}$. However, if we want to test $H_0 : \beta_m = \beta_m^0$ (here β_m^0 is a number of interest), the statistics must be redefined: $t_{N-(k+1)} =$ $(\hat{\beta}_m - \beta_m^0)/\sqrt{\widehat{\mathrm{var}}\,\hat{\beta}_m}$. Again, following the rule of thumb, if $|\, t_{N-(k+1)} \,| > 2$, we reject H_0.

These „nonstandard" tests are often applied in the theory of elasticity. The essential idea is that elasticity measures *how* sensitive is consumption to prices. If prices matter very little, changes in price only will have small impacts on our willingness to buy or sell. For example, the price elasticity of demand (in $q^d = \beta_0 + \beta_1 p$) is computed as the percentage change in the quantity demanded divided by the percentage change in price:

$$Elast(q^d)(p) = \frac{dq^d}{dp} \cdot \frac{p}{q^d} \approx \frac{\left(q^d(p+\Delta p) - q^d(p)\right)/\, q^d(p)}{\Delta p \,/\, p} = \frac{\Delta q^d \,/\, q^d}{\Delta p \,/\, p}$$

Generally, the elasticity depends (just as a derivative) on p but some curves are constant elasticity curves (specifically, whatever is x, the elasticity of y with respect to x in $\log y = \beta_0 + \beta_1 \log x \Leftrightarrow y = \mathrm{e}^{\beta_0} x^{\beta_1}$ is always the same, namely, β_1). If the elasticity numbers exceed one, we say that demand and/or supply is *elastic*. If the numbers are less than one, we say that demand or supply is *inelastic*. If elasticity equals one, we say that demand or supply is *unit elastic*. Note that price elasticity of demand is always negative (whereas the price elasticity of supply is always positive) and in fact the above definitions apply to the moduli of elasticity.

2.3 example. We shall consider four annual time series, 1923-1939, in File| Open data| Sample file...| Gretl| theil (or in Theil.txt):

`year`	year
`consume`	volume of textile consumption per capita (base 1925=100)
`income`	real Income per capita (base 1925=100)
`relprice`	relative price of textiles (base 1925=100)

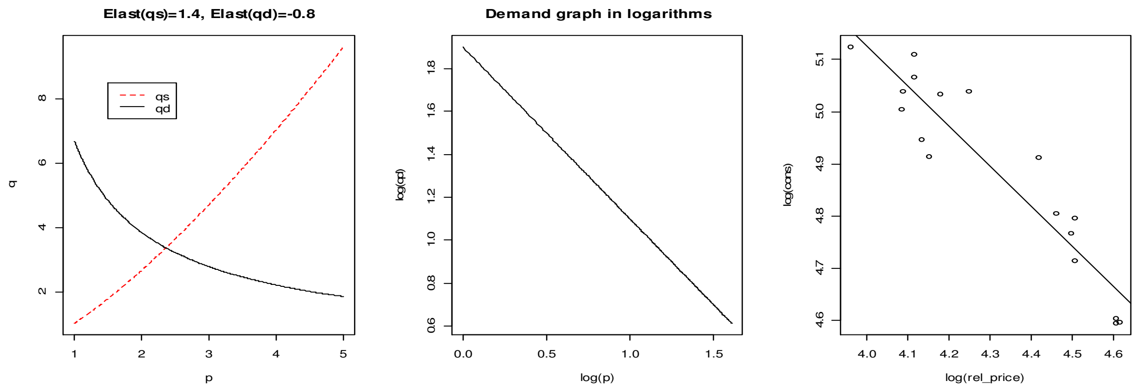


Figure 2.5. Constant elasticity demand and supply curves (left), constant demand line in logarithms (center), and the `log(relprice)` and `log(consume)` scatter diagram (right)

The output of the model[16] $\log(consume_t) = \beta_0 + \beta_1 \log(income_t) + \beta_2 \log(relprice_t) + \varepsilon_t$ is as follows:

```
mod=lm(log(consume)~log(income)+log(relprice))
summary(mod)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.16355    0.70480   4.489  0.00051 ***
log(income)      1.14316    0.15600   7.328 3.74e-06 ***
log(relprice)   -0.82884    0.03611 -22.952 1.65e-12 ***
```

The estimates of both coefficients are close (in modulus) to 1, but are the coefficients in DGP equal to 1? We shall test the hypotheses $H_0 : \beta_1 = 1$ and $H_0 : \beta_2 = -1$ in R first. One can test $H_0 : \beta_1 = 1$ in two different ways.

1) The $t-$statistics $t_{17-3} = (1.14316 - 1)/0.15600 = 0.9176923$ which is considerably less than 2, therefore there is no ground to reject $H_0 : \beta_1 = 1$ (more specifically, the $p-$value of this test with the one-sided alternative $H_1 : \beta_1 > 1$ equals `1-pt(0.9176923, 17-3)` (=0.1871597 >0.05)).

2) Another possibility to test $H_0 : \beta_1 = 1$ with the two-sided alternative $H_0 : \beta_1 \neq 1$ is to use the F test (see LN, Sec. 4.6) – the unrestricted model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ while the restricted $Y = \beta_0 + 1 \cdot X_1 + \beta_2 X_2 + \varepsilon$, the F statistics equals $f = \dfrac{(RSS_R - RSS_{UR})/1}{RSS_{UR}/(17-(2+1))}$.

```
library(car)
linearHypothesis(mod,"log(income)=1")

Hypothesis:
log(income) = 1
Model 1: restricted model
```

[16] This is a two-variate model but the syntax is essentially the same.

```
Model 2: log(consume) ~ log(income) + log(relprice)

  Res.Df       RSS Df  Sum of Sq      F Pr(>F)
1     15 0.014432
2     14 0.013613  1 0.00081884 0.8421 0.3743
```

Note that both tests give the same answer (since $t^2 = f$, 0.91769^2 = 0.8421) and also the one-sided p – value in the t – test (=0.187...) should be multiplied by 2 to get the p – value in the two-sided F – test (=0.3743)

2.10 exercise. Test the hypothesis $H_0 : \beta_2 = -1$. ◀◀

To repeat the analysis with GRETL, create the same log-log model as above and in the model window go to Tests| Linear restrictions, type in `b[2]=1`| OK. The output (of the F – test) coincides with that of R:

```
Restriction:
b[l_income] = 1
Test statistic: F(1, 14) = 0.842112, with p-value = 0.374332
```

2.11 exercise. Žinoma Cobb'o ir Douglas'o (netiesinė) gamybos funkcija suriša produkcijos apimtį (Y) su gamybos veiksniais, pvz., darbu (L) ir kapitalu (K):

$$Y = \alpha L^{\beta} K^{\gamma} \exp(\varepsilon)$$

(šis modelis nėra įprastinis regresinis modelis $Y = \alpha L^{\beta} K^{\gamma} + \varepsilon$, tačiau artimas jam). Išlogaritmavę šį (netiesinį!) reiškinį, gautume tokį tiesinį log-log modelį :

$$\log(Y) = \beta_0 + \beta_1 \log(L) + \beta_2 \log(K) + \varepsilon .$$

Tokio tipo reiškiniai ekonomikoje sutinkami labai dažnai. Pvz., lentelė TABLE9.10.txt turi 25 įrašus (eilutes) ir 5 (bevardžius) stulpelius, kuriuose pateikti Quantity Indexes of Capital (`K`), Labor (`L`), Energy (`E`), Other Intermediate Products (`M`), Gross Output (`Y`) for U.S. Manufacturing. Aišku, kad tai daugelio kintamųjų regresijos uždavinys, mes kol kas sudarysime paprastesnį vieninį modelį. Importuokite nurodytą lentelę ir sudarykite keturis modelius

$$\begin{gathered}
\log Y = \beta_0 + \beta_1 \log K + \varepsilon \\
\log Y = \beta_0 + \beta_1 \log L + \varepsilon \\
\log Y = \beta_0 + \beta_1 \log E + \varepsilon \\
\log Y = \beta_0 + \beta_1 \log M + \varepsilon
\end{gathered}$$

Kuris modelis yra tiksliausias (atsako kintamojo $\log Y$ aprašymo tikslumo prasme)? Kuriame iš šių modelių prognozinio kintamojo elastingumas yra didžiausias? Koks šiame modelyje hipotezės $H_0 : \beta_1 = 1$ likimas? Ką galite pasakyti apie `L` elastingumą modelyje $\log Y = \beta_0 + \beta_1 \log K + \beta_2 \log L + \beta_3 \log E + \beta_4 \log M + \varepsilon$? ◀◀

2.4 example. In LN, p. 3-26, we have examined the data set containing the percent of the total US population that lives on farms (in GRETL go to File| Open data| Sample file...| Ramanathan| data6-6. We shall create a few more models of the time series (`fp=`) `farmpop` now. The graph of the variable resembles descending exponent or power function (see Fig.2.6, blue line), therefore we shall analyse two models (first add logarithms to your list of variables):

$$\text{Mod.1:}\quad \log fp_t = \beta_0 + \beta_1 year + \varepsilon_t \qquad \Rightarrow\quad \widehat{fp}_t = \exp(\hat{\beta}_0 + \hat{\beta}_1 year) \qquad \text{(exponential)}$$

$$\text{Mod.2:}\quad \log fp_t = \beta_0 + \beta_1 \log year + \varepsilon_t \ \Rightarrow\ \widehat{fp}_t = \exp(\hat{\beta}_0)\, year^{\hat{\beta}_1} \qquad \text{(power)}$$

```
Model 1: OLS, using observations 1948-1991 (T = 44)
Dependent variable: l_fp
             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       107.158       1.61218        66.47    3.48e-044 ***
  year         -0.0535770   0.000818557   -65.45    6.61e-044 ***

R-squared            0.990291   Adjusted R-squared   0.990060
Log-likelihood       56.26349   Akaike criterion    -108.5270
Schwarz criterion  -104.9586   Hannan-Quinn        -107.2037
rho                  0.753045   Durbin-Watson        0.553432
```

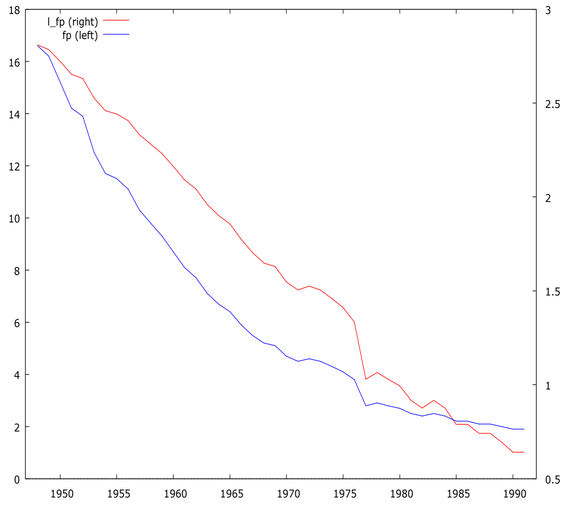


Figure 2.6. The graphs of `fp` (blue, left axis) and `l_fp` (red, right axis)

The Durbin-Watson test informs that the errors are autocorrelated, therefore we can use the HAC standard errors (in the model window check the Robust standard errors box):

```
Dependent variable: l_fp
HAC standard errors, bandwidth 2 (Bartlett kernel)

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       107.158       2.46916       43.40    1.60e-036 ***
  year         -0.0535770   0.00125881   -42.56    3.57e-036 ***
```

We got different `std.errors` now, but they have no significant effect on $p-$values.

The graph of the forecast depends on the scaling of y axis: it will be a straight line in `year` – `l_fp` coordinates and exponent in `year` – `fp` coordinates. Note that we must use the following correction for the latter[17] forecast: $\widehat{fp}_t = \exp(\hat{\beta}_0 + \hat{\beta}_1 year + s_\varepsilon^2/2)$. The gretl script is below:

```
ols l_fp 0 year
scalar ss = $sigma
series yhat = $yhat
series uhat = $uhat
series yhat_cor = exp(yhat+ss^2/2)
gnuplot fp yhat_cor uhat --output=display --time-series --with-lines
```

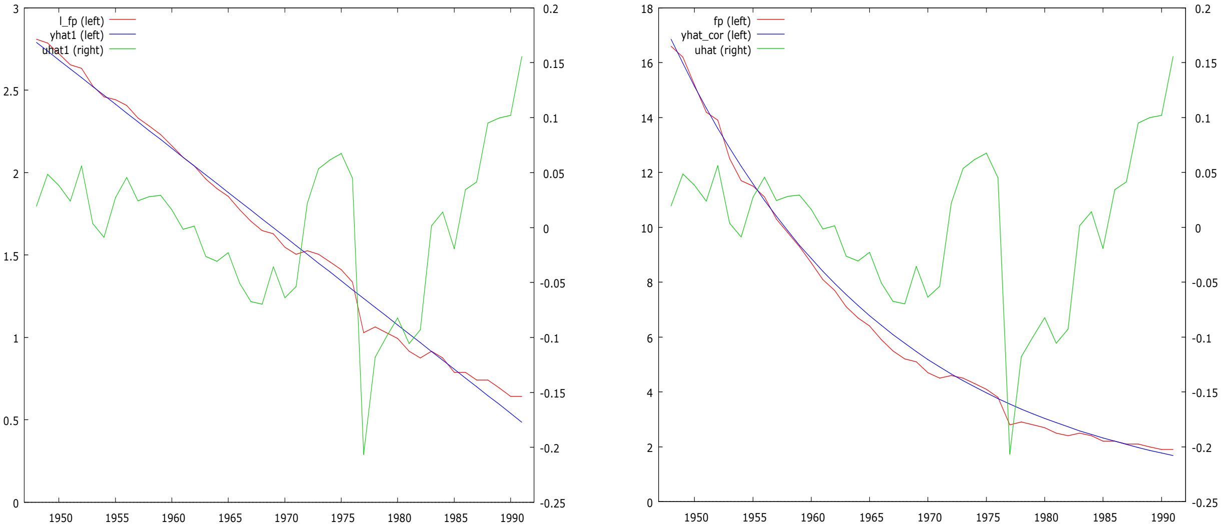


Figure 2.7. Forecast for `l_fp` (left) and corrected forecast for `lp` (right) (note the *Y* axes scales)

a) Draw the same (as in Fig.2.7) graphs from the pull-down menus.
b) Estimate the derivative and elasticity of `fp` in the year 1980.
c) Repeat the same analysis with the power model.

A few comments on the power model

```
Dependent variable: l_farmpop
HAC standard errors, bandwidth 2 (Bartlett kernel)

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        802.161      18.2434      43.97    9.37e-037 ***
  l_year      -105.533       2.40637    -43.86    1.04e-036 ***
```

[17] The correction is used when the model is for $\log(Y)$, but we want to describe *Y*.

Recal that the meaning of the cofficient β_1 in this model is elasticity: 1% change in X leads to an (approximately) β_1% change in Y. This claim is true when β_1 is not very big (roughly, <5), but now (when $\hat{\beta}_1 = -105.5$) it is senseless. The true effect of `year` on `farmpop` equals

$$\frac{farmpop_{new} - farmpop_{old}}{farmpop_{old}} = \frac{\exp(\hat{\beta}_0)(1.01 \cdot year)^{\hat{\beta}_1} - \exp(\hat{\beta}_0)\, year^{\hat{\beta}_1}}{\exp(\hat{\beta}_0)\, year^{\hat{\beta}_1}} = 1.01^{-105.533} - 1 = -0.65$$

which means that 1% increase in `year` leads to 65% decrease in the percentage of the farm population (for example, in in the time period between 1970 and 1989.7 (the 20 years or 1% increase) the model predicts the decrease from 4.7 to 1.645 what is close to the true value of 1.9).

d) Forecast `fp` for the comming 10 years with the power model. ◀◀

2.12 exercise. The data set newbroiler.txt contains 52 annual observations, 1950-2001:

`q`	per capita consumption of boneless chicken, pounds
`y`	per capita real disposable income, 1996 = 100
`p`	real price (index) of fresh chicken
`pb`	real price (index) of beef
`pcorn`	real price (index) of feed corn
`pf`	real price (index) of broiler feed
`qprod`	estimate of aggregate production of boneless chicken
`lexpts`	log of estimate of exports of boneless chicken
`popgro`	population growth rate

1. Draw a (p, q) and $(\log p, \log q)$ scatter diagrams
2. Create two demand models: $q = \beta_0 + \beta_1 p + \varepsilon$ and $\log q = \beta_0 + \beta_1 \log p + \varepsilon$
3. What is the elasticity of both models at the `price` median point?
4. Plot their residuals. Test for normality.
5. Compare the coefficients of determination of the two models (for the second model estimate the coefficient by the formula $R^2 = (cor(q, \hat{q}))^2$ where $\hat{q} = \exp(\widehat{\log q}) e^{\hat{\sigma}^2/2}$).
6. Which model do you prefer?
7. Using the 52 annual observations, 1950–2001, estimate the reciprocal model $q = \beta_0 + \beta_1 (1/p) + \varepsilon$. Plot the fitted value of q versus p. How well does the estimated relation fit the data?
8. Using the estimated relation in part (7), compute the elasticity of q with respect to p when the real price is its median, \$1.31, and quantity q is taken to be the corresponding value on the fitted curve. Compare this estimated elasticity to the estimate found in part (3) with the log-log functional form.
9. Estimate the poultry demand using the linear-log functional form $q = \beta_0 + \beta_1 \log p + \varepsilon$. Plot the fitted values of q versus p. How well does the estimated relation fit the data?

10. Using the estimated relation in part (9), compute the elasticity of q with respect to p when the real price is its median, \$1.31. Compare this estimated elasticity to the estimate from the log-log model and from the reciprocal model in part (7).
11. Evaluate the suitability of the log-log, linear-log, and reciprocal models for fitting the poultry consumption data. Which of them would you select as best, and why?

2.13 exercise. (Monte Carlo exercises). Assume that $X, X_1, X_2, ...$ are independent identically distributed random variables (iidrv's). The Law of Large Numbers (LLN) claims that $(X_1 + ... + X_N)/N \to EX$ and, as a consequence, $(1_{X_1 > a} + ... + 1_{X_N > a})/N \to E1_{X>a} = P(X > a)$. In the case where $P(X > a)$ is difficult to estimate, one can generate "many" copies of X and to count how many times $X_i > a$ – the average number of "successes" will be close to $P(X > a)$ (this is called a Monte Carlo method (MCM) or MC simulations).

Estimation of π

We begin with explaining how one can use MCM to estimate the number $\pi (= 3.14159...)$. Generate a sequence of two dimensional random vectors $\vec{\alpha}_1, \vec{\alpha}_2, ...$ having a uniform distribution in the square S with vertices at (-1,-1), (1,-1), (1,1), and (-1,1) (one copy of such a vector is generated with `runif(2,-1,1)`). Recall that in the case of uniform distribution, the probality $P(\vec{\alpha} \in A) = L(A)/L(S) = L(A)/4$ where $L(A)$ is the Lebesque measure of A, i.e., just its area. This implies that $P(\vec{\alpha} \in C) = \pi/4$ (here C is a unit circle in plane). If the number of MC experiments is „big" (=10000, 100000 etc), the relative frequency of experiments which led to $\alpha_i \in C$ is approximately equal to $\pi/4$. The following code in R allows to approximately estimate π:

```
xx <- c(-1,1,1,-1,-1)
yy <- c(-1,-1,1,1,-1)
plot(xx,yy,type="l", main="500 points")  # Draw a square S
x <- cos(seq(0,2*pi,length=100))
y <- sin(seq(0,2*pi,length=100))  # Circumference in polar coordinates
polygon(x,y,col=3)                # Colour the circle
points(runif(500,-1,1),runif(500,-1,1),pch="*") # see Fig. 2.8, left
##############
set.seed(10)
xxx <- runif(100000,-1,1) # We perform 100000 MC experiments, that is
yyy <- runif(100000,-1,1) # throw 100000 points into S
s <- numeric(500)
print(s)                  # Vector of 500 zeros
# Now loop: estimate relative frequency using
# the batches of 200*i points
for(i in 1:500) {s[i] <-
4*sum(ifelse(xxx[1:(200*i)]^2+yyy[1:(200*i)]^2<=1,1,0))/(200*i)}
print(s)             # A vector of relative frequances times 4 (≈π)
plot(1:500,s,type="l")
abline(pi,0)
s[500]               # The final estimate
[1] 3.14252
```

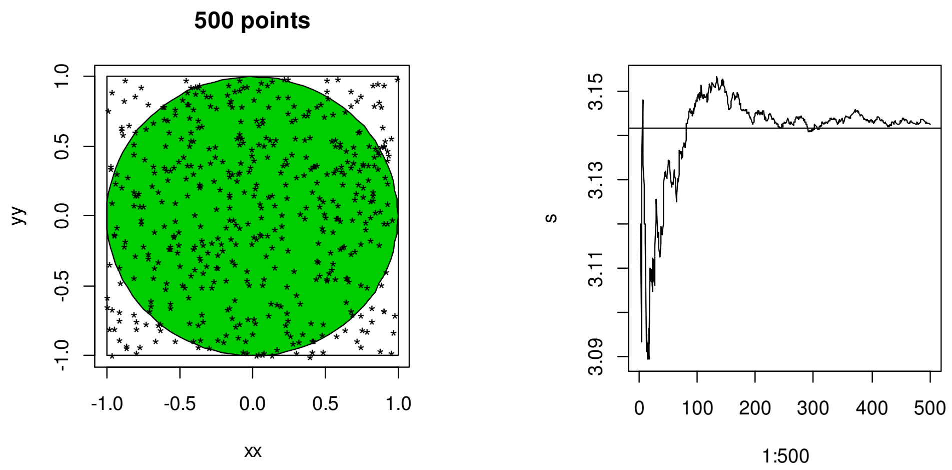


Figure 2.8. The first 500 points (left); relative frequency approaches π (right)

We repeated the experiment of throwing the point 100000 times, but the accuracy is still not very high. This is common for the MCM, but the method makes sense when the direct calculation of the probability in question is complicated. First 2000 digits of π can be found in `library(UsingR);pi2000`.

Testing $H_0 : \beta_1 = \beta_1^0$

Assume that a researcher has a sample $(Y_1, X_1),...,(Y_N, X_N)$, she expects that respective DGP is $Y = \beta_0 + \beta_1 X + \varepsilon$ and also that U2 and U3 holds. She uses OLS and calculates $\hat{\beta}_1^{OLS} = 0.15$. Is it compatible with an inside information that $\beta_1 = 0.5$? She knows that too large a deviation of $\hat{\beta}_1^{OLS}$ from 0.5 is a sign of the failure of $H_0 : \beta_1 = 0.5$; how large is "too large"? To make it clear, we shall pick a particular model by taking $\beta_0 = 1$, $\beta_1 = 0.5$, $\sigma_\varepsilon^2 = 1$ and try to answer the questions whether 1) $\hat{\beta}_1^{OLS}$ is a "good" estimator of β_1 (for example, is it unbiased, i.e., $E(\hat{\beta}_1^{OLS} \mid \vec{X}) = 0.5$) and 2) what deviations $\mid \hat{\beta}_1^{OLS} - 0.5 \mid$ do not contradict H_0?

To answer the 1st question, we shall simulate our sample "many times" and calculate $(\hat{\beta}_1 + ... + \hat{\beta}_{NN}) / NN$ - if it is „very close" to 0.5, then the OLS procedure is generally acceptable. We start with generating a conditioning random vector $\vec{X} = (X_1,..., X_N)$. This can be done in many diverse ways and we begin with a rather complicated process[18] $X_i = c + \phi X_{i-1} + u_i$, $i = 1,..., N$, where $\{u_i\}$ is iid $N(0,1)$ and $X_0 \sim N\left(\frac{c}{1-\phi}, \sqrt{\frac{\sigma_\varepsilon^2}{1-\phi^2}}^{\,2}\right)$, $c = 3$, $\phi = 0.7$. This fixes the joint distribution of (Y, X) from which a large number of sam-

[18] This a stationary AR(1) process (that is, $EX_t \equiv c/(1-\phi)$, var $X_t \equiv 1/(1-\phi^2)$; see PE.II, Lecture Notes, 2.8).

ples will be drawn. To code the simulation, note that

$$X_i = c + \phi(c + \phi X_{i-2} + u_{i-1}) + u_i = ... = \phi^i X_0 + (1 + \phi + ... + \phi^{i-1})c + (u_i + \phi u_{i-1} + ... + \phi^{i-1} u_1).$$

The simplest and fastest way to simulate many samples is to use the matrix algebra:

$$\begin{array}{ccccccc} \vec{X} & = & \vec{r} \cdot X_0 & + & \vec{d} & + & \mathbf{A} \quad \vec{u} \\ (N \times 1) & & (N \times 1) & & (N \times 1) & & (N \times N)(N \times 1) \end{array}$$

where $\vec{r} = \begin{pmatrix} \phi \\ \phi^2 \\ ... \\ \phi^N \end{pmatrix}, \vec{d} = \begin{pmatrix} c \\ (1+\phi)c \\ .... \\ (1+\phi+...+\phi^{N-1})c \end{pmatrix}, \mathbf{A} = \begin{pmatrix} 1 & 0\ 0...\ 0 \\ \phi & 1\ 0...\ 0 \\ \phi^2 & \phi\ 1\ ...\ 0 \\ .... & \\ \phi^{N-1} & \phi^{N-2} ....1 \end{pmatrix}, \vec{u} = \begin{pmatrix} u_1 \\ u_2 \\ ... \\ u_N \end{pmatrix}$.

```
N=16          # sample size
beta0=1
beta1=0.5
sigma=1
cc=3
phi=0.7
set.seed(1)
####
X=numeric(N)
X0=rnorm(1,mean=cc/(1-phi),sd=1/sqrt(1-phi*phi)) # starting value of X
rr=phi^(1:N)
dd=cc*(1-phi^(2:(N+1)))/(1-phi)
AA=matrix(0,N,N)
col1=phi^(0:(N-1))
for(i in 1:N) AA[i:N,i]=col1[1:(N-(i-1))]          # create matrix A
X=rr*X0 + dd + AA%*%rnorm(N)                        # create a sample of X
plot(X,type="l");abline(cc/(1-phi),0,col=2)
```

In the **first** variant of simulation, we shall keep $\vec{X} = (X_1,...,X_N)$ fixed and change only the errors $\vec{\varepsilon}$:

```
NN=1000
beta=numeric(NN)
for(i in 1:NN)
{
Y=beta0+beta1*X+rnorm(N) # the errors satisfy U2, U3
beta[i]=coef(lm(Y~X))[2]
}
mean(beta) # =0.502 -> OLS is unbiased
```

Thus, if H_0 is true, $\hat{\beta}_1$ fluctuates around 0.5. To answer the 2nd question, we have to estimate how improbable is the event $\hat{\beta}_1 \le 0.15$ under the given condition $\vec{X}$.

```
sum(ifelse(beta<=0.15,1,0))/NN  # relative frequency of β̂1 ≤0.15
[1] 0.089   # empirical p-value
```

In statistics, 0.05 is the standard probability of impossible or improbable event, therefore 0.089 means that such a deviation from 0.5 can be explained just by random fluctuations of ε's (but not because H_0 is false). Note that in LN, 3.7, we have already mentioned that $T = \left(\hat{\beta}_1 - 0.5\right) / \sqrt{\widehat{\text{var}}\hat{\beta}_1} = \dfrac{\hat{\beta}_1 - 0.5}{s.e.\hat{\beta}_1}$ has the Student T_{N-2} distribution. Since

```
 > pt((0.15-0.5)/sd(beta),N-2)
[1] 0.099 # one-sided theoretical p-value
```

our simulation "proves" the above claim.

In the **second** variant of simulation, we shall again keep $\vec{X} = (X_1,...,X_N)$ fixed, but now change the distribution of the errors $\vec{\varepsilon}$:

```
NN=1000
beta=numeric(NN)
set.seed(5)
for(i in 1:NN)
{
Y=beta0+beta1*X+runif(N,-sqrt(3),sqrt(3)) # uniform distribution with sd=1
beta[i]=coef(lm(Y~X))[2]
}
mean(beta) # 0.494
sum(ifelse(beta<=0.15,1,0))/NN # empirical rel. frequency = 0.103
pt((0.15-0.5)/sd(beta),N-2)    # probability = 0.102
```

This proves that the distribution of T is not very sensitive to the distribution of ε.

So far, $\vec{X}$ was fixed throughout the loop for Y. The **third** variant of simulation calculates the unconditional distribution of the $t-$ratio (in each replication, we generate a new $\vec{X}$).

```
X=numeric(N)
NN=10000
beta=numeric(NN)
set.seed(2)
for(i in 1:NN)
{
X0=rnorm(1,mean=cc/(1-phi),sd=1/sqrt(1-phi*phi))
X=rr*X0 + dd + AA%*%rnorm(N)
Y=beta0+beta1*X+rnorm(N)
beta[i]=coef(lm(Y~X))[2]
}
mean(beta)                     # 0.497
sum(ifelse(beta<=0.15,1,0))/NN # empirical rel. frequency = 0.04
pt((0.15-0.5)/sd(beta),N-2)    # probability =0.05
```

Again, MC modeling gives results close to theorethical. Thus, when necessary, we can replays theoretical considerations by empirical modeling. On the other hand, if you have a valid formula, there is no need for MC experimenting.

The **fourth** variant, instead of generating $\vec{X}$ with stationary AR(1), uses a simple procedure `X = rnorm(N)` or `X = runif(N,-sqrt(3),sqrt(3))`. Repeat the previous variants in this new setting. What about the two-sided alternatives? Also, investigate the conditional case where only one value of $\vec{X}$, say, the last one, is fixed. Use the Kolmogorov-Smirnov test (`ks.test`) to verify that $t-$ratio has the T_{N-2} distribution. ◀◀

2.4. Codes for the Ch.3 of the Lecture Notes

- **Figures 3.4 and 3.5**

To illustrate the properties of unbiasedness, we return to our DGP $Y=-2+3X+\varepsilon$. Clearly, the estimate $\hat{\beta}_1$ depends on sample. In Fig. 3.4 of LN, one can see four different (out of 5000 generated) samples and four different estimates of regression line (etc…)

```
# 4 regression lines
set.seed(11)
par(mfrow=c(2,2))
for(i in 1:4)
{
X=rnorm(20,sd=5)
Y=-2+3*X+rnorm(20,sd=6)
plot(X,Y)
abline(-2,3)
abline(lm(Y~X),col=2,lty=2)
}

# sample mean
set.seed(11)
beta1=numeric(5000)
for(i in 1:5000)
{
X=rnorm(20,sd=5)
Y=-2+3*X+rnorm(20,sd=6)
mod=lm(Y~X)
beta1[i]=mod$coef[2]
}
mean(beta1)

# sample variance
set.seed(11)
beta1=numeric(5000)
for(i in 1:5000)
{
X=rnorm(1000,sd=5)
Y=-2+3*X+rnorm(1000,sd=6)
mod=lm(Y~X)
beta1[i]=mod$coef[2]
}
var(beta1) #=0.00145

# 3 histograms

par(mfrow=c(1,3))
for(j in 1:3)
{
```

```
set.seed(11)
beta1=numeric(5000)
for(i in 1:5000)
{
X=rnorm(10^j,sd=5)
Y=-2+3*X+rnorm(10^j,sd=6)
mod=lm(Y~X)
beta1[i]=mod$coef[2]
}
hist(beta1)
}
```

- **Figure 3.17**

```
library(car)
data(USPop)
head(USPop)
#attach(USPop)
mod.exp=nls(population~a+b*10^(-12)*exp(c/1000*year),
start=list(a=0,b=10,c=15),data=USPop[1:15,])
# the  iterative  procedure in nls converges better if the parameters are
# of similar order
summary(mod.exp)
mod.logist=nls(population~a/(1+exp(b*(year-1916))),
start=list(a=100,b=0),data=USPop[1:15,])
# convergence is better when the mean of explanatory variable is close to 0
summary(mod.logist)
par(mfrow=c(1,2))
plot(year,population,type="l",main="Exponential",ylim=c(0,300))
points(year,predict(mod.exp,newdata=data.frame(year=seq(1790,2000,by=10))),
col=2)
plot(year,population,type="l",main="Logistic")
points(year,predict(mod.logist,newdata=data.frame(year=seq(1790,2000,
by=10))),col=2)
```

3. MULTIVARIATE REGRESSION MODELS

3.1. Simple model

3.1 example. The data set set andy.txt contains 75 observations in different cities of three variables:

`sales`	Monthly hamburger sales revenue ($1000s)
`price`	A price index for all products sold in a given month (in dollars)
`advert`	Expenditure on advertising ($1000s)

In a two-variables case, a scatter diagram is very informative about the relationship between Y and X. However, in multivariate case, there is no useful analogue to the diagram. Some information (in R) is provided by the command `plot(andy)` (see Fig.3.1, left) or still better by (see Fig.3.1, right)

```
pairs(andy,upper.panel=panel.smooth,diag.panel=panel.hist,
lower.panel=panel.cor)   # consult ?pairs
```

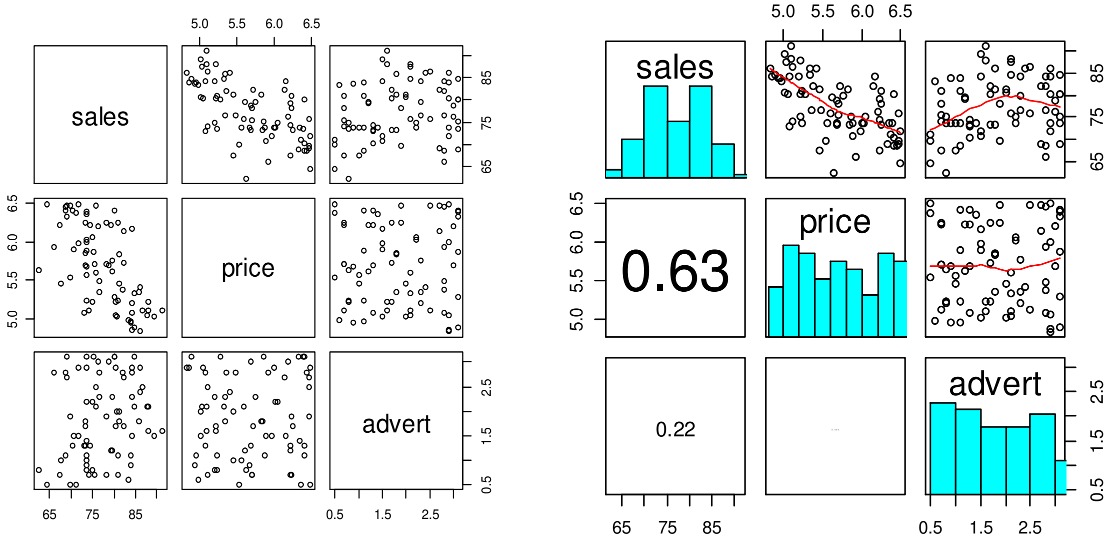


Figure 3.1. The rhs plot suggests linear dependence of `sales` on `price` and parabolic on `advert` (the function `panel.smooth` uses the `loess` procedure, see LN, p.2-6)

We shall analyze the model

$$sales = \beta_0 + \beta_1 advert + \beta_2 price + \varepsilon$$

where ε may include weather, the behavior of competitors, a new Surgeon General‘s report on the deadly effects of fat intake, and so on. We begin with a simple, univariate model. The graph suggests that sales do not individually depend on advert. Indeed,

```
Model 2: OLS, using observations 1-75
Dependent variable: sales

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        74.1797       1.79898      41.23     2.56e-052 ***
  advert        1.73262      0.890324      1.946    0.0555    *

Log-likelihood      -244.2731   Akaike criterion     492.5463
Schwarz criterion    497.1813   Hannan-Quinn         494.3970
```

However, if we create a full model,

```
Model 3: OLS, using observations 1-75
Dependent variable: sales

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       118.914        6.35164      18.72     2.21e-029 ***
  advert        1.86258      0.683195      2.726    0.0080    ***
  price        -7.90785      1.09599      -7.215    4.42e-010 ***

Log-likelihood      -223.8695   Akaike criterion     453.7390
Schwarz criterion    460.6915   Hannan-Quinn         456.5151

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 0.989504
  with p-value = 0.609722
```

`advert` becomes significant, Akaike‘s criterion smaller, and errors normal[1] (as a general note, if you omit a relevant variable in the model, namely price in the second model, the coefficients in the simplified model become biased; thus it is recommended to begin with the most general model). The estimate `1.86258` of the coefficient at `advert` means that if `advert` increases by 1 unit, `sales` increase by 1.86258 units (what is the meaning of -7.90785?).

Remark. A word of caution is in order about interpreting regression results. The negative sign attached to `price` implies that reducing the price will increase sales revenue. If taken literally, why should we not keep reducing the price to zero? Obviously that would not keep increasing total revenue. This makes the following important point: estimated regression models describe the relationship between the economic variables for values similar to those found in the sample data. Extrapolating the results to extreme values is generally not a good idea. Predicting the value of the dependent variable for values of the explanatory variables far from the sample values invites disaster. ◀◀

We can also begin with a still more general model[2]

$$sales = \beta_0 + \beta_1 advert + \beta_2 advert^2 + \beta_3 price + \beta_4 price^2 + \beta_5 advert * price + \varepsilon$$

[1] In the model window, go to Tests| Normality of residual.

[2] First you have to create three new variables through the Add menu.

(this is a model with the *interaction* term `advert*price`).

```
Model 4: OLS, using observations 1-75
Dependent variable: sales

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       238.297       83.8885       2.841    0.0059  ***
  price       -54.7116      29.3081      -1.867    0.0662  *
  advert       17.0995       7.58396      2.255    0.0273  **
  sq_advert    -3.07490      0.951730    -3.231    0.0019  ***
  sq_price      4.24943      2.55016      1.666    0.1002
  ad_pr        -0.696829     1.18378     -0.5886   0.5580

R-squared            0.530341   Adjusted R-squared   0.496308
Log-likelihood      -217.8292   Akaike criterion     447.6584
Schwarz criterion    461.5634   Hannan-Quinn         453.2105
```

Note that despite the fact that some coefficients of Model 4 are insignificant and the number of variables is greater, its Akaike criterion is smaller then that of Model 3 (thus Model 4 is preferable); on the other hand, the more strict Schwarz criterion suggests Model 3. We can remove three insignificant variables manually, variable-by-variable, but it is also possible to automate the procedure: in Model 4 window, go to Tests| Omit variables| and check the „Sequential elimination[3] of variables ...“ box:

```
Model 5: OLS, using observations 1-75
Dependent variable: sales

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        248.839      81.5712       3.051    0.0032  ***
  price        -57.0629     28.8988      -1.975    0.0523  *
  advert        13.1624      3.55823      3.699    0.0004  ***
  sq_advert     -3.08965     0.946949    -3.263    0.0017  ***
  sq_price       4.33638     2.53397      1.711    0.0915  *

R-squared            0.527983   Adjusted R-squared   0.501010
Log-likelihood      -218.0171   Akaike criterion     446.0341
Schwarz criterion    457.6216   Hannan-Quinn         450.6609
```

This model is better then Model 4 (in both Akaike and Schwarz sense), one can still improve it by omitting `sq_price` but it will change the model only marginally.

The graph on the right (see Fig.3.2) shows that there exists a critical ammount of expenditure on advertising where `sales` starts to diminish. To find the point, differentiate the model[4]

```
 ^sales = 249 - 57.1*price + 13.2*advert - 3.09*sq_advert + 4.34*sq_price
        (81.6) (28.9)       (3.56)        (0.947)          (2.53)

n = 75, R-squared = 0.528 (standard errors in parentheses)
```

[3] Recall that the automated sequential elimination procedure is risky. It does not apply to our case but always keep in your mind the rule: if `sq_x` is significant, never remove the linear term `x`.

[4] To get this expression of the model, in the model window go to File| View as equation.

in respect of `advert`: $\frac{\partial \widehat{sales}}{\partial advert} = 13.2 - 3.09 \cdot 2 \cdot advert = 0$; thus the optimal expenditure on advertising equals 2.14.

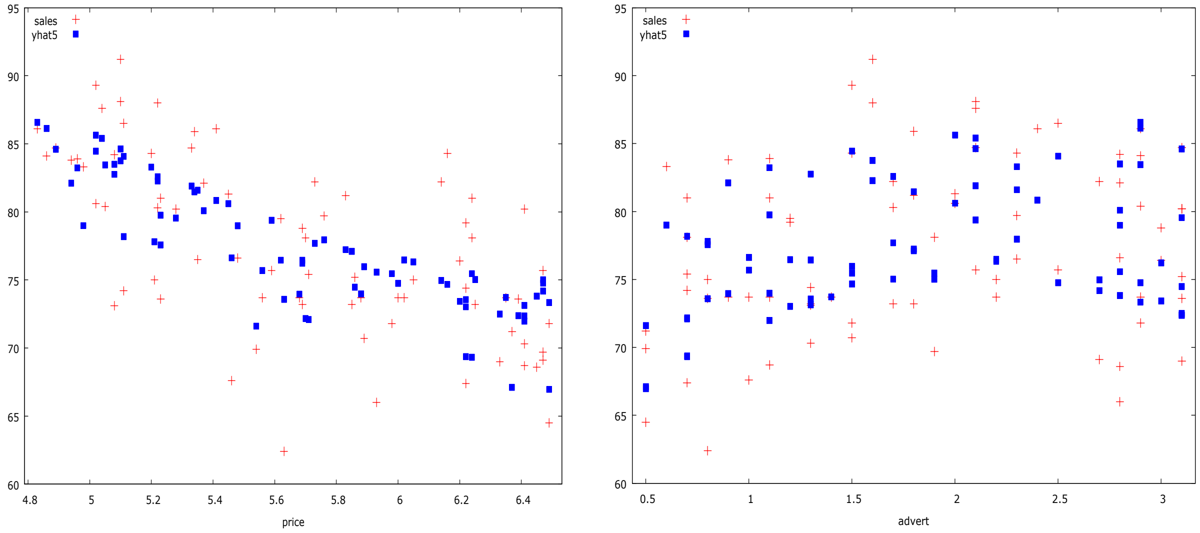


Figure 3.2. `yhat5` is for the fitted values in Model 5 (watch x axes!); the parabolic dependence of `yhat5` on, say, `advert` in the right graph of the panel is not very parabolic because `yhat5` depends also on `price`; note that the blue "parabolas“ differ (the left parabola is almost a straight line and the branches of the right one go downwards; why?)

Let us return to the simplified Model 3:

```
^sales = 119 - 7.91*price + 1.86*advert
        (6.35)(1.10)       (0.683)
```

One hypothesis of interest is whether an increase in advertising expenditure will bring an increase in sales revenue that is sufficient to cover the increased cost of advertising. Since such an increase will be achieved if $\beta_2 > 1$, we set up the hypotheses[5] $H_0 : \beta_2 \le 1$ and $H_1 : \beta_2 > 1$.

```
Restriction:
b[advert] = 1
Test statistic: F(1, 72) = 1.59409, with p-value = 0.210817
```

The p – value of this F –test can also be obtained by calculating the probability $P\left(|T_{75-3}| > \frac{1.8626-1}{0.6832}\right) = 1.263 = 0.2108$. Note that the same result can be achieved through

```
ols sales 0 price advert
restrict
    b[advert] = 1
end restrict
```

[5] To test them, in the model window go to Tests| Linear restrictions| `b[advert] = 1` or

In fact, we are interested in one-sided p – value: $P(T_{72} > 1.263) = 0.105\,(= 0.2108/2)$ which is >0.05, therefore we have no ground to reject H_0: despite the fact that $\hat{\beta}_2 = 1.86$, there is insufficient evidence in our sample to conclude that advertising will be cost effective. The same conclusion can be made using more adequate Model 5: the derivative 13.2 – 3.09*2*`advert` depends on `advert` and is often less than (rather than greater than) one.

Andy's marketing adviser claims that dropping the price by just 15 cents will be more effective for increasing sales revenue than increasing advertising expenditure by $500. To test the claim, we begin with the simple Model 3 where this proposition is equivalent to

$$H_0 : sales(price - 0.15, advert) - sales(price, advert + 0.5) \geq 0 \qquad (*)$$

what is the same as $H_0 : \beta_1 \cdot (-0.15) - \beta_2 \cdot 0.5 \geq 0$ with alternative $H_1 : ... < 0$. Note that her proposal is based on the estimates, i.e., on the inequality -7.91*(-0.15) - 1.86*0.5 = 0.26 > 0. To make sure that it holds in general, we have to test (*) with, say, 5% significance.

```
ols sales 0 price advert
restrict
 -0.15*b[price] - 0.5*b[advert] = 0
end restrict
```

```
Test statistic: F(1, 72) = 0.461561, with p-value = 0.499074
```

Since 0.499/2=0.25>0.05, we have no ground to reject H_0.

3.1 exercise. Test similar hypothesis for Model 5. Now the difference in (*) also depends on the values of `price` and `advert`, therefore test the hypothesis at the medians of these two variables.

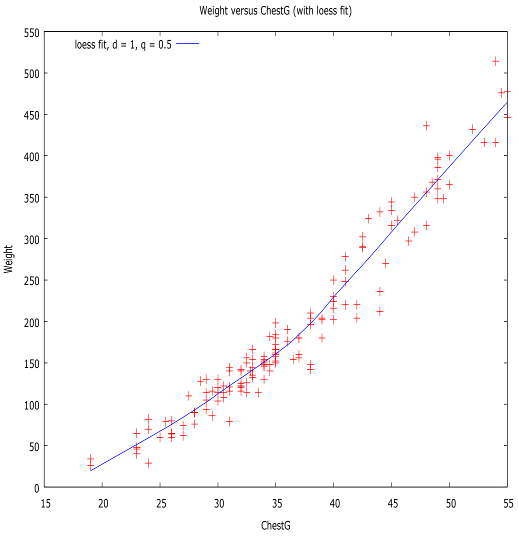
3.2 example. Data set Bears.csv contains 143 observations on 9 variables. The set originates from the study on wild bears aimed to help hunters to estimate the weight of a bear based on other measurements (this would be used because in the forest it is easier to measure, say, the length of a bear than to weigh it). Wild bears were anesthetized, and their bodies were measured and weighed. The nine variables are

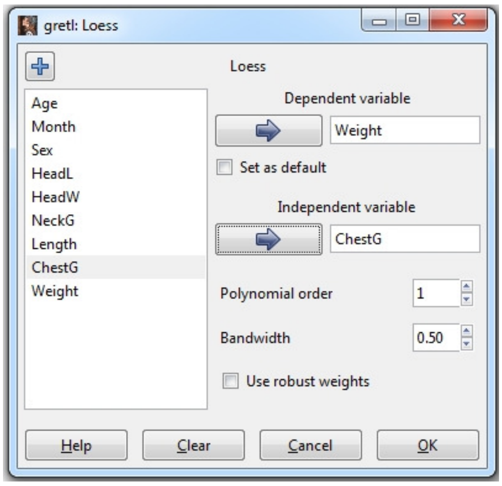
`Age`	is in months
`Month`	is the month of measurement
`Sex`	is coded with 1 = male and 2= female
`HeadL`	is head length (inches)
`HeadW`	is width of head (inches)
`NeckG`	Girth (distance around) the neck, in inches
`Length`	is length of body (inches)
`ChestG`	Girth (distance around) the chest, in inches
`Weight`	is measured in pounds

It will be convenient in the future if we sort all the data by `ChestG`: go to Datal Sort data… and select sort key `ChestG`.

The correlation matrix shows (click Ctrl+A, right click and choose Correlation matrix) that Weight correlates best with ChestG, therefore we shall start with Weight vs ChestG regression model.

```
Correlation Coefficients, using the observations 1 - 143
(missing values were skipped)

          Age         Month           Sex         HeadL         HeadW
       1.0000        0.0219        0.1188        0.6869        0.6692  Age
                     1.0000       -0.0885        0.0572        0.0114  Month
                                   1.0000       -0.2834       -0.2957  Sex
                                                 1.0000        0.7436  HeadL
                                                               1.0000  HeadW

        NeckG        Length        ChestG        Weight
       0.7338        0.6906        0.7342        0.7740  Age
       0.1212        0.0819        0.1284        0.0859  Month
      -0.3476       -0.2568       -0.2600       -0.2972  Sex
       0.8624        0.8952        0.8543        0.8333  HeadL
       0.8054        0.7363        0.7560        0.7556  HeadW
       1.0000        0.8730        0.9399        0.9433  NeckG
                     1.0000        0.8887        0.8746  Length
                                   1.0000        0.9660  ChestG
                                                 1.0000  Weight
```





Dependent variable: Weight

□ Set as default

Independent variable: ChestG

Polynomial order: 1

Bandwidth: 0.50

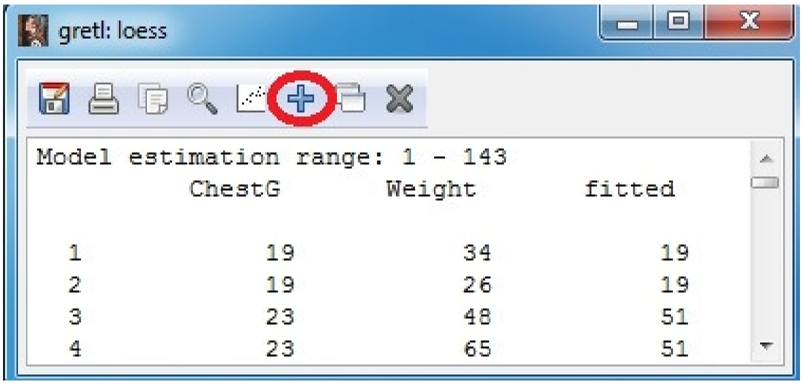
□ Use robust weights

Figure 3.3. The loess fit (left) and respective box (right)

The command

```
gnuplot Weight ChestG --output=display --loess-fit
```

displays the nonparametric loess regression curve (see Fig. 3.3, left, and LN, p.2-5). The same picture with some more flexible options can be pro-



```
Model estimation range: 1 - 143
            ChestG        Weight        fitted

  1             19            34            19
  2             19            26            19
  3             23            48            51
  4             23            65            51
```

duced through the menu bars: go to Model| Robust estimation| Loess…| OK (see Fig. 3.3, right) (to save the fitted values, click on the plus sign as shown above). The table consisting of the ChestG and fitted columns can already be used for estimating weight in the forest.

The loess curve resembles a broken or segmented line, therefore we can also use OLS and try to describe the model via

$$Weight = \begin{cases} \beta_0 + \beta_2 ChestG + \varepsilon & \text{for } ChestG \le 37 \\ (\beta_0 + \beta_1) + (\beta_2 + \beta_3) ChestG + \varepsilon & \text{for } ChestG > 37 \end{cases} =$$
$$= (\beta_0 + \beta_1\, big) + (\beta_2 + \beta_3\, big)\, ChestG + \varepsilon =$$
$$= \beta_0 + \beta_1\, big + \beta_2 ChestG + \beta_3\, bigChG + \varepsilon$$

The new variables big and the interaction term bigChG = big*ChestG are defined here as

```
series big = ChestG>37
series bigChG = big*ChestG
```

(thus Weight is described by one line untill ChestG≤37 and another one for bigger values).

```
Model 1: OLS, using observations 1-143
Dependent variable: Weight

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -159.819      18.3052      -8.731    7.13e-015 ***
  big        -290.817      33.5913      -8.658    1.08e-014 ***
  ChestG        9.11035     0.585834    15.55     3.20e-032 ***
  bigChG        7.68684     0.853276     9.009    1.45e-015 ***

Log-likelihood      -648.8385   Akaike criterion     1305.677
Schwarz criterion    1317.528   Hannan-Quinn         1310.493
```

The premium values β_1 and β_3 are quite significant, thus it is sensible to use this model. After saving fitted values as ols_fit, we can draw Fig. 3.4 (both model give almost the same predicted values therefore we can use any of them).

Similar results can be obtained with R (use the function loess or the segmented function from the “segmented” package). Repeat the OLS variant with R using the dummy variable big.

3.2 exercise. Use the same data, begin with a full model (i.e., the one with all the original data Age, Month etc) and end with the “best” model (when in the full model window, go to Tests| Omit variables and eliminate the least significant variables one by one; as an alternative, check the “Sequential elimination …” button).

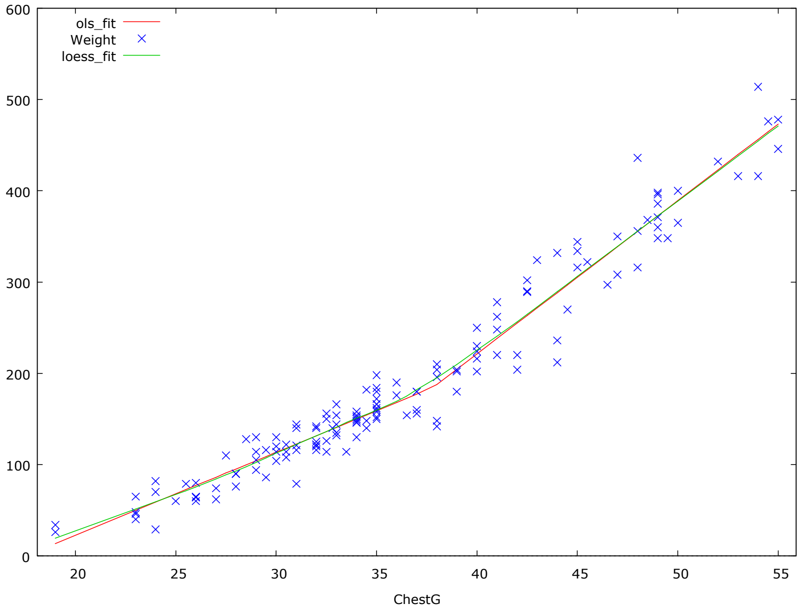


Figure 3.4. Red (OLS) and green (loess) predictions of `Weight` almost coincide.

3.3 exercise. Import the time series in shipm.txt which contains value of shipments, in millions of dollars, monthly from January, 1967 to December, 1974 (this represents manufacturers' receipts, billings, or the value of products shipped, less dicounts, and allowances, and excluding freight charges and excise taxes; shipments by foreign subsidiaries are excluded, but shipments to a foreign subsidiary by a domestic firm are included). Find its trend as a broken line and extend the trend 12 months ahead. Isn't the model for logarithms better? (keep in mind possible heteroskedasticity and normality of residuals). ◀◀

3.4 exercise. Is it possible to predict graduation rates `grad` from freshman test scores `sat`? Based on the average SAT score of entering freshmen at a university, can we predict the percentage of those freshmen who will get a degree there within 6 years? We use a random sample of 20 universities from the 248 national universities listed in the 2005 edition of *America's Best Colleges,* published by *U.S.News & World Report* (see the file univ1.txt).

1. Draw a `sat`-`grad` scatter diagram.
2. Create a linear model $\texttt{grad} = \beta_0 + \beta_1\texttt{sat} + \varepsilon$ and add a regression line to the graph.
3. What is the meaning of β_1?
4. Are the residuals normal?
5. Transform `PrivState` variable into a dummy one and include it into a model. What can you tell about the differences between the models?
6. Draw a graph presented in Fig. 3.3; why the black line is higher?

```
univ1=read.table(file.choose(),header=TRUE) # go to univ1.txt
attach(univ1)
plot(sat,grad,col=ifelse(PrivState=="P",1,2),pch=ifelse(PrivState=="P",15,16))
points(sat,grad+3,pch=ifelse(PrivState=="P","P","S"),col=ifelse(PrivState=="P",
1,2))
abline(lm(grad~sat,data=univ1[PrivState=="P",]))
abline(lm(grad~sat,data=univ1[PrivState=="S",]),col=2)
legend(750,95,c("P","S"),lty=1,col=1:2)
```

7. A $100(1-\alpha)\%$ prediction interval for a future Y observation to be made when $X = X^*$ is $\hat{\beta}_0 + \hat{\beta}_1 X^* \pm t_{\alpha/2,n-2} \cdot s_\varepsilon \sqrt{1+1/N+(X^*-\bar{X})^2/\sum(X_i-\bar{X})^2}$. Estimate the interval for the state universities if `sat`=1200.

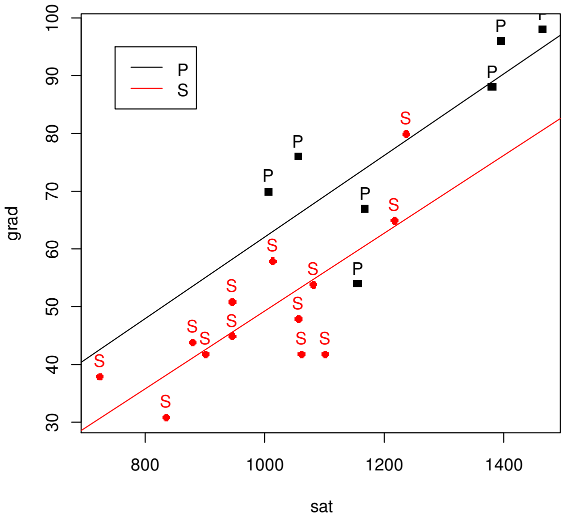


Figure 3.5. `sat-grad` scatter diagram for private (black) and state (red) universities together with two regression lines for both groups.

3.2. Multicollinearity

Consider the problem faced by the marketing executives at Big Andy's Burger Barn when they try to estimate the increase in sales revenue attributable to advertising that appears in newspapers and the increase in sales revenue attributable to coupon advertising. Suppose that it has been common practice to coordinate these two advertising devices, so that at the same time that advertising appears in the newspapers there are flyers distributed containing coupons for price reductions on hamburgers. If variables measuring the expenditures on these two forms of advertising appear on the right-hand side of a sales revenue equation such as in Model 3, then the data on these two variables will show a systematic, positive relationship. Although it is clear that total advertising expenditure increases sales revenue, however, because the two types of advertising expenditure move together (i.e., they are highly correlated or collinear), it may be difficult to sort out their separate effects on sales revenue.

3.3 example. 1970 m. JAV Energijos departamentas atliko tyrimą, susijusį su naftos produktų skirstymu visoms 50 valstijų pagal produktų ankstesnį naudojimą, gyventojų skaičių, autokelių ilgį ir pan. Duomenų rinkinyjs gas10.txt yra pateikta dalis surinktos informacijos. Čia

`PCON`	naftos produktų poreikis valstijoje (trilijonais BTU)
`POP`	gyventojų skaičius
`REG`	registruotų automobilių skaičius (tūkstančiais)

TAX	benzino akcizo mokestis (centais už galoną)
UHM	greitkelių ilgis (myliomis)

Iš ekonominių samprotavimų aišku, kad naftos produktų poreikio funkcija turėtų elgtis taip:

$$\text{PCON} = f(\underset{+}{\text{POP}}, \underset{+}{\text{REG}}, \underset{+}{\text{UMH}}, \underset{-}{\text{TAX}})$$

t.y., kai valstijos POP auga, PCON taip pat turėtų augti (taigi regresijos koeficientas prie POP turėtų būti + ir t.t.). Antra vertus, POP, REG ir UHM yra tikriausiai smarkiai koreliuoti (t.y., ir multikolinearūs). Iš tikrųjų, taip ir yra:

```
Correlation Coefficients, using the observations 1 - 50
5% critical value (two-tailed) = 0.2787 for n = 50

          POP           REG           UHM
       1.0000        0.9806        0.9645  POP
                     1.0000        0.9786  REG
                                   1.0000  UHM
```

Analizę pradėkime paprastu modeliu

```
Model 1 - Dependent variable: PCON

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       21.6762      58.2181        0.3723   0.7113
  POP          0.125892     0.00873175   14.42     4.57e-019 ***

R-squared           0.812405   Adjusted R-squared   0.808497
Log-likelihood     -354.5496   Akaike criterion     713.0993
Schwarz criterion   716.9233   Hannan-Quinn         714.5555
```

Kiek įtarimų kelia nereikšmingas laisvasis narys, tačiau vietoje to, kad gilintumėmės į modelio be laisvojo nario analizę, į modelį iš karto įjunkime (mažai su POP koreliuotą) mokesčių TAX narį:

```
Model 2 - Dependent variable: PCON

             coefficient    std. error    t-ratio    p-value
  ----------------------------------------------------------
  const      617.318       199.837         3.089    0.0034    ***
  POP          0.119839      0.00827744   14.48     6.05e-019 ***
  TAX        -56.2393       18.1762       -3.094    0.0033    ***

R-squared           0.844151   Adjusted R-squared   0.837519
Log-likelihood     -349.9148   Akaike criterion     705.8296
Schwarz criterion   711.5657   Hannan-Quinn         708.0139
```

Naujas modelis yra geresnis tiek Akaike's, tiek Schwarz'o prasmėmis, o jo koeficientų ženklai yra tokie, kokių tikėjomės. Pabandykime dar pagerinti šį modelį, į jį įjungdami visus kintamuosius:

```
Model 3 - Dependent variable: PCON

             coefficient     std. error    t-ratio   p-value
  -----------------------------------------------------------
  const      387.631         146.200         2.651    0.0110    **
  POP         -0.00663039      0.0294278    -0.2253   0.8228
  REG         -0.0524422       0.0579811    -0.9045   0.3706
  TAX        -36.3691         13.3000       -2.735    0.0089    ***
  UHM         61.0537         10.4475        5.844    5.32e-07  ***

R-squared            0.924275   Adjusted R-squared   0.917544
Log-likelihood      -331.8702   Akaike criterion     673.7405
Schwarz criterion    683.3006   Hannan-Quinn         677.3810
```

Jo `R-squared` dar padidėjo, tačiau, modelį papildžius naujais kintamaisiais, taip būna visuomet. Svarbiau yra tai, kad padidėjo `Adjusted R-squared`, o Akaike's ir Schwarz'o statistikos sumažėjo, taigi formaliai žiūrint paskutinis modelis yra „geriausias". Antra vertus, koeficientų ženklai dabar „neteisingi", o anksčiau buvęs reikšmingu `POP` tapo nereikšmingu. Visa tai aiškūs multikolinearumo požymiai, o tuo įsitikinti galima taip:

1. apskaičiuosime šio modelio koeficientų įvertinių dispersijos daugiklius:

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

            POP    26.379
            TAX     1.117
            REG    43.937
            UHM    25.255
```

(modelio lange nueikite į Tests| Collinearity; kadangi „didumo" kintamųjų VIF'ai dideli, tai kintamųjų nereikšmingumas gali būti tariamas);

2. valstijos didumą nusakantys keli kolinearūs kintamieji individualiai yra "nereikšmingi", tačiau jų bendras poveikis tikrai nėra nulinis – jungtinę hipotezę $H_0 : \beta_1 = 0, \beta_2 = 0, \beta_4 = 0$ tikriname taip: Model 3 lange nuvairuokite į Tests| Linear restrictions ir įrašykite

```
b[POP]=0
b[REG]=0
b[UHM]=0
```

paspaudę OK, pamatysite atsakymą:

```
Test statistic: F(3, 45) = 153.549, with p-value = 1.19407e-023
```

taigi H_0 reikia neabejotinai atmesti.

A variant: go to Tests| Omit variables and choose `POP`, `REG`, and `UHM`.

Su R tą patį gautume su tokiu skriptu:

```
gas10=read.table(file.choose(),header=T)
```

```
mod3=lm(PCON~POP+REG+TAX+UHM,data=gas10)
mod3.be=lm(PCON~TAX,data=gas10)
anova(mod3,mod3.be)

Analysis of Variance Table

Model 1: PCON ~ POP + REG + TAX + UHM
Model 2: PCON ~ TAX
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     45  1704795
2     48 19156148 -3 -17451353 153.55 < 2.2e-16 ***
```

(priminsime, kad šis testas atliekamas taikant F – testą, t.y., lyginant ribotojo ir neribotojo modelių RSS'us, žr. LN, 4.7 skyrelį).

Tolimesnis mūsų elgesys priklauso nuo tyrimo tikslo. Jei analizės tikslas būtų prognozuoti naftos poreikį valstijoje, tai paskutinis modelis visai tinkamas. Antra vertus, jei tyrimo tikslas būtų kiekvienos kintamųjų grupės (viena grupė susijusi su valstijos didumu – ją sudaro `POP`, `UHM` ir `REG`, o kitą sudaro `TAX`) įtakos nustatymas, tai šią analizę kol kas geriausiai atlieka Modelis 2 su „didumą" aprašančiu `POP` (Modelis 3 su visais „didumą" aprašančiais kintamaisiais dėl multikolinearumo yra blogas). Šią analizę patikslintume, jei į modelį įtrauktume ne individualius „didumo" kintamuosius, o jų pagrindines komponentes [ČM2, 244 psl.] (angl. principal components). GRETL'o lange nueikite į View| Principles components ir pasirinkite `POP`, `REG` ir `UHM`[6]:

```
Principal Components Analysis

Eigenanalysis of the Correlation Matrix

Component  Eigenvalue  Proportion   Cumulative
    1        2.9491       0.9830       0.9830
    2        0.0356       0.0119       0.9949
    3        0.0153       0.0051       1.0000

Eigenvectors (component loadings)

               PC1      PC2      PC3
POP          0.577    0.684    0.448
REG          0.579    0.044   -0.814
UHM          0.576   -0.729    0.370
```

Kitaip sakant, pirmoji pagrindinė komponentė $\texttt{PC1} = 0.577\texttt{POP} + 0.579\texttt{REG} + 0.576\texttt{UHM}$ yra sukaupusi 98,3% informacijos apie visą trijų kintamųjų sistemą, todėl ji geras jungtinis didumo koeficientas.

```
Model 4 - Dependent variable: PCON

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        1095.74       158.520      6.912     1.11e-08  ***
  TAX           -48.7474      15.3711    -3.171     0.0027    ***
  PC1           351.299       19.7814    17.76      1.76e-022 ***
```

[6] Išsaugoti šiuos naujus kintamuosius galima komponenčių lange paspaudus ✚ ; komandinis variantas yra toks: `pca POP REG UHM --save-all`

```
R-squared            0.889642   Adjusted R-squared   0.884946
Log-likelihood      -341.2857   Akaike criterion     688.5714
Schwarz criterion    694.3075   Hannan-Quinn         690.7557
```

Šis modelis panašus į Modelį 2, tačiau pagal visus kriterijus „geresnis“ už jį. Taigi renkamės šį modelį ir darome išvadą, kad valstijos „didumui“ padidėjus 1, naftos produktų poreikis padidėja 351.299 vienetais, o benzino mokesčius padidinus 1, poreikis sumažėja 48.747 vienetais. ◀◀

3.5 exercise. The file cars.txt contains observations on the following variables for 392 cars:

MPG	miles per gallon
CYL	number of cylinders
ENG	engine displacement in cubic inches
WGT	vehicle weight in pounds

Suppose we are interested in estimating the effect of CYL, ENG, and WGT on MPG. All the explanatory variables are related to the power and size of the car. Although there are exceptions, overall we would expect the values for CYL, ENG, and WGT to be large for large cars and small for small cars. They are variables that are likely to be highly correlated and whose separate effect on MPG may be difficult to estimate.

1. What about the correlation between explanatory variables?
2. Estimate the regression of MPG on CYL (Model 1). What sign do you expect at CYL?
3. Draw a scatter diagram of MPG vs CYL together with the regression line. Is the straight line a good approximation to your data?
4. Estimate the regression model of MPG on all the explanatory variables (Model 2). Is it a better model? Do you see any signs of collinearity? Individually, CYL and ENG are insignificant. Are they jointly insignificant?
5. CYL is in fact a group name but not a continuous variable. To convert it to a set of dummy variables, first make it discrete (select it, right-click on it, choose Edit attributes and check the “Treat this variable as discrete” box) and then go to Add| Dummies for selected discrete variables| OK.
6. Create a regression of MPG on DCYL_2,…,DCYL_5 (Model 3). Why we do not include DCYL_1 into the model? Is Model 3 better than Model 1?
7. Find two principal components of ENG and WGT. Estimate the regression model of MPG on DCYL_2,…,DCYL_5 and PC1 (Model 4). Is it better than other models?
8. What is the meaning of all the coefficients in Model 4? Derive out of this model a formula of MPG for the cars with 3 cylinders.
9. Repeat the analysis with R. To recreate Model 3, use mod=lm(MPG~factor(CYL), data=CARS). To calculate AIC, use AIC(mod). To find principal components, use

```
aaa=prcomp(~ ENG+WGT, data = CARS, scale = TRUE)
PC1=aaa$x[,1]
```
◀◀

3.6 exercise. Use the data in hsb.txt for this exercise.

1. Estimate the principle components of the system of five variables `RDG`, `WRTG`, `MATH`, `SCI`, and `CIV`. Use the first one, i.e., `PC1`, as the indicator of student's achievment.
2. Estimate a regression model relating `PC1` to all explanatory variables (first, transform qualitative variables to dummy variables).
3. Go to Tests| Omit variables and check the "Sequential elimination …" box. Comment the results.

3.7 exercise. [redacted] Is it true that the same regression model describes college grade point averages for male and female college athletes? Use the data in GPA3.txt, where

`term`	fall = 1, spring = 2
`sat`	SAT score
`tothrs`	total hours prior to term
`cumgpa`	cumulative GPA
`season`	=1 if in season
`frstsem`	=1 if student's 1st semester
`crsgpa`	weighted course GPA
`verbmath`	verbal SAT to math SAT ratio
`trmgpa`	term GPA
`hssize`	size high school graduating class
`hsrank`	rank in h.s. class
`id`	student identifier
`spring`	=1 if spring term
`female`	=1 if female
`black`	=1 if black
`white`	=1 if white
`ctrmgpa`	change in trmgpa
`ctothrs`	change in total hours
`ccrsgpa`	change in crsgpa
`ccrspop`	change in crspop
`cseason`	change in season
`hsperc`	100*(rank/hssize)
`football`	=1 if football player

Consider two models:

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + \varepsilon$$

and

$$\begin{gathered} cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \\ \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + \varepsilon \end{gathered}$$

The null hypothesis that *cumgpa* follows the same model for males and females is stated as

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0 .$$

1. Create both models.
2. What is the meaning of the δ coefficients?
3. Test the null hypothesis with both gretl and R. Comment.

4. In many cases, it is more interesting to allow for an intercept difference between the groups and then to test for the equality of the slopes. In other words, test $H_0: \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$.

3.8 exercise. The data set `aa` is given as

```
aa = structure(list(x1 = c(0.4, 2.8, 4, 6, 1.1, 2.6, 7.1, 5.3, 9.7, 3.1,
9.9, 5.3, 6.7, 4.3, 6.1, 9, 4.2, 4.5, 5.2, 4.3), x2 = c(19.7, 19.1, 18.2,
5.2, 4.3, 9.3, 3.6, 14.8, 11.9, 9.3, 2.8, 9.9, 15.4, 2.7, 10.6, 16.6, 11.4,
18.8, 15.6, 17.9), y = c(19.7, 19.3, 18.6, 7.9, 4.4, 9.6, 8, 15.7, 15.4,
9.8, 10.3, 11.2, 16.8, 5.1, 12.2, 18.9, 12.2, 19.3, 16.5, 18.4)), .Names =
c("x1", "x2", "y"), class = "data.frame", row.names = c(NA, -20L))
```

Try to best describe `y` in terms of `x1` and `x2`. Hint: [redacted] ◀◀

3.3. Heteroskedasticity

If the homoskedasticity assumption $\operatorname{var}\varepsilon_i \equiv \sigma^2$ (more precisely, assumption $\operatorname{var}(\varepsilon \mid \mathbf{X}) \equiv \sigma^2$) fails, we say that errors (or the regression model) are *heteroskedastic*. Heteroskedasticity does not cause bias or inconsistency in the $\hat{\beta}_m^{OLS}$. However, without the homoskedasticity assumption $\hat{\beta}_m^{OLS}$ are no longer the best, the OLS estimators of the $\operatorname{var}\hat{\beta}_m$ are biased, and the usual OLS t- statistics do not have Student‘s distribution even in large samples. Similarly, F statistic no longer has Fisher‘s distribution, and the *LM* statistic no longer has an asymptotic χ^2 distribution. Thus, we have to take into account the fact that $\operatorname{var}\varepsilon_i \not\equiv \sigma^2$.

To analyze the model for heteroskedasticity

- create OLS model and visually examine its residuals for heteroskedasticity or/and apply Breusch-Pagan test or/and White test
- correct the model with weights: use the WLS or gls procedures
- alternatively, use the White correction (robust standard errors)

3.9 exercise. The file data8-2.txt contains aggregate personal income and expenditures on domestic travel (1993) for the U.S. states and Washington, D.C. (51 observation):

`exptrav`	travel expenditures in billions of dollars
`income`	personal income in billions of dollars
`pop`	population in millions

Consider the following Engel curve relation:

$$exptrav = \beta_0 + \beta_1 income + \varepsilon \qquad (*)$$

We might expect that the variances of the errors of the OLS model (*) are heteroskedastic and increase together with population (a sensible specification is therefore $\sigma_i = \sigma pop_i$). We use the Harvey-Godfrey procedure to test this claim: create an auxiliary regression $\log \hat{\varepsilon}^2 = \alpha_0 + \alpha_1 pop + u$; in order to test the hypothesis $H_0 : \alpha_1 = 0$ (i.e., the errors of (*) are homoskedastic with respect to `pop`), estimate its *LM* statistics (= 51* R^2_{aux}) and calculate respective $p-$value $P(\chi^2_1 > LM)$.

```
ols exptrav 0 income
Model 1: OLS, using observations 1-51
Dependent variable: exptrav
             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        0.498120     0.535515      0.9302   0.3568
  income       0.0555731    0.00329311   16.88     4.70e-022 ***
series logres = log($uhat^2)
ols logres 0 pop
pvalue X 1 $nobs*$rsq
Chi-square(1): area to the right of 10.0207 = 0.00154792
(to the left: 0.998452)
```

We see that the $p-$value is much less than 0.05, therefore we divide each term of (*) by `pop` and hopefully obtain a homoskedastic model

$$\frac{exptrav}{pop} = \beta_0 \frac{1}{pop} + \beta_1 \frac{income}{pop} + \varepsilon^* \qquad (**)$$

which can be estimated by OLS (broadly speaking, if population has a role in a model, it is generally a good practise to express the model in per-capita terms). Recall that to apply the OLS to (**) means to find b_0 and b_1 such that the expression

$$\sum \left(\frac{exptrav_i}{pop_i} - \left(b_0 \frac{1}{pop_i} + b_1 \frac{income_i}{pop_i} \right) \right)^2 = \sum \frac{1}{pop_i^2} \left(b_0 + b_1\, income_i \right)^2$$

attains its minimum. Note that rhs expression stands for the WLS with $(1/pop)^2$ as weight.

```
series pcexp = exptrav/pop
series pcincm = income/pop
series invpop = 1/pop
ols pcexp invpop pcincm     # OLS with no intercept

Model 3: OLS, using observations 1-51
Dependent variable: pcexp

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  invpop       0.736824     0.332260      2.218    0.0312    **
  pcincm       0.0585518    0.0122610     4.775    1.66e-05 ***
```

It should be pointed out that while OLS is applied to the transformed equation, the interpretation of the coefficients is for the original equation. Thus, the estimated coefficient of `1/pop` is that of the intercept term, and the estimated coefficient 0.0586 of `income/pop` is that of the marginal propensity to spend on travel with respect to income.

The OLS model 3 is equivalent to the WLS model 4:

```
series wtvar = 1/(pop^2)        # create weights
wls wtvar exptrav const income  # WLS with ww as weights
```

```
Model 4: WLS, using observations 1-51
Dependent variable: exptrav
Variable used as weight: wtvar

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       0.736824     0.332260      2.218    0.0312   **
  income      0.0585518    0.0122610     4.775    1.66e-05 ***
```

To finalize our analysis – Model 1 suggests that once `income` increases by 1 (billion), the `exptrav` will increase on average by 0.056 (billions). Do not forget that this number is only an estimate of marginal propensity, the true value of it is somewhere in the interval (0.056-2*0.003,0.056+2*0.003). Similarly, Model 3 suggests the interval (0.059-2*0.012,0.059+ 2*0.012) (surprisingly, now the standard deviation is even bigger, but these values are only estimates). Both methods give slightly different estimates of β_1, but you should keep in your mind that two unbiased estimates do not necessarily coincide. The final point – if you correct standard deviations in Model 1 taking heteroskedasticity-robust standard errors with

```
ols exptrav const income --robust
```

you get (0.056-2*0.005,0.056+2*005), thus the advice in LN to "take care of heteroskedasticity only in severe cases" seems quite reasonable. ◀◀

3.4 example. The below-presented code generates a sample of 50 elements described by the equation $Y_i = 0 + 0 * i + \varepsilon_i$ where $\varepsilon_i \sim N(0, i^2)$. Since the errors are heteroskedastic, we estimate β_0 and β_1 twice, with OLS and WLS. To verify the claim that OLS is not efficient (i.e., there exists a method, namely WLS, with a smaller variance of estimators), we repeat the procedure 500 times and calculate sample variances of $\hat{\beta}_1^{OLS}$ and $\hat{\beta}_1^{WLS}$ - indeed, the weighted least squares estimate of $\beta_1 (= 0)$ (see Fig.3.3) has a smaller variance.

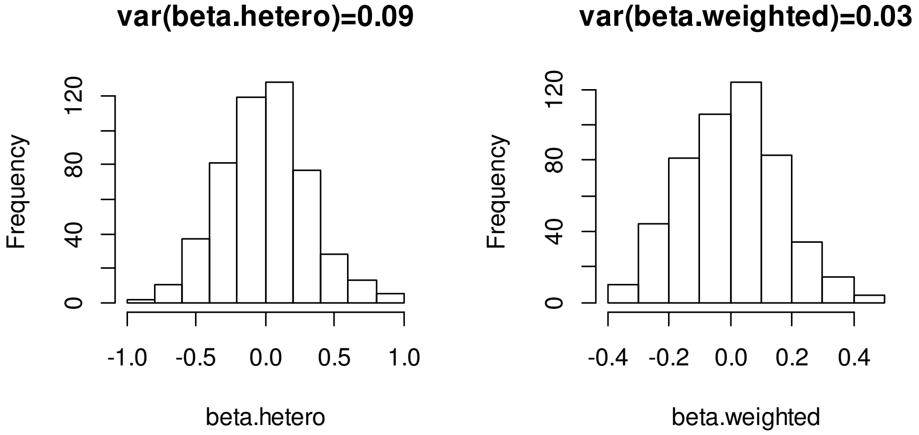


Figure 3.6. $\hat{\beta}_1^{WLS}$ has smaller spread of values around $\beta_1 = 0$

```
N=50
y=numeric(N)
beta.hetero=numeric(500)
beta.weighted=numeric(500)
ii=1:N

for(j in 1:500)
{
set.seed(j)
y=rnorm(N,sd=1:N)
mod.hetero=lm(y~ii)
beta.hetero[j]=mod.hetero$coef[2]

mod.weighted=lm(y~ii,weights=1/(ii^2))
beta.weighted[j]=mod.weighted$coef[2]
}

cat("var(beta.heterosk)=",var(beta.hetero),"\n")
cat("var(beta.weigted)=",var(beta.weighted),"\n")
par(mfrow=c(1,2))
hist(beta.hetero,main="var(beta.hetero)=0.09")
hist(beta.weighted,main="var(beta.weighted)=0.03")
windows() # open a new window
par(mfrow=c(1,1))
plot(ii,y) # y is from 500th loop
abline(0,0);abline(0,1.5,lty=2);abline(0,-1.5,lty=2)
```

3.10 exercise. In the above 3.4 example, we have proved that $\text{var}\,\hat{\beta}_1^{WLS} < \text{var}\,\hat{\beta}_1^{OLS}$. Examine whether it implies that in any of 500 pairs $(\hat{\beta}_1^{WLS} - 0)$ is less than $(\hat{\beta}_1^{OLS} - 0)$.

3.11 exercise. Redo 3.8 exercise from the menu lines. Redo the exercise with R. Add some graphs to your report (for example, draw the scatter diagram of the residuals of the OLS model (*) vs pop and another scatter diagram of the residuals of the OLS model (**) vs pop; explain the difference). ◀◀

The GLS procedure discussed in the previous example consists of dividing each variable (including the constant term) by σ_i and then applying OLS to the resulting transformed model. As the structure of the heteroskedasticity is generally unknown, a researcher must first obtain estimates of σ_i by some means and then use the WLS procedure (it is called the feasible or estimable GLS, FGLS or EGLS). The natural method of estimating the error variance is to exploit the information contained in the auxiliary equation:

1) Regress Y against a constant term and $X_1,...,X_k$ and obtain the residuals $e_i = \hat{\varepsilon}_i^{OLS}$.

2) Regress e_i^2 or $|e_i|$, or $\log e_i^2$ against a constant and $W_1,...,W_p$ (this is called an auxiliary regression) where W's are all or some X's, together with some (significant) squares and cross products and also some relevant outside variables. The FGLS uses the fitted value $\hat{e}_i^2$ as an estimate of σ_i^2, i.e., $wtvar_i = 1/\hat{e}_i^2$ as weights. A problem with the first two variants is that there is no guarantee that the predicted variances will be positive for all *i*. If any of them is negative, we cannot take the square root. If this situation arises for some observations, then we can use the original e_i^2 and take their positive square roots.

The FGLS estimates obtained in this way are consistent, as are OLS estimates. However, unlike OLS, the estimated variances of the estimates are consistent here, too. It should be noted that conventional methods of calculating R^2 are not valid (compute it as the square of the correlation between observed and fitted values of the original dependent variable).

3.5 example. The file data8-3.txt contains four columns of data for the U.S. states and Washington, D.C., for 1993:

`exphlth`	aggregate expenditures on health care in billions of dollars
`income`	aggregate personal income
`pop`	population in millions
`seniors`	percent 65 and over

The same data is present in GRETL, File| Open data| Sample file...| Ramanathan| data8-3. We have seen that population often induces heteroskedasticity and therefore expressing variables as per-capita, by dividing by population, is a useful way of reducing that effect:

$$\frac{exphlth}{pop} = \beta_0 + \beta_1 \frac{income}{pop} + \beta_2 seniors + \varepsilon .$$

We use GRETL| File| Script files| Practice file...| ps8-8:

```
open data8-3
genr y=exphlth/pop
genr x=income/pop
# estimate model by OLS and save absolute residuals, squared residuals,
# and their logs
ols y const x seniors
genr absuhat=abs($uhat)
genr usq=$uhat*$uhat
genr lnusq=ln(usq)
# generate square and cross product variables; the flag -o generates cross
# product
square x seniors -o
# Testing and estimation for the Glesjer approach
ols absuhat const x seniors sq_x sq_seniors
# estimate residual s.d. from the auxiliary regression
genr sigmahat=absuhat-$uhat
# compute LM test statistic and its p-value
genr LM1=$nrsq
pvalue X 4 LM1
# print sigmahat and note that only one estimate is negative
print sigmahat
# replace negative value with original sigmahat and get weights
genr d1=(sigmahat>0.0)
genr sigma2=(d1*sigmahat)+((1-d1)*absuhat)
genr wt1=1/(sigma2^2)
# Estimate model by FGLS which is the same as WLS
wls wt1 y const x seniors
# Testing and estimation for the Breusch-Pagan approach
ols usq const x seniors sq_x sq_seniors
# estimate residual s.d. from the auxiliary regression
genr usqhat1=usq-$uhat
# compute LM test statistic and its p-value
genr LM2=$nrsq
pvalue X 4 LM2
# print usqhat and note that several estimates are negative
print usqhat1
# replace negative values with original usqhat and get weights
genr d2=(usqhat1>0.0)
genr usqhat2=(d2*usqhat1)+((1-d2)*usq)
genr wt2=1/usqhat2
# Estimate model by FGLS which is the same as WLS
wls wt2 y const x seniors
# Testing and estimation for White's procedure
ols usq const x seniors sq_x sq_seniors x_seniors
genr usqhat3=usq-$uhat
# compute LM test statistic and its p-value
genr LM3=$nrsq
pvalue X 5 LM3
# print usqhat and note that several estimates are negative
print usqhat3
# replace negative values with original usqhat and get weights
genr d3=(usqhat3>0.0)
genr usqhat4=(d3*usqhat3)+((1-d3)*usq)
genr wt3=1/usqhat4
# Estimate model by FGLS which is the same as WLS
wls wt3 y const x seniors
# Test using the Harvey-Godfrey approach
ols lnusq const x seniors sq_x sq_seniors
# compute LM test statistic and its p-value
genr LM4=$nrsq
# since the p-value is high, we do not reject homoscedasticity
pvalue X 4 LM4
# Because the coefficients for x and x-squared are significant, another LM
# test is done with just these
ols lnusq const x sq_x
genr lnusqhat=lnusq-$uhat
# compute LM test statistic and its p-value
genr LM5=$nrsq
```

```
pvalue X 2 LM5
# since the p-value is acceptable, we reject homoscedasticity and
# procede with WLS estimation
genr usqhat5=exp(lnusqhat)
genr wt4=1/usqhat5
wls wt4 y const x seniors
```

```
WLS, using observations 1-51
Dependent variable: y
Variable used as weight: wt4

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       0.0237519    0.437873     0.05424   0.9570
  x           0.0947065    0.0174083    5.440     1.77e-06 ***
  seniors     0.0743781    0.0166254    4.474     4.71e-05 ***
```

A simplified version of the above script (it carries out only the $\log e_i^2$ case) can be performed with the `hsk` command:

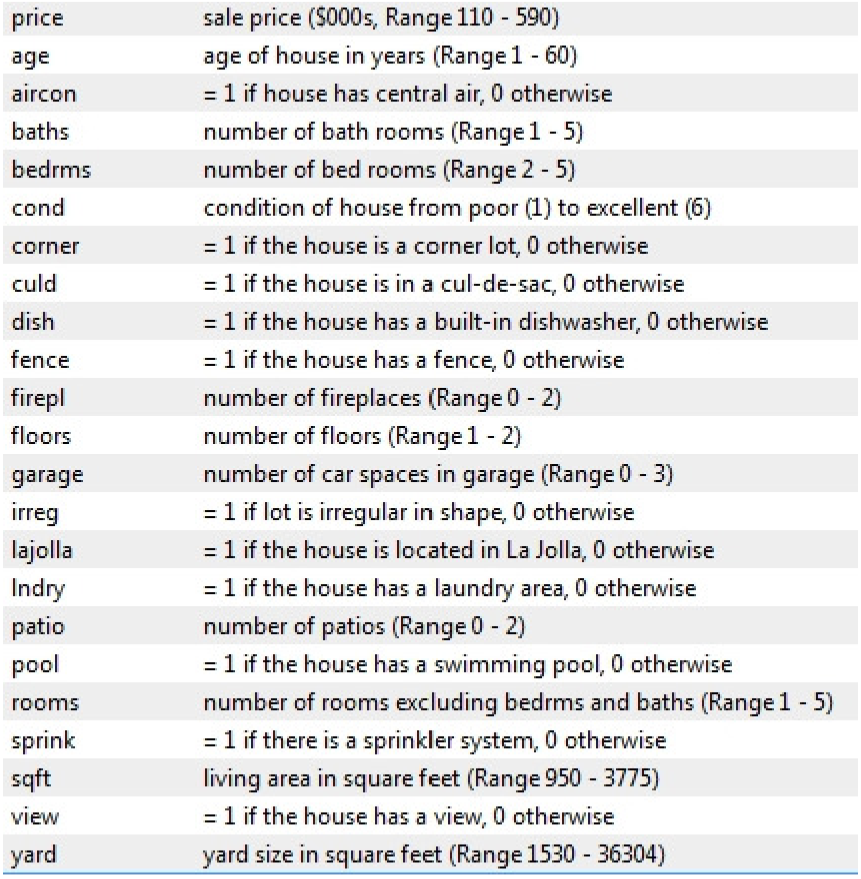
```
hsk y 0 x seniors
```

```
Heteroskedasticity-corrected, using observations 1-51
Dependent variable: y

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       0.0338012    0.384074     0.08801   0.9302
  x           0.0986142    0.0167724    5.880     3.83e-07 ***
  seniors     0.0677264    0.0114200    5.931     3.20e-07 ***
```

Both WLS regression results are very close. ◀◀

3.12 exercise. Repeat the WLS analysis without the DC observation (verify that this observation is an “outlier” in a sense). Compare it with the OLS regression. Do the same from the menu-driven windows. Do the same analysis with R.

3.13 exercise. The data set data7-11 (go to GRETL| Open data| Sample file…| Ramanathan| data7-11) contains the following variables:



price	sale price ($000s, Range 110 - 590)
age	age of house in years (Range 1 - 60)
aircon	= 1 if house has central air, 0 otherwise
baths	number of bath rooms (Range 1 - 5)
bedrms	number of bed rooms (Range 2 - 5)
cond	condition of house from poor (1) to excellent (6)
corner	= 1 if the house is a corner lot, 0 otherwise
culd	= 1 if the house is in a cul-de-sac, 0 otherwise
dish	= 1 if the house has a built-in dishwasher, 0 otherwise
fence	= 1 if the house has a fence, 0 otherwise
firepl	number of fireplaces (Range 0 - 2)
floors	number of floors (Range 1 - 2)
garage	number of car spaces in garage (Range 0 - 3)
irreg	= 1 if lot is irregular in shape, 0 otherwise
lajolla	= 1 if the house is located in La Jolla, 0 otherwise
lndry	= 1 if the house has a laundry area, 0 otherwise
patio	number of patios (Range 0 - 2)
pool	= 1 if the house has a swimming pool, 0 otherwise
rooms	number of rooms excluding bedrms and baths (Range 1 - 5)
sprink	= 1 if there is a sprinkler system, 0 otherwise
sqft	living area in square feet (Range 950 - 3775)
view	= 1 if the house has a view, 0 otherwise
yard	yard size in square feet (Range 1530 - 36304)

1. Consider the following model for real estate values:

$$price = \beta_0 + \beta_1 sqft + \beta_2 yard + \beta_3 pool + \varepsilon$$

1a. You suspect that the error term ε might be heteroskedastic and that the variance of ε is proportional to `sqft`. Describe step-by-step how you should use the WLS procedure to take care of the problem. Be sure to state the transformation you need to and the regression to be run. Write down the assumptions on ε and the properties of the WLS estimates. Carefully explain why your properties hold.
1b. Suppose you did not know the nature of heteroskedasticity and want to use the White test for it. Describe carefully all the steps needed to perform the test.
1c. Use data7-11 to estimate the above model. Repeat this process after adding more variables to the model.

2. Consider the model

$$price = \beta_0 + \beta_1 sqft + \beta_3 \log(sqft) + \beta_4 yard + \beta_5 \log(yard) + \varepsilon$$

2a. You suspect that the marginal effect of `sqft` (`yard`) on price decreases as `sqft` (respectively, `yard`) increases. If these hypotheses were true, what signs would you expect for β_3 and β_5? Justify your answer.
2b. You know from past studies that the variance of ε is proportional to to the size of the `sqft`. Describe step-by-step how you would apply the WLS prosedure that makes use of this information. Be sure to state what variables to generate and what regressions to run.

2c. In what way is the WLS procedure better than the OLS procedure?
2d. In the basic model, suppose the nature of heteroskedasticity is unknown. Describe the steps to be taken to perform the Harvey-Godfrey test for the model. To do this, first write the auxiliary equation for the error variance and state the null hypothesis of no heteroskedasticity. Then describe the regressions to run, how you will compute the test statistic, and what its distribution and d.f. are.
2e. Use the data of data7-11 to estimate the model in **2**. and implement the WLS and FGLS procedures discussed above. Repeat this process after adding more variables to the model.

3.14 exercise. [redacted] The data file cps4_small.txt contains 1000 observations on the following variables:

wage	earnings per hour
educ	years of education
exper	post education years experience
hrswk	usual hours worked per week
married	= 1 if married
female	= 1 if female
metro	= 1 if lives in metropolitan area
midwest	= 1 if lives in midwest
south	= 1 if lives in south
west	= 1 if lives in west
black	= 1 if black
asian	= 1 if asian

<u>Note on educ variable</u>. CPS reports educational attainment by category. For the purpose of illustrations, we assign the following numerical values for educ:

```
00. Less than 1st grade
03. 1st,2nd,3rd,or 4th grade
03. 5th or 6th grade
08. 7th and 8th grade
09. 9th grade
10. 10th grade
11. 11th grade
12. 12th grade no diploma
12. High school graduate - high school diploma or equivalent
13. Some college but no degree
14. Associate degree in college - occupation/vocation program
14. Associate degree in college - academic program
16. Bachelor's degree (for example: BA,AB,BS)
18. Master's degree (for example:MA,MS,MENG,MED,MSW, MBA)
21. Professional school degree (for example: MD,DDS,DVM,LLB,JD)
21. Doctorate degree (for example: PHD,EDD)
```

1a. Using the data in cps4_small.txt estimate the following wage equation with OLS and heteroskedasticity-robust standard errors:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4\, exper * educ + \varepsilon$$

Report the results.
1b. Add married to the equation and re-estimate. Holding educ and exper constant, do married workers get higher wages? Using a 1% significance level, test a null hypothesis that

wages of married workers are less than or equal to those of unmarried workers against the alternative that wages of married workers are higher.

1c. Plot the residuals from part 1a against the two values of `married`. Is there evidence of heteroskedasticity?

1d. Estimate the model in part 1a twice – once using observations on only married workers and once using observations on only unmarried workers. Use a two variances test (see GRETL| Tools| Test statistic calculator| 2 variances) to test whether the error variances for married and unmarried workers are different.

1e. Find GLS of the model in part 1a. Compare the estimates and standard errors with those obtained in part 1a.

1f. Find two 95% interval estimates for the marginal effect $\partial E\log(wage)/\partial educ$ for a worker with 25 years of experience. (Hint: the interval equals $\hat{\beta}_1+\hat{\beta}_4 exper \pm 2\sqrt{\mathrm{var}\left(\hat{\beta}_1+\hat{\beta}_4 exper\right)}\,|_{exper=25}$). To find the matrix $\mathrm{var}\,\hat{\hat{\beta}}$, use the line

```
ols l_wage 0 educ exper sq_exper exp_ed --vcv
```

****

2a. Consider the same data set and the same wage equation. Plot the OLS residuals against `educ` and against `exper`. What do they suggest?

2b. Test for heteroskedasticity using a Breusch-Pagan test where the variance depends on `educ`, `exper` and `married`. What do you conclude at a 5% significance level?

2c. Estimate a variance function that includes `educ`, `exper`, and `married` and use it to estimate the standard deviation for each observation.

2d. Find GLS estimates of the wage equation. Compare the estimates and standard errors with those obtained from OLS estimation with heteroskedasticity-robust standard errors.

2e. Find two 95% interval estimates for the marginal effect $\partial E\log(wage)/\partial exper$ for a worker with 20 years of experience. Use OLS with heteroskedasticity-robust standard errors for one interval and the results from 2d for the other. Comment on any differences.

3.15 exercise. White's test is a special case of the Breusch-Pagan test using a particular choice of auxiliary regressors. In R, the Breusch-Pagan test is available in `bptest()` from the lmtest package or `ncvTest()` from the car package. A worked example on how to perform the White test with `bptest` is provided in `help(CigarettesB,package=AER)`, based on an example from Baltagi's "Econometrics" textbook:

```
library(AER); ?CigarettesB
data(CigarettesB, package = "AER")
cig_lm2 <- lm(packs ~ price + income, data = CigarettesB)
bptest(cig_lm2, ~income*price+I(income^2)+I(price^2),data= CigarettesB)

        studentized Breusch-Pagan test
data:  cig_lm2
BP = 15.6564, df = 5, p-value = 0.007897
```

Note that the term `income*price` in the formula is the same as `income+price+I(income*price)`. Now, since White's test rejects the homoskedasticity hypothesis, we have to correct standard errors etc. Here is the R's function that returns regression results using White's standard errors (analyze the text!):

```
summaryw <- function(model)
{
s <- summary(model)
X <- model.matrix(model)
u2 <- residuals(model)^2
XDX <- 0
## here one needs essentially to calculate X'DX. But due to the fact that D
## is huge (NxN), it is better to do it with a cycle.
for( i in 1:nrow( X)) {
XDX <- XDX + u2[i]*X[i,]%*%t( X[i,])
}
XX1 <- solve( t( X)%*%X)
varcovar <- XX1 %*% XDX %*% XX1
stdh <- sqrt(diag(varcovar))
t <- model$coefficients/stdh
p <- 2*pnorm(-abs(t))
results <- cbind(model$coefficients, stdh, t, p)
dimnames(results) <- dimnames( s$coefficients)
results
}
```

Compare the usual `summary` and the new `summaryw`:

```
> summary(cig_lm2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.2997     0.9089   4.730 2.43e-05 ***
price        -1.3383     0.3246  -4.123 0.000168 ***
income        0.1724     0.1968   0.876 0.385818

> summaryw(cig_lm2)
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  4.299662  1.0589100  4.0604599 4.897616e-05
price       -1.338335  0.3319823 -4.0313441 5.545877e-05
income       0.172386  0.2287646  0.7535518 4.511184e-01
```

The difference in standard errors and the $p-$values is not very significant.

The exercice assignment itself is as follows: export `CigarettesB` as a text file (use `library(MASS); write.matrix(CigarettesB,”cig.txt”))` and repeat the calculations with GRETL. ◀◀

3.4. Serial Correlation (or Autocorrelation)

So far we have considered the case where (part of the) condition M3, namely, $\operatorname{cov}(\varepsilon_t, \varepsilon_s) = E\varepsilon_t\varepsilon_s = 0$ for all $t \neq s$, holds. What if this condition fails? The general case $\operatorname{cov}(\varepsilon_t, \varepsilon_s) = f(t,s)$ is unfeasible; the commonly used simplification assumes that $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ (here u_t is white noise (WN), i.e., it satisfies M3) or $\varepsilon_t = \rho_1\varepsilon_{t-1} + ... + \rho_r\varepsilon_{t-r} + u_t$ (the first process ε_t, $t = 2,...,T$, is termed AR(1) and the second AR(r)).

If the OLS procedure is used to estimate the unknown coefficients $\beta_m, m = 0,...,k,$ in $Y_t = \beta_0 + \beta_1 X_{1,t} + ... + \beta_k X_{k,t} + \varepsilon_t$, then the estimates $\hat{\beta}_m^{OLS}$ are still unbiased and consistent; however, they are no longer efficient.

The estimated variances of $\hat{\beta}_m^{OLS}$ will be biased and inconsistent, hence the tests of hypotheses are invalid.

3.6 example. This example contains two codes – the first is performed in GRETL and the second in R.

GRETL

The below-presented GRETL code generates a series of 60 elements described by the equation $Y_t = (\beta_0 + \beta_1 t + \varepsilon_t =) 0 + 1 * t + \varepsilon_t$, $\varepsilon_t = (\rho \varepsilon_{t-1} + u_t =) 0.9\, \varepsilon_{t-1} + u_t$, $t = 2,...,60$, $\varepsilon_1 = 0$, $Y_1 = 1$, where $u \sim WN(0, 2^2)$.

```
nulldata 60                 # create an empty session file of length 60
set seed 10000              # where to begin our random series
setobs 1 1 --time-series    # supply a time series structure to session:
                            # start=1, freq=1
series u = 2*normal()       # create a normal series with sd=2
series eps = 0
eps = 0.9*eps(-1) + u       # create AR(1) series of disturbances
series y = index + eps      # create a series
ols y 0 index               # OLS model
```

```
Model 1: OLS, using observations 1-60
Dependent variable: y

             coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const       -1.63939     0.710466     -2.307     0.0246    **
  index        0.948428    0.0202564    46.82      8.89e-048 ***

rho                   0.673803   Durbin-Watson          0.647530
```

```
series epshat = $uhat       # create residuals
gnuplot epshat --time-series --with-lines --output=display
```

The residuals (see Fig 3.6, left) show some persistence which is a clear sign that they do not make a WN. The DW statistics is far from 2, which once again indicates that residuals are, probably, an AR(1) process (that is, we have to correct the model for serial correlation). Recall that 1) the estimate $\hat{\beta}_1^{OLS}$ is not efficient (it may depart from its true value of 1 "too much") and 2) its `std.error` is calculated incorrectly (in fact, it is not 0.0203). To correct the estimates for serial correlation, we can use either the CORC or the Hildreth-Lu procedures (see below).

The main difference between the OLS model and the one which takes into consideration the autoregressive structure of errors is forecasting. The original model can be rewritten as the dynamic[7] model with "good" errors:

$$Y_t = \beta_0(1-\rho) + \beta_1\rho + \beta_1(1-\rho)\,t + \rho Y_{t-1} + u_t\,.$$

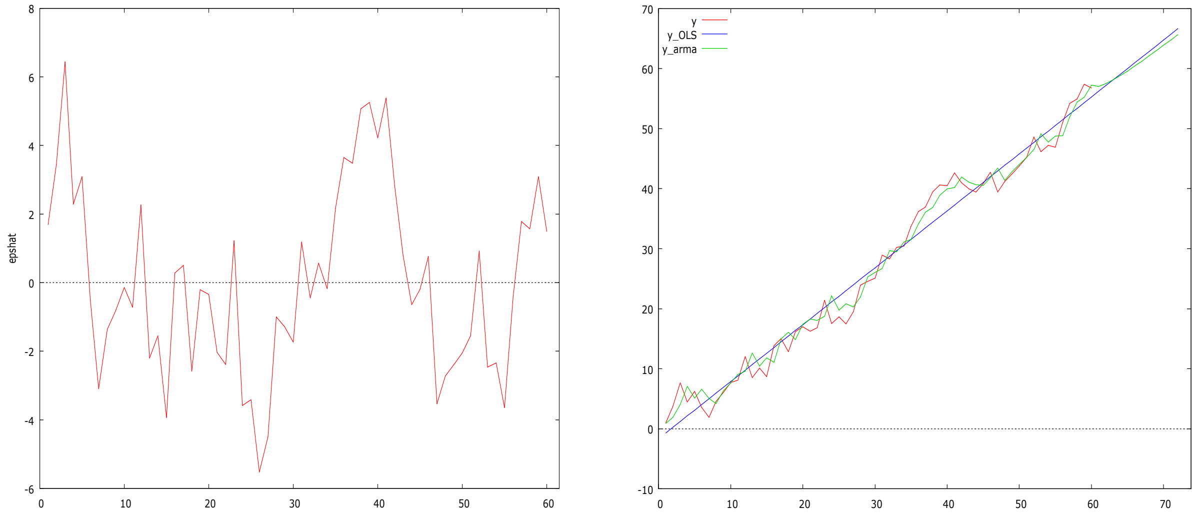


Figure 3.7. Residuals of the OLS model (left); 12-months-ahead forecast of Y_t with the OLS (blue) and corrected models (green)

This new model can be estimated with OLS and the fitted model is clearly <u>not</u> of the form $\hat{Y}_t = \hat{\gamma}_0 + \hat{\gamma}_1 t$. To forecast both the OLS and corrected model for 12 moments ahead, we continue with the following code:

```
dataset addobs 12     # we extend the time limits 12 months ahead
fcast y_OLS           # y_OLS is the 72 time moments forecast
smpl 1 60             # return to original sample
arma 1 0 ; y index    # the model assumes that errors are AR(1)
smpl full             # go to extended sample
fcast y_arma
gnuplot y y_OLS y_arma --time-series --with-lines --output=display
                      # see Fig. 3.4, right
```

Here, instead of the CORC method, we have used its more contemporary variant, namely, the FGLS inplemented in the `arma` function (see below).

3.16 exercise. Perform this example from the menu lines. ◀◀

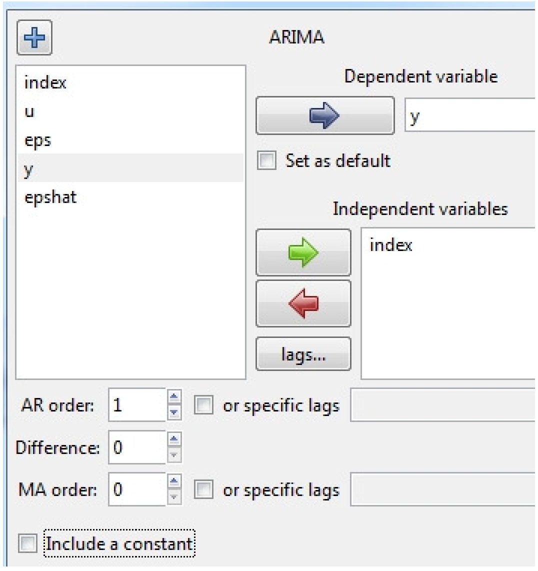
Recall that the OLS procedure has some drawbacks. To correct the estimates for serial correlation, we can use either the CORC or the Hildreth-Lu procedures. In the first case, go to Modell Time seriesl Cochrane-Orcutt…:

[7] „Dynamic" means that the rhs of the model contains lag or lags of Y_t. To forecasts Y_t such a model uses not only the present time information, but also the information contained in the past.

```
Performing iterative calculation of rho...

                 ITER       RHO        ESS
                   1      0.71468   203.634
                   2      0.71484   203.634

Model 3: Cochrane-Orcutt, using observations 2-50 (T = 49)
Dependent variable: y
rho = 0.714845

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  index        0.897247     0.0324263     27.67    3.72e-031 ***

Statistics based on the rho-differenced data:

[…]
rho                    -0.140911   Durbin-Watson        2.265139
```

The last line reports that DW statistics of $\hat{u}_t$ is quite close to 2, thus the model with AR(1) residuals ε_t is satisfactory and we can rely on its `std.error`.

In fact, the CORC procedure is interesting only for historical reasons – nowadays, the maximum likelihood method is preferred and it is implemented in `arima` function: go to Model| Time series| ARIMA… and fill in the boxes as shown on the right.



Dependent variable: y

□ Set as default

Independent variables: index

AR order: 1 □ or specific lags

Difference: 0

MA order: 0 □ or specific lags

□ Include a constant

```
Function evaluations: 19
Evaluations of gradient: 5

Model 4: ARMAX, using observations 1-50
Estimated using Kalman filter (exact ML)
Dependent variable: y
Standard errors based on Hessian

             coefficient   std. error      z       p-value
  --------------------------------------------------------
  phi_1        0.701244     0.0966655     7.254   4.04e-013 ***
  index        0.897466     0.0304986    29.43    2.52e-190 ***
```

Thus, the model $Y_t = \beta_1 t + \varepsilon_t$ which takes into account the first order serial correlation of residuals ε_t is given by

$$\begin{cases} \hat{Y}_t = 0.897\,t + \varepsilon_t \\ \qquad (0.03) \\ \varepsilon_t = 0.70\varepsilon_{t-1} + u_t \end{cases}$$

where the number in parenthesis is the standard error.

R

Now we shall perform another exercise – we compare the OLS and `arima` estimates of β_1. To generate 500 sequences of $Y's$, we use the following R's code:

```
N=60                              # the length of Y
y=numeric(N)
eps=numeric(N)
beta.OLS=numeric(500)
beta.arima=numeric(500)
index=1:N                         # index stands for t
for(j in 1:500)
{
set.seed(j)
for(i in 2:N) eps[i] = 0.9*eps[i-1]+rnorm(1,sd=2) # create AR(1) residuals
y=index+eps
beta.OLS[j] = lm(y~index)$coef[2]
library(forecast)
beta.arima[j] = Arima(y, order = c(1, 0, 0),
xreg=index-25,include.mean=T)$coef[3] # it is adviced to make mean(xreg)≈0
}
cat("mean(beta.OLS)=",mean(beta.OLS),"\n")
cat("mean(beta.arima)=",mean(beta.arima),"\n")
cat("var(beta.OLS)=",var(beta.OLS),"\n")
cat("var(beta.arima)=",var(beta.arima),"\n")
par(mfrow=c(1,3))
hist(beta.OLS,main="var(beta.OLS)=0.013")
hist(beta.arima,main="var(beta.arima)=0.009")
plot(y,type="l",main="The last realization of y")
Arima(y, order = c(1, 0, 0),xreg=index-25,include.mean=T)
```

```
mean(beta.OLS)= 0.9996
mean(beta.arima)= 0.9997
var(beta.OLS)= 0.0090
var(beta.arima)= 0.0060
```

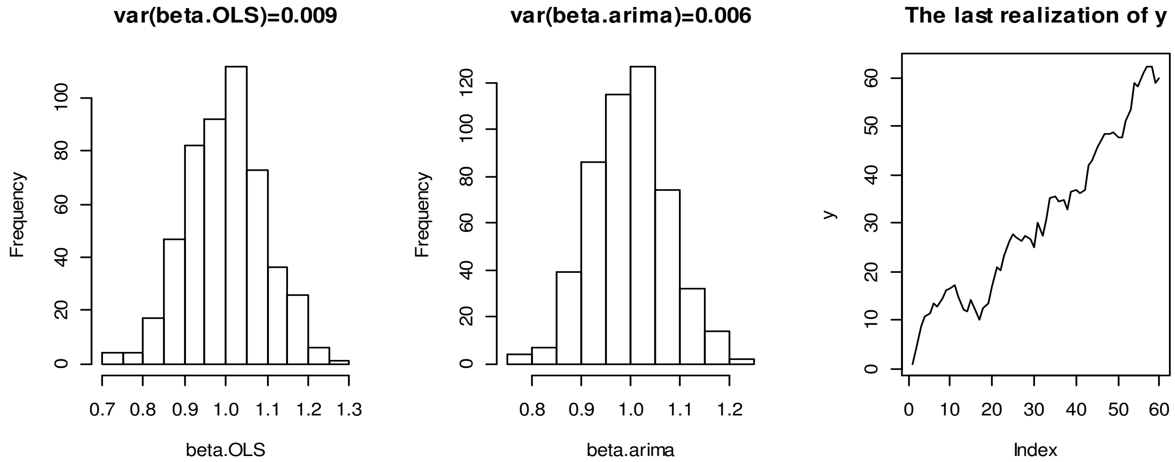


Figure 3.8. Typical trajectory of Y_t with serially correlated errors (right)

We see that $\widehat{\text{var}}\hat{\beta}_1^{OLS} > \widehat{\text{var}}\hat{\beta}_1^{arima}$. The output of the `arima` function is given by

```
arima(y, order = c(1, 0, 0),xreg=index-25,include.mean=T)
```

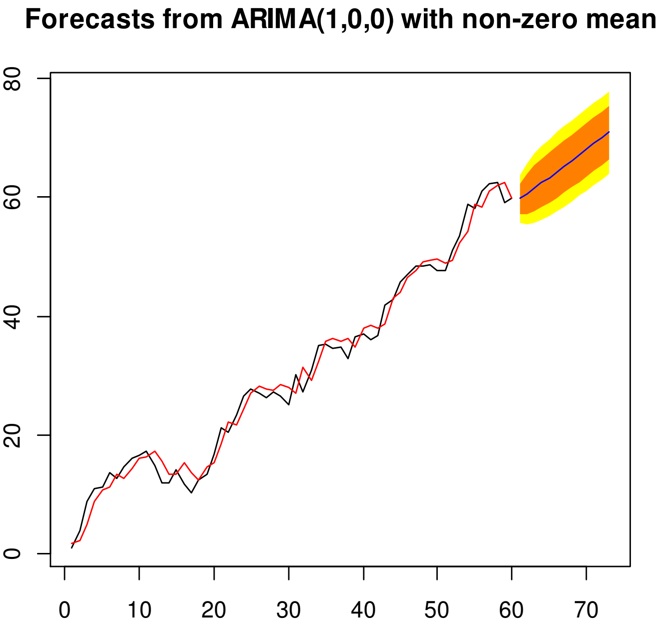
```
         ar1  intercept  index - 25
       0.819     25.395        0.967
s.e.    0.069          1.404
0.068
```

Thus, the 500th model is

$$\begin{cases} Y_t = 25.395 + 0.819\,(t-25) + \varepsilon_t \\ \varepsilon_t = 0.819\varepsilon_{t-1} + u_t, \quad u \sim WN \end{cases}.$$

To predict the model 12 steps ahead, use the following lines:

```
windows()
mod.serial = Arima(y,
order = c(1, 0, 0),
xreg=index-25,
include.mean=T)
plot(forecast(mod.serial,h=
12,xreg=(60:72)-25))
lines(1:60,
fitted(mod.serial),col=2)
```



To analyze the model for serial correlation

- create OLS model and visually examine its residuals for persistency or/and apply DW test or/and Breusch-Godfrey test
- correct the model with FGLS: use the CORC or arima or gls procedures
- alternatively, use the HAC estimators

3.17 exercise. The data file morta.txt consists of 508 lines (observations) and three columns (these are weekly averages of respective variables):

`mort`	daily mortality in Los Angeles County
`temp`	temperature
`part`	particulate pollution

```
morta=read.table(file.choose(),header=TRUE) # go to morta.txt
head(morta)
matplot(morta,type="l",lty=1,col=1:3)       # plot all the series
windows()                                   # open new graph window
plot(morta)                       # scatter diagrams; note that mort
                                  # decreases with time
tt=1:dim(morta)[1]
attach(morta)
temp1=temp - mean(temp)           # adjust temperature for its mean to
                                  # avoid scaling problems
temp2=temp1^2
mod1=lm(mort~tt)
summary(mod1);AIC(mod1)
mod2=lm(mort~tt+temp1)
summary(mod2);AIC(mod2)
mod3=lm(mort~tt+temp1+temp2)
```

```
summary(mod3);AIC(mod3)
mod4=lm(mort~tt+temp1+temp2+part)# the scatter diagram suggests to
summary(mod4)                    # include temp2 (why?)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.592238   1.102148   74.03  < 2e-16 ***
tt          -0.026844   0.001942  -13.82  < 2e-16 ***
temp1       -0.472469   0.031622  -14.94  < 2e-16 ***
temp2        0.022588   0.002827    7.99 9.26e-15 ***
part         0.255350   0.018857   13.54  < 2e-16 ***

AIC(mod4)  # the OLS model mod4 is the best according to its AIC
[1] 3332.28

plot(mod4$res,type="l");abline(0,0) # some signs of inertia

library(lmtest)       # most tests on linear models are in the lmtest
                      # package
dwtest(mod4)
DW = 1.31             # residuals are not WN; maybe, AR(1) or AR(2)

bgtest(mod4)          # testing for AR(1)

        Breusch-Godfrey test for serial correlation of order up to 1
data:  mod4
LM test = 63.4323, df = 1, p-value = 1.66e-15

bgtest(mod4,order=2)  # testing for AR(2); both tests reject WN

        Breusch-Godfrey test for serial correlation of order up to 2
data:  mod4
LM test = 127.086, df = 2, p-value < 2.2e-16
```

We omit accurate proof of the fact that `mod4`'s residuals make AR(2); the relevant FGLS model is obtained with

```
mod.arima = arima(mort,order=c(2,0,0),xreg=cbind(tt,temp1,temp2,part))
mod.arima
# Always follow the rule: if the top term in polinomial regression is sig-
# nificant, do not remove any lower term

Coefficients:
         ar1     ar2  intercept       tt    temp1   temp2    part
      0.3848  0.4326    87.6338  -0.0292  -0.0190  0.0154  0.1545
s.e.  0.0436  0.0400     2.7848   0.0081   0.0495  0.0020  0.0272

AIC=3114.07
```

The corrected model has a smaller AIC and quite different[8] coefficients (when compared with mod4).

```
plot(mort[400:508],type="l",ylab="origial and fitted",lwd=2)
lines(fitted(mod4)[400:508],col=2)
lines(fitted(mod.arima)[400:508],col=3)
legend(80,105,c("mort","mort.4","mort.arima"),lty=1,col=1:3)
```

[8] Just because two estimates have the same expected values does not mean that they will be identical.

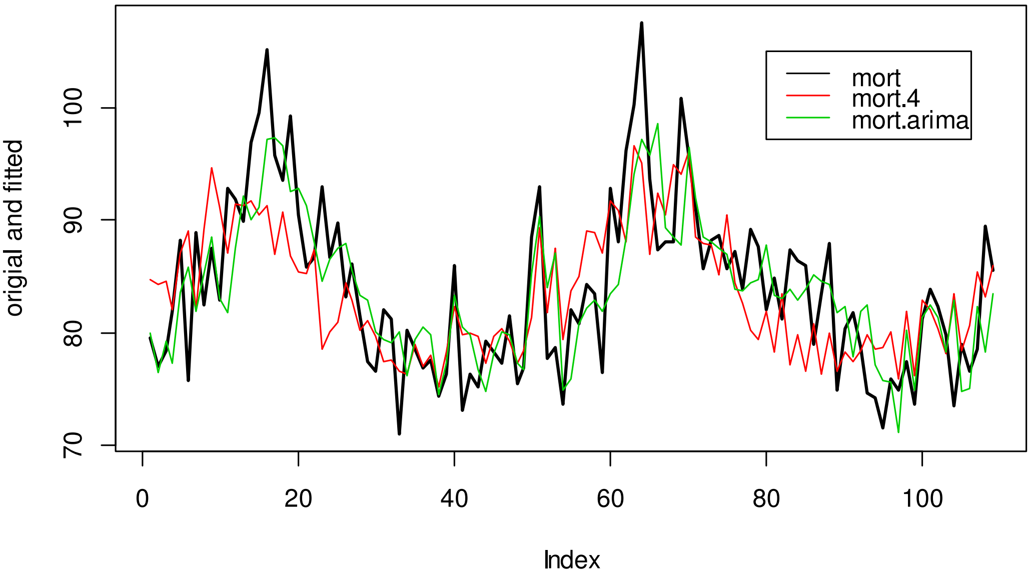


Figure 3.9. `mod.arima` is a better fit than `mod.4`

Another approach is to take, instead of `arima`, a more universal approach and use the FGLS method implemented as the `gls` function in the nlme package.

```
library(nlme)
mod.gls <- gls(mort~tt+temp1+temp2+part, cor = corARMA(p=2))
summary(mod.gls)
```

```
       AIC      BIC    logLik
  3140.488 3174.253 -1562.244

Correlation Structure: ARMA(2,0)
 Formula: ~1
 Parameter estimate(s):
     Phi1      Phi2
0.3939043 0.4381177

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 87.86181 2.9143056 30.148454  0.0000
tt          -0.02918 0.0088414 -3.300541  0.0010
temp1       -0.00970 0.0432201 -0.224484  0.8225
temp2        0.01537 0.0020206  7.606690  0.0000
part         0.15014 0.0249122  6.026867  0.0000
```

Note that `mod2.gls`, described by

$$\begin{cases} mort_t = (\beta_0 + \beta_1 t + \beta_2 temp1_t + \beta_3\, temp2_t + \beta_4 part_t + \varepsilon_t =) \\ \qquad\qquad 87.86 - 0.03t - 0.01\, temp1_t + 0.02\, temp2_t + 0.15\, part_t + \varepsilon_t \\ \varepsilon_t = (\rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t =) \\ \qquad 0.397\varepsilon_{t-1} + 0.438\varepsilon_{t-2} + u_t \end{cases}$$

is almost identical to the `mod.arima`.

Some caution is needed when interpreting the model: `mod2.gls` can be expressed as

$$\begin{aligned} mort_t = (1-\rho_1-\rho_2)\beta_0 + (\rho_1+2\rho_2)\beta_1 + \beta_1(1-\rho_1-\rho_2)t + \\ \beta_2 temp1_t + (\beta_2\rho_1)temp1_{t-1} + (\beta_2\rho_2)temp1_{t-1} + \\ \beta_3 temp2_t + (\beta_3\rho_1)\,temp2_{t-1} + (\beta_3\rho_2)\,temp2_{t-2} + \\ \beta_4 part_t + (\beta_4\rho_1)\,part_{t-1} + (\beta_4\rho_2)\,part_{t-2} + u_t \end{aligned}$$

thus, say, $\beta_4(=0.14)$ means that if the pollution $part_t$ increases by 1 and other variables remain the same (this is the *ceteris paribus* condition), then $mort_t$ increases by 0.15 (the words *ceteris paribus* have different meaning in the first equation of the system above and in the latest model where we can also speak about the isolated effect of $part_{t-1}$ (or even $part_{t-2}$) on $mort_t$).

Note that in any case we cannot extend `mort` forecast into future; we could do this only if we knew future `temp` and `part`. ◀◀

3.18 exercise. Redo this example with GRETL. ◀◀

There are a number of reasons why GLS should not be applied every time that the Durbin-Watson test indicates the likelihood of serial correlation in the residuals of an equation. When autocorrelation is detected, the cause may be an omitted variable or a poor choice of functional form (for example, the use of X instead of $\log X$). In case of uncorrelated omitted variables or improper functional form, it can be shown that OLS is superior to GLS for estimating an incorrectly specified equation. The Newey-West technique directly adjusts the standard errors to take account of serial correlationwithout changing the $\hat{\beta}s$ themselves in any way.

3.19 exercise. [redacted] Consider the annual consumption of chicken in the United States, 1951-1994 (the data is placed in chick6.txt):

`Y`	per capita chicken consumption (in pounds)
`PC`	the price of chicken (in cents per pound)
`PB`	the price of beef (in cents per pound)
`YD`	US per capita disposable income (in hundreds of dollars)

1. Create the OLS model $Y = \beta_0 + \beta_1 PC + \beta_2 PB + \beta_3 YD + \varepsilon$.
2. What conclusion can you do about serial correlation on the basis of the DW statistics?
3. Assume that your residuals follow AR(1). Use any two of the FGLS procedures to reesti mate the above model.
4. Plot in one graph $Y, \hat{Y}^{OLS}$, and $\hat{Y}^{GLS}$. Comment.

3.20 exercise. [redacted] The R package car contains a data set `Hartnagel`.

1. Comment on the data.
2. Plot the all-pairs scatter diagrams.
3. Create the OLS model `fconvict ~ tfr + partic + degrees + mconvict`.
4. Plot `fconvict` together with the fitted value. Plot the residuals. Estimate DW statistics. Do the residuals make a white noise?
5. Assume that residuals follow an AR(2) process and create any relevant FGLS model. Plot in one graph `fconvict` and its OLS and FGLS estimates. ◄◄

3.5. Regression Model Specification tests

The RESET (Regression Specification Error Test) test is a popular diagnostic for correctness of functional form. The basic assumption is that under the alternative the model can be written in the form $Y = \vec{X}\cdot\vec{\beta}+\vec{Z}\cdot\vec{\gamma}+\varepsilon$ where $\vec{Z}$ is generated by taking the second or third powers either of the fitted response, the regressor variables, or the first principal component of $\vec{X}$. A standard $F-$test (or its *LM* version) is then applied to determine whether these additional variables have significant influence. In R, we can use the `resettest` function from the lmtest package or `durbinWatsonTest` in car package where it is implemented through a bootstrap approach. We shall redo here the 4.6 example from the Lecture Notes.

3.7 example. Consider the hprice.txt dataset with 88 observations. In LN, 4.3 example, we found that the model $price = \beta_0 + \beta_1 bdrms + \beta_2 lotsize + \beta_3 sqrft + \varepsilon$ is heteroskedastic which was, probably, detected because of the misspecification of the model (the model, probably, lacks quadratic terms).

```
head(hprice)
    price assess bdrms lotsize sqrft colonial   lprice  lassess llotsize   lsqrft
1 300.000  349.1     4    6126  2438        1 5.703783 5.855359 8.720297 7.798934
2 370.000  351.5     3    9903  2076        1 5.913503 5.862210 9.200593 7.638198
3 191.000  217.7     3    5200  1374        0 5.252274 5.383118 8.556414 7.225482
4 195.000  231.8     3    4600  1448        1 5.273000 5.445875 8.433811 7.277938
5 373.000  319.1     4    6095  2514        1 5.921578 5.765504 8.715224 7.829630
6 466.275  414.5     5    8566  2754        1 6.144775 6.027073 9.055556 7.920810
```

```
attach(hprice)
mod1=lm(price~bdrms+lotsize+sqrft+colonial)
summary(mod1)
library(MASS)
library(help=MASS)
?stepAIC
mod2=stepAIC(mod1)    #in a stepwise manner we are looking for min-AIC model
summary(mod2)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
bdrms        1.385e+01  9.010e+00   1.537  0.12795
lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***

Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6607
```

```
plot(data.frame(mod2$res^2,bdrms,lotsize,sqrft))

# we want to remove two observations with biggest mod2$res^2
N=2                                # detect 2 biggest elements
tail(order(mod2$res^2), N)         # their numbers are 42,76
hprice2=hprice[-c(42,76),]
detach(hprice)
attach(hprice2)
mod3=lm(price~bdrms+lotsize+sqrft+colonial)
summary(mod3)
mod4=stepAIC(mod3)
summary(mod4)                      # another three variables

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.8850279 21.7761655  -0.041  0.96768
lotsize      0.0017888  0.0005504   3.250  0.00168 **
sqrft        0.1273091  0.0098452  12.931  < 2e-16 ***
colonial    24.4281525 11.8917003   2.054  0.04314 *

Multiple R-squared:  0.7085,    Adjusted R-squared:  0.6979

plot(data.frame(mod4$res^2,lotsize,sqrft,colonial))
library(lmtest)
?resettest
resettest(mod4)                    # do we need square terms?

RESET = 3.7725, df1 = 2, df2 = 80, p-value = 0.02719 # we need square terms

mod3sq=lm(price~(bdrms+lotsize+sqrft+colonial)^2+I(bdrms^2)+I(lotsize^2)+
I(sqrft^2))
summary(mod3sq)
mod4sq=stepAIC(mod3sq)
summary(mod4sq)              # model with smallest AIC

resettest(mod4sq)            # no need for new square terms
shapiro.test(mod4sq$res)     # errors are normal, mod4sq is the best model
```

3.21 exercise. Create the model $lprice = \beta_0 + \beta_1 bdrms + \beta_2 colonial + \beta_3 llotsize + \beta_4 lsqrft + \varepsilon$, test with `resettest`, simplify, if necessary, and test for heteroskedasticity and normality of errors.

3.22 exercise. Analyze the following three examples:

```
x <- c(1:30)
y1 <- 1 + x + x^2 + rnorm(30)
y2 <- 1 + x + rnorm(30)
resettest(y1 ~ x , power=2, type="regressor")
resettest(y2 ~ x , power=2, type="regressor")

?growthofmoney
modelHetzel <- TG1.TG0 ~ AG0.TG0
lm(modelHetzel, data=growthofmoney)
dwtest(modelHetzel, data=growthofmoney)
resettest(modelHetzel, data=growthofmoney)
resettest(modelHetzel, power=2, type="regressor", data=growthofmoney)
resettest(modelHetzel, type="princomp", data=growthofmoney)
```

Principal component analysis (PCA) has a number of different interpretations. The simplest is a projection method finding projections of maximal variability. That is, it seeks linear combinations of the columns of the design matrix X with maximal (or minimal) variance. The first k principal components span a subspace containing the 'best' k dimensional view of the K-dimensional, $K > k$, data. It best approximates the original points in the sense of minimizing the sum of squared distances from the points to their projections. The first few principal components are often useful to reveal structure in the data.

```
?princomp
data(stackloss)
?stackloss
summary(lm.stack <- lm(stack.loss ~ .,data=stackloss)) # model with all
                                                       # variables
(pc.cl  <- princomp(stackloss))
names(pc.cl)
pc.cl$load
pc.cl$load[,1]
summary(pc.cl)
predict(pc.cl)
# now the model with only the first principal component - it is even better
summary(lm.princ <- lm(stack.loss~predict(pc.cl)[,2],data=stackloss)
```

3.6. Instrumental Variables

If any X_m in the linear regression model $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$ correlates with ε, i.e., $\mathrm{cov}(X_m, \varepsilon) = EX_m\varepsilon \neq 0$, then all the OLS estimators $\hat{\beta}_i^{OLS}$ are biased and inconsistent. When faced with such a situation, we must consider alternative estimation procedures.

3.8 example. We shall repeat 4.7 example from Lecture Notes with R.

```
library(Ecdat)
data(Mroz)
head(Mroz)
modOLS=lm(log(hearnw)~educw+experience+I(experience^2),data=Mroz,
subset=work=="no")
summary(modOLS)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.5220406  0.1986321  -2.628  0.00890 **
educw             0.1074896  0.0141465   7.598 1.94e-13 ***
experience        0.0415665  0.0131752   3.155  0.00172 **
I(experience^2) -0.0008112  0.0003932  -2.063  0.03974 *
```

We shall use instrumental variables to correctly estimate the influence of `educw` on `log(hearnw)`. A mother's education `educwm` does not belong to the daughter's wage equation, and it is reasonable to propose that more educated mothers are more likely to have more educated daughters.

Stage 1.

```
attach(Mroz)
Mroz.no=Mroz[work=="no",]
mod.s1 = lm(educw~experience+I(experience^2)+educwm,data=Mroz.no)
summary(mod.s1)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       9.775103   0.423889  23.061   <2e-16 ***
experience        0.048862   0.041669   1.173    0.242    
I(experience^2) -0.001281   0.001245  -1.029    0.304    
educwm            0.267691   0.031130   8.599   <2e-16 ***
```

Stage 2.

```
mod.s2 = lm(log(hearnw)~mod.s1$fit+experience+I(experience^2),data=Mroz.no)
summary(mod.s2)
```

```
Coefficients:
                  Estimate  Std. Error t value Pr(>|t|)   
(Intercept)       0.1981861  0.4933427   0.402  0.68809   
mod.s1$fit        0.0492630  0.0390562   1.261  0.20788   
experience        0.0448558  0.0141644   3.167  0.00165 **
I(experience^2) -0.0009221  0.0004240  -2.175  0.03019 * 
```

Note that the coefficient 0.049 is in fact the *IV* coefficient for `educw` (not `mod.s1$fit`). Also, both stages can be performed in one step with the `ivreg` function:

```
library(AER)
mod2SLS = ivreg(log(hearnw) ~ educw + experience + 
I(experience^2) | experience + I(experience^2) + educwm, data = Mroz.no)
summary(mod2SLS)
```

```
Coefficients:
                  Estimate  Std. Error t value Pr(>|t|)   
(Intercept)       0.1981861  0.4728772   0.419  0.67535   
educw             0.0492630  0.0374360   1.316  0.18891   
experience        0.0448558  0.0135768   3.304  0.00103 **
I(experience^2) -0.0009221  0.0004064  -2.269  0.02377 * 
```

◀◀

3.9 example. [redacted] We are interested in the demand elasticity of cigarettes. One tool in the quest for reducing illnesses and deaths from smoking – and the costs, or externalities, imposed by those illnesses on the rest of society – is to tax cigarettes so heavily that current smokers cut back and potential new smokers are discouraged from taking up the habit. But precisely how big a tax hike is needed to make a dent in cigarette consumption? For example, what would the after tax sales pricc of cigarettes need to be to achieve a 20% reduction in cigarette consumption?

The answer to this question depends on the elasticity of demand for cigarettes. If the elasticity is -1, then the 20% target in consumption can be achieved by a 20% increase in price. If the elasticity is -0.5, then the price must rise 40% to decrease consumption by 20%. Of course, we do not know what the demand elasticity of cigarettes is in the abstract: we must estimate it from data on prices and sales. But, because of the interactions between supply and demand, the elasticity of demand for cigarettes cannot be estimated consistently by an OL S regression of log quantity on log price, i.e., the variable $\log(\texttt{price})$ is endogeneous in $\log(\texttt{packs}) = \beta_0 + \beta_1 \log(\texttt{price}) + \varepsilon$.

We therefore use 2SLS to estimate the elasticity of demand for cigarettes using annual data for the 48 continental U.S. states for 1985-1995 (see `CigarettesSW` in package AER):

`state`	Factor indicating state.
`year`	Factor indicating year
`cpi`	Consumer price index
`population`	State population.
`packs`	Number of packs per capita.
`income`	State personal income (total, nominal)
`tax`	Average state, federal and average local excise taxes for fiscal year
`price`	Average price during fiscal year, including sales tax
`taxs`	Average excise taxes for fiscal year, including sales tax

```
   state year   cpi population     packs   income      tax     price      taxs
1     AL 1985 1.076    3973000 116.48628  46014968 32.50000 102.18167  33.34834
2     AR 1985 1.076    2327000 128.53459  26210736 37.00000 101.47500  37.00000
3     AZ 1985 1.076    3184000 104.52261  43956936 31.00000 108.57875  36.17042
.......................................................................................
47    WV 1985 1.076    1907000 112.84740  20852964 33.00000 108.91125  38.18625
48    WY 1985 1.076     500000 129.39999   7116756 24.00000  93.46667  24.00000

49    AL 1995 1.524    4262731 101.08543  83903280 40.50000 158.37134  41.90467
50    AR 1995 1.524    2480121 111.04297  45995496 55.50000 175.54251  63.85917
51    AZ 1995 1.524    4306908  71.95417  88870496 65.33333 198.60750  74.79082
.......................................................................................
95    WV 1995 1.524    1820560 115.56883  32611268 41.00000 166.51718  50.42550
96    WY 1995 1.524     478447 112.23814  10293195 36.00000 158.54166  36.00000
```

The data set consists of annual data for 48 continental U.S. states for the years 1985 and 1995. Quantity consumed is measured by annual per capita cigarette sales in packs per fiscal year as derived from state tax collection data. The `price` is the real (that is, inflation-adjusted) average retail cigarette price per pack during the fiscal year, including taxes. `income` is real per capita income. The general sales `tax` is the average tax, in cents per pack, due to the broad-based state sales tax applied to all consumption goods. The cigarette-specific tax, `taxs`, is the tax applied to cigarettes only. All prices, income, and taxes used in the regression in this example are deflated by the Consumer Price Index and thus are in constant (real) dollars.

The instrumental variable `tdiff` (see below) is the portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack (in real dollars, deflated by the Consumer Price Index). Cigarette consumption, `packs`, is the number of packs of cigarettes sold per capita in the state, and the `price` is the average real price per pack of cigarettes including all taxes.

Before using TSLS it is essential to ask whether the two conditions for instrument `tdiff` validity hold. First consider instrument relevance. Because a high sales tax increases the total sales `price`, the sales tax per pack plausibly satisfies the condition for instrument relevance.

Next consider instrument exogeneity. For the sales tax to be exogenous, it must be uncorrelated with the error in the demand equation; that is, the sales tax must affect the demand for cigarettes only indirectly through the price.This seems plausible: general sales tax rates vary from state to state, but they do so mainly because different states choose different mixes of sales, income, property, and other taxes to finance public undertakings. Those choices about public finance are driven by political considerations, not by factors related to the demand for cigarettes.

```
library(AER)
data("CigarettesSW")
CigarettesSW$rprice = with(CigarettesSW, price/cpi)       # real price
CigarettesSW$rincome = with(CigarettesSW, 
income/population/cpi)                                    # real income per
                                                          # capita
CigarettesSW$tdiff = with(CigarettesSW, (taxs - tax)/cpi) # IV tdiff
c1995 = subset(CigarettesSW, year == "1995")

## convenience function: HC1 covariances
## vcovHC is a heteroskedasticity-consistent estimation of the covariance
## matrix of the coefficient estimates in regression models
## HC1 is a refinement of the White estimator
hc1 = function(x) vcovHC(x, type = "HC1")
fm_ivreg <- ivreg(log(packs) ~ log(rprice) | tdiff, data = c1995)
coeftest(fm_ivreg, vcov = hc1)
```

```
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  9.71988    1.52832  6.3598 8.346e-08 ***
log(rprice) -1.08359    0.31892 -3.3977  0.001411 ** 
```

This 2SLS cigarette demand model with heteroskedasticity-robust standard errors (we use instrument `tdiff` for endogenous `log(rprice)`) is surprisingly elastic: an increase in the price of 1% reduces consumption by 1.08%. But do not take this estimate too seriously – there still might be omitted variables that are correlated with the sales tax per pack. A leading candidate is income (it is an exogenous variable, therefore we include it into the list of IV).

```
fm_ivreg2 <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + 
tdiff, data = c1995)
coeftest(fm_ivreg2, vcov = hc1)
```

```
             Estimate Std. Error t value  Pr(>|t|)    
(Intercept)   9.43066    1.25939  7.4883 1.935e-09 ***
log(rprice)  -1.14338    0.37230 -3.0711  0.003611 ** 
log(rincome)  0.21452    0.31175  0.6881  0.494917    
```

Here the dependent variable is `log(packs)`, the endogenous regressor is `log(rprice)`, the included exogenous variable is `log(rincome)`, and the instrument is `log(rincome)`(now the elasticity equals to 1.14).

This regression uses a single instrument `rincome` but, in fact, another candidate instrument is available, the cigarette specific taxes (they increase the price of cigarettes paid by the consumer, so it meets the condition for instrument relevance; if it is uncorrelated with the error term, it is an exogenous instrument).

```
fm_ivreg3 <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + 
tdiff + I(tax/cpi), data = c1995)
```

```
coeftest(fm_ivreg3, vcov = hc1)

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)    9.89496    0.95922 10.3157 1.947e-13 ***
log(rprice)   -1.27742    0.24961 -5.1177 6.211e-06 ***
log(rincome)   0.28040    0.25389  1.1044    0.2753
```

Now the elasticity has risen to 1.28 and the standard errors have diminished by one third, thus the model is quite satisfactory. Note that we will not pursue the matter of the validity of the intruments, GRETL seems to be more adjusted to such an analysis. ◄◄

3.23 exercise. Repeat the modelling with GRETL. Import the data set cigarett.gta as a STATA file.

3.24 exercise. The 500 values of x, y, z_1, and z_2 in ivreg2.txt were generated artificially. The variable $y = \beta_0 + \beta_1 x + \varepsilon = 3 + 1 \cdot x + \varepsilon$. Do this exercise with GRETL and/or R.

(a) The explanatory variable x follows a normal distribution with mean zero and variance $\sigma_x^2 = 2$. The random error ε is normally distributed with mean zero and variance $\sigma_\varepsilon^2 = 1$. The covariance between x and ε is 0.9. Using the algebraic definition of correlation, determine the correlation between x and ε.
(b) Given the values of y and x, and the values of $\beta_0 = 3$ and $\beta_1 = 1$, solve for the values of the random disturbances ε. Find the sample correlation between x and ε and compare it to your answer in (a).
(c) In the same graph, plot the value of y against x, and the regression function $E(y \mid x) = 3 + 1 \cdot x$. Note that the data do not fall randomly about the regression function.
(d) Estimate the regression model $y = \beta_0 + \beta_1 x + \varepsilon$ by OLS using a sample consisting of the first N = 10 observations on y and x. Repeat using N = 20, N = 100, and N = 500. What do you observe about the least squares estimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not?
(e) The variables z_1 and z_2 were constructed to have normal distributions with means zero and variances one, and to be correlated with x but uncorrelated with ε. Using the full set of 500 observations, find the sample correlations between z_1, z_2, x, and ε. Will z_1 and z_2 make good instrumental variables? Why? Is one better than the other? Why?
(f) Estimate the model $y = \beta_0 + \beta_1 x + \varepsilon$ by instrumental variables using a sample consisting of the first N = 10 observations and the instrument z_1. Repeat using N = 20; N = 100, and N = 500. What do you observe about the IVestimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not?
(g) Estimate the model $y = \beta_0 + \beta_1 x + \varepsilon$ by instrumental variables using a sample consisting of the first N = 10 observations and the instrument z_2. Repeat using N = 20; N = 100, and N = 500. What do you observe about the IVestimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not? Comparing the results using z_1 alone to those using z_2 alone, which instrument leads to more precise estimation? Explain.

(h) Estimate the model $y = \beta_0 + \beta_1 x + \varepsilon$ by instrumental variables using a sample consisting of the first N = 10 observations and the instruments z_1 and z_2. Repeat using N = 20, N = 100, and N = 500. What do you observe about the IV estimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not? Is estimation more precise using two instruments than one, as in parts (f) and (g)?

3.25 exercise. During the 1880s, a cartel known as the Joint Executive Committee (JEC) controlled the rail transport of grain from the Midwest to eastern cities in the United States.The cartel preceded the Sherman Antitrust Act of 1890 and it legally operated to increase the price of grain above what would have been the competitive price. From time to time, cheating by members of the cartel brought about a temporary collapse of the collusive price-setting agreement. In this exercise, you will use variations in supply associated with the cartel's collapses to estimate the elasticity of demand for rail transport of grain.

The data is presented in the JEC.dta file in Stata format (import it with GRETL), its description is given below. Suppose that the demand curve for rail transport of grain is specified as

$$\log(Q_i) = \beta_0 + \beta_1 \log(P_i) + \beta_2 ice_i + \sum\nolimits_{j=1}^{12} \beta_{2+j} seas_{j,i} + \varepsilon_i$$

where $\mathtt{Q}_i$ is the total tonnage of grain shipped in week *i*, $\mathtt{P}_i$ is the price of shipping a ton of grain by rail, $\mathtt{ice}_i$ is a binary variable that is equal to 1 if the Great Lakes are not navigable because of ice, and `seas` is a binary variable that captures seasonal variation in demand. `ice` is included because grain could also be transported by ship when the Great Lakes were navigable.

Variable Definitions

Variable	Definition
`week`	Week of observation: =1 if 1880.01.01-1880.01.07,=2 if 1880.01.08-1880.01.14,…=328 for final week
`price`	weekly index of price of shipping a ton of grain by rail
`ice`	1 if Great Lakes are impassible because of ice, 0 otherwise
`cartel`	1 if railroad cartel is operative, 0 otherwise
`quantity`	total tonnage of grain shipped in the week

`seas1-seas13`	thirteen “month” binary variables. To match the weekly data, the calendar has been divided into 13 periods, each approximately 4 weeks long. Thus
	seas1=1 if date is January 1 through 28, =0 otherwise
	seas2=1 if date is January 29 through February 25, =0 otherwise
	………
	seas13=1 if date is December 4 through December 31, =0 otherwise

a. Estimate the demand equation by OLS. What is the estimated value of the demand elasticity and its standard error?
b. Explain why the interaction of supply and demand could make the OLS estimator of the elasticity biased.
c. Consider using the variable `cartel` as instrumental variable for log(`P`). Use economic reasoning to argue whether `cartel` plausibly satisfies the two conditions for a valid instru-

ment.

d. Estimate the first-stage regression. Is `cartel` a weak[9] instrument?

e. Estimate the demand equation by instrumental variable regression. What is the estimated demand elasticity and its standard error?

f. Does the evidence suggest that the cartel was charging the profit-maximizing monopoly price? Explain. (*Hint:* What should a monopolist do if the price elasticity is less than 1?)

3.10 example. We repeat example from Lecture Notes, p. 4-49, and correct it by introducing IV.

```
library(MASS); set.seed(2);N=100;ro=0.7
par(mfrow=c(1,2))
### OLS
Sigma=matrix(c(3^2,3*1*ro,1*3*ro,1^2),2,2);Sigma
Xeps=mvrnorm(N,c(0,0),Sigma)
X=Xeps[,1]; eps=Xeps[,2] # Endogenous X correlates with eps
Y=2+0.3*X+eps  # DGP
plot(X,Y)
mod=lm(Y~X);summary(mod)
abline(2,0.3); abline(mod,lty=2)
legend(-6.5,6,c("true","OLS estimate"),lty=c(1,2))
### IV
set.seed(2)
roXeps=ro
roXZ=0.6
roZeps=0 # Instrumental variable Z correlates with X, but not with eps
Sigma2=matrix(c(3^2,3*1*roXZ,3*0.5*roZeps,1*3*roXZ,1^2,1*0.5*roXeps,
0.5*3*roZeps,0.5*1*roXeps, 0.5^2),3,3);Sigma2
ZXeps=mvrnorm(N,c(0,0,0),Sigma2)
Z=ZXeps[,1];X=ZXeps[,2];eps=ZXeps[,3]

cor.test(Z,eps)
cor.test(X,eps)
cor.test(Z,X)
### 2SLS
### Step 1
mod2=lm(X~Z)
Xfit=fitted(mod2)
Y=2+0.3*X+eps; plot(X,Y)
mod=lm(Y~X);summary(mod)
abline(2,0.3); abline(mod,lty=2)
### Step 2
modIV=lm(Y~Xfit);abline(modIV,col=2)
legend(-2.2,3.6,  c("true","OLS estimate","IV estimate"),  lty  =  c(1,2,1),
col = c(1,1,2))
# in a synonymous manner:
library(AER)
modIV.2 <- ivreg(Y ~ X | Z)
summary(modIV.2)
```

[9] Instruments that explain little of the variation in endogenous X are called weak instruments. If the instruments are weak, then the normal distribution provides a poor approximation to the sampling distribution of the 2SLS estimator and 2SLS is no longer reliable. One way to check for weak instruments when there is a single endogenous regressor is to compute the $F-$statistic testing the hypothesis that the coefficients of the instruments are all zero in the first stage regression of 2SLS. You do not need to worry about the weakness of the instruments if the first stage $F-$statistic exceeds 10. [redacted]

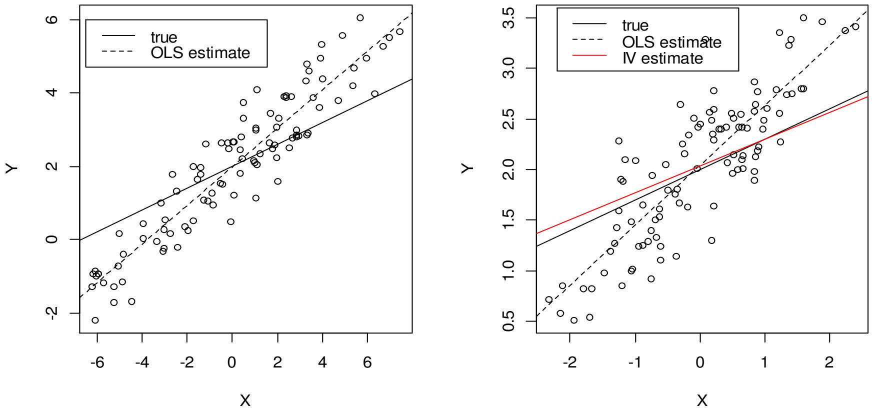


Figure 3.10. OLS estimate and IV estimate of the regression line

3.26 exercise. The data set in prodfun.txt contains log-levels of output (LNY), capital (LNK), and labor (LNL) for two different industrial sectors (SECTOR = 4 or 9) across 83 firms from a developing country (India). Each observation describes the output and inputs of a single firm, thus this is a cross-section dataset. Assume that each sector's output can be described by the Cobb-Douglas production function, i.e., either $Y_i^{(4)} = C^{(4)} K_i^{\alpha 4} L_i^{\beta 4} \exp(\varepsilon_i^{(4)})$ or $Y_i^{(9)} = C^{(9)} K_i^{\alpha 9} L_i^{\beta 9} \exp(\varepsilon_i^{(9)})$. By taking logarithms,

A. Estimate the unknown parameters for each sector using multiple regression models with observations on $\log(Y_i)$, $\log(K_i)$, and $\log(L_i)$.

B. Evaluate the goodness of fit of each sector's fitted production function. Be sure to comment on whether the behavior of the fitted residuals indicates any concerns about the validity of the Cobb-Douglas production function.

C. Test for constant returns to scale (that is, $H_0 : \alpha + \beta = 1$) in each of the estimated production functions. What is the meaning of this property?

D. Test the null hypothesis that the production functions are identical across the two sectors, that is, test $H_0 : C^{(4)} = C^{(9)}, \alpha 4 = \alpha 9, \beta 4 = \beta 9$. (*Hint*: dummify SECTOR and create the OLS model $\log(Y) = c + \alpha \log(K) + \beta \log(L) + \varepsilon$ with all the observations, use the GRETL's `chow` function or the Tests section of the OLS regression window with the dummy variable `DSECTOR_1`). Be sure to state explicitly how the Chow test works.

To do the same in R, use the following script (analyze and run it):

```
prodfun=read.table(file.choose(),header=TRUE)
head(prodfun)
attach(prodfun)
plot(prodfun, pch=SECTOR+6,col=SECTOR-3)
```

```
mod4=lm(LNY~LNK+LNL,subset=SECTOR==4)
summary(mod4)

mod9=lm(LNY~LNK+LNL,subset=SECTOR==9)
summary(mod9)

mod=lm(LNY~LNK+LNL)
summary(mod)      # mod is a model for pooled observations

mod.int=lm(LNY~fS+LNK*fS+LNL*fS,data=prodfun)
summary(mod.int)  # mod.int is a model with interactions

# mod.int is the same as:
fS=as.numeric(factor(SECTOR))-1     # dummify SECTOR
fS
summary(lm(LNY~fS+LNK+I(LNK*fS)+LNL+I(LNL*fS)))


anova(mod,mod.int)
# chow test - tests whether all interaction variables are insignificant
# p-value is 0.0268, the same as in GRETL (reject H0)
```

4. Discrete Response Models

Let *Y* takes on only two values: 1 (=succes) and 0 (=failure). We want to create a model

$$P(Y=1)=F(\beta_0+\beta_1X_1+...+\beta_kX_k)$$

where *F* is any distribution function (usually it is either logistic distribution function Λ (logit regression) or normal d.f. Φ (probit regression)). We use the maximum likelihood method to estimate the coefficients.

We begin by reproducing Fig. 5.1 from LN where the model

$$P(coke=1)=\beta_0+\beta_1 pratio$$

and its extensions are estimated.

4.1 example. The R code for Figure 5.1 in LN:

```
ccoke=read.table(file.choose(),header=T)    # navigate to coke.txt
head(ccoke)
COKE=ccoke[order(ccoke$pratio),]            # sort lines by pratio
head(COKE)
attach(COKE)

par(mfrow=c(1,2))
# compare logit and probit
xx=seq(-2,2,by=0.01)
plot(xx,pnorm(xx),type="l",ylab="F",xlab="x",lwd=2,
main="Logistic and normal cdf's")
lines(xx,exp(xx)/(exp(xx)+1),col=2,lwd=2)
abline(0.5,0,lty=2)
legend(-1.8,0.9,c("normal","logistic"),lty=1,col=1:2,lwd=2)

# other coke models
plot(jitter(pratio,amount=0.2),jitter(coke,amount=0.05),ylab="coke",
xlab="pratio",main="Binary response variable")

# linear
COKE.lm=lm(coke~pratio)
abline(COKE.lm,lwd=2,col=3)                 # green line

# nls, no weights
COKE.nls1=nls(coke ~ exp(a + b*pratio)/(1+exp(a+b*pratio)),
start = list(a = 1, b = -1))
summary(COKE.nls1)
lines(pratio,predict(COKE.nls1),lwd=2)

# nls, with weights
Cpred=predict(COKE.nls1)
www=1/(Cpred*(1-Cpred))
COKE.nls2=nls(coke ~ exp(a + b*pratio)/(1+exp(a+b*pratio)), weight=www,
start = list(a = 1, b = -1))
```

```
summary(COKE.nls2)
lines(pratio,predict(COKE.nls2),lwd=2,lty=2)

# logit
xxx=seq(min(pratio)-0.2,max(pratio)+0.15,by=0.01)
COKE.logit=glm(coke~pratio,family=binomial(link="logit"),data = COKE)
summary(COKE.logit)
new=data.frame(pratio=xxx)
lines(xxx,predict(COKE.logit,type="response",newdata=new),lwd=2,col=2)
abline(1,0,col=1)
abline(0,0,col=1)
abline(0.5,0,lty=2)                        # 0.5 threshold line
ppp = tapply(pratio,coke,mean)             # mean value of pratio in two
                                           # coke groups (red squares)
points(ppp[2],1,pch=15,col=2,cex=2)
points(ppp[1],0,pch=15,col=2,cex=2)
yyy=exp(2.5251-2.7108*xxx)
lines(xxx,yyy/(yyy+1),col=3)

legend(1.7,0.93,c("linear","nls","nls-w","logit"),
col=c(3,1,1,2),lty=c(1,1,2,1),lwd=2)
```

A simplified GRETL version of the above code is given below:

```
dataset sortby pratio                      # sort by pratio
series coke_j = coke + 0.02*normal()       # jitter
series pratio_j = pratio + 0.05*normal()   # jitter
gnuplot coke_j pratio_j --output=display   # scatter diagram
# linear probability model
ols coke 0 pratio
series cokeOLS = $yhat
gnuplot coke_j cokeOLS pratio --output=display --with-lines=cokeOLS
# logit model
logit coke 0 pratio
series cokeLOGIT = $yhat
gnuplot coke_j cokeLOGIT pratio --output=display --with-lines=cokeLOGIT
```

```
OLS, using observations 1-1140
Dependent variable: coke

             coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const        1.02035     0.0519576     19.64    3.58e-074 ***
  pratio      -0.557783    0.0487203    -11.45    8.41e-029 ***

Log-likelihood      -758.9132   Akaike criterion     1521.826
Schwarz criterion    1531.904   Hannan-Quinn         1525.632

***************************

Logit, using observations 1-1140
Dependent variable: coke
Standard errors based on Hessian

             coefficient   std. error      z        slope
  ----------------------------------------------------------
  const        2.52508      0.271574      9.298
  pratio      -2.71081      0.266631    -10.17    -0.666411
```

```
McFadden R-squared    0.082656   Adjusted R-squared   0.080105
Log-likelihood       -719.0694   Akaike criterion     1442.139
Schwarz criterion     1452.216   Hannan-Quinn         1445.945

Number of cases 'correctly predicted' = 755 (66.2%)
f(beta'x) at mean of independent vars = 0.246
Likelihood ratio test: Chi-square(1) = 129.582 [0.0000]

             Predicted
                 0     1
  Actual 0   508   122
         1   263   247
```

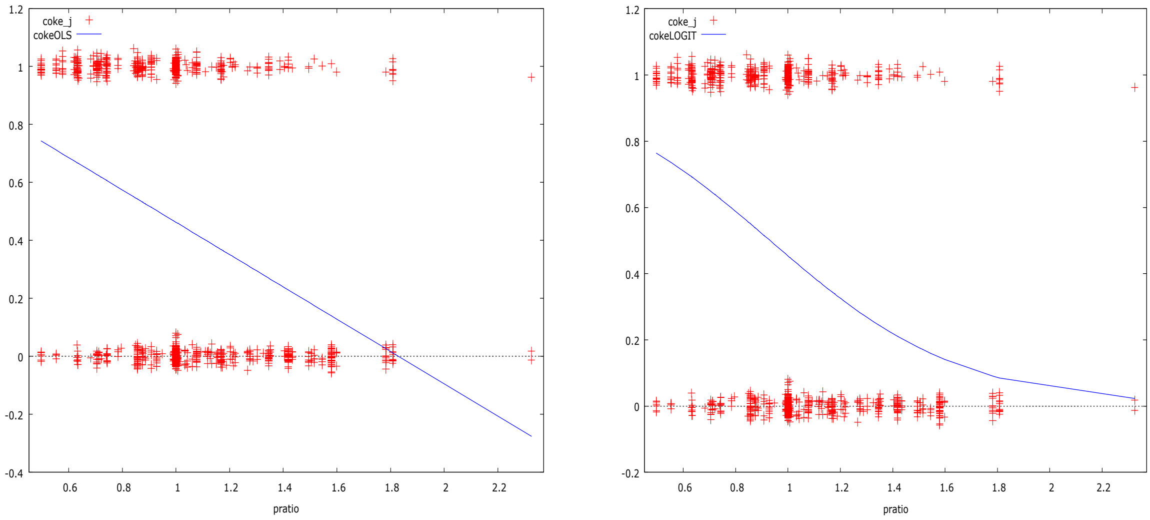


Figure 4.1. Linear probability model (left) and logit model (right)

4.1 exercise. Repeat the example through GUI.

4.2 example. Išduodami paskolas bankai visuomet rizikuoja. Šiame pavyzdyje bandysime ištirti nuo ko priklauso paskolos negrąžinimo (kredito defolto) tikimybė. Dalis mums reikalingų duomenų yra R duomenų rinkinyje `credit` (žr. paketą `Fahrmeir`). Deja, ten yra ne visi mums reikalingi stulpeliai, todėl duomenis, surinktus viename pietų Vokietijos banke, teks importuoti iš originaliojo rinkinio kredit.txt (kintamųjų aprašas yra faile kredit.var.html).

```
credit=read.table(file.choose(),header=TRUE) # 1000 eilučių, 21 stulpelis
head(credit)
```

```
  kredit laufkont laufzeit moral verw hoehe sparkont beszeit rate famges buerge
1      1        1       18     4    2  1049        1       2    4      2      1
2      1        1        9     4    0  2799        1       3    2      3      1
3      1        2       12     2    9   841        2       4    2      2      1
4      1        1       12     4    0  2122        1       3    3      3      1
5      1        1       12     4    0  2171        1       3    4      3      1
6      1        1       10     4    0  2241        1       2    1      3      1
```

```
  wohnzeit verm alter weitkred wohn bishkred beruf pers telef gastarb
1        4    2    21        3    1        1     3    1     1       1
2        2    1    36        3    1        2     3    2     1       1
3        4    1    23        3    1        1     2    1     1       1
4        2    1    39        3    1        2     2    2     1       2
5        4    2    38        1    2        2     2    1     1       2
6        3    1    48        3    1        2     2    2     1       2
```

Mums rūpimas binarinis atsako kintamasis yra kredito defoltas `y` (`=1-kredit`): jis =0, jei paskola buvo grąžinta laiku, ir =1, jei klientas paskolos negrąžino).

```
attach(credit)
y=1-kredit
table(y)
y
  0   1
700 300   # 700 grąžino, 300 negrąžino
```

Pradėsime paprastu logitiniu modeliu, aprašančiu defolto priklausomybę nuo amžiaus (vok. `alter`). Lygties[1] `E(y|alter)=P(y=1|alter)=exp(c(1)+c(2)alter)/(1+exp(c1)+c(2)alter)` koeficientus apskaičiuosime su `glm` funkcija:

```
c.log = glm(y ~ alter, family=binomial(link="logit"),data=credit)
summary(c.log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.198456   0.233333  -0.851   0.3950      # =c(1)
alter       -0.018512   0.006449  -2.870   0.0041 **   # =c(2)

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1213.1  on 998  degrees of freedom
AIC: 1217.1

Number of Fisher Scoring iterations: 4
```

Kadangi `alter` koeficientas yra neigiamas, o ryšio (logitinė) funkcija ir jos atvirkštinė[2] monotoniškai didėja, todėl didėjant amžiui defolto tikimybė mažėja (žr. 4.2 pav.). Kairėje šio paveikslo pusėje išbrėžtos dvi stulpelinės diagramos – čia geltona spalva žymime defoltų dažnius, o raudona – grąžintų kreditų dažnius. Matyti, kad defolto tikimybė (ji lygi geltono stulpelio aukščio santykiui su abiejų stulpelių aukščių suma) su amžiumi mažėja. Šis mažėjimas gal ir nėra labai akivaizdus, tačiau logitinės regresijos kreivė dešinėje patvirtina mūsų teiginį.

```
par(mfrow=c(1,2))
barplot(table(y,alter),xlab="alter",col=c(2,7))
```

[1] Ją ekvivalenčiai galima užrašyti taip: `logit(P(y=1|alter))=c(1)+c(2)alter`.

[2] Lygties dešinėje užrašyta logistinė skirstinio funkcija.

```
plot(alter,predict(c.log,type="response"),ylab="P(y=1|alter)")
```

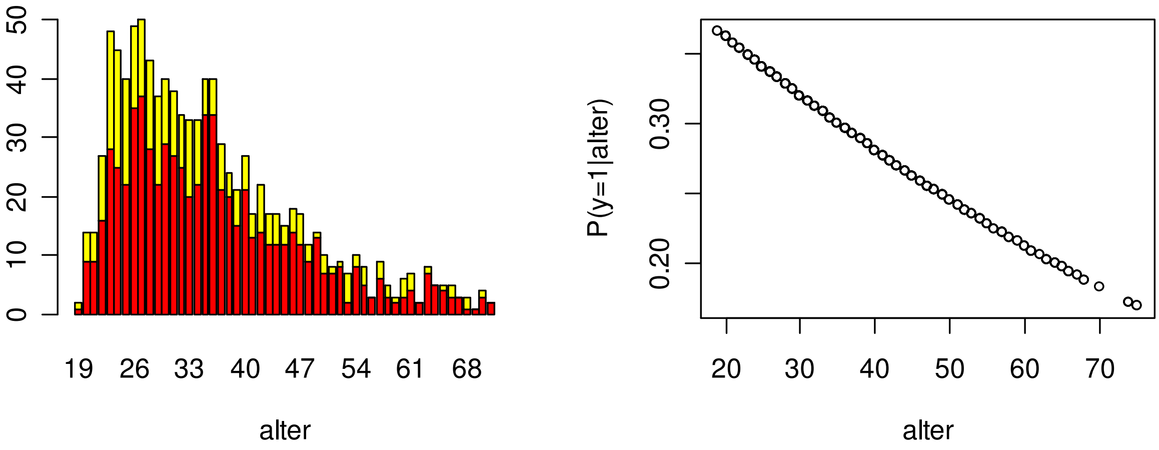


Figure 4.2. Grąžintų ir negrąžintų kreditų dažniai (atitinkamai raudoni ir geltoni stulpeliai) (kairėje) ir logitinės regresijos kreivė (dešinėje) - darome išvadą, kad defolto tikimybė su amžiumi mažėja

Modelio `c.log` lentelėje matyti, kad koeficientas prie `alter` yra reikšmingas (jo *p* reikšmė 0.0041). Bendresnis būdas patikrinti `alter` reikšmingumą yra palyginti šį modelį su kitu, sudarytu tik iš konstantos. Tam GLM modeliuose paprastai vartojamas liekanų kvadratų sumos analogas, vadinamoji (normuotoji) deviacija (angl. (scaled) deviance). Jei turime du įdėtuosius modelius, tai jų deviacijų skirtumas yra dydis, ekvivalentus tikėtinumo funkcijų santykiui. Šis skirtumas turi χ^2 skirstinį su laisvės laipsnių skaičiumi, lygiu kintamųjų skaičiaus šiuose modeliuose skirtumui.

```
anova(c.log)
```

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: y
Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev
NULL                    999    1221.73
alter  1     8.61       998    1213.11
```

```
1-pchisq(8.61,1)
```

```
[1] 0.003343223        # <0.05
```

Taigi modelis su `alter` yra akivaizdžiai pranašesnis už modelį vien iš konstantos.

Modeliai su skirtingomis ryšio funkcijomis negali būti tiesiogiai palyginti dėl skirtingų mastelių (pvz., standartinės logistinės skirstinio funkcijos dispersija lygi $\pi^2/3$, o normaliosios – 1; modelius galėsime palyginti, jei vieną kurią nors ryšio funkciją atitinkamai normuosime). 11.8 pav. išbrėžta standartinė logistinė kreivė (atvirkštinė logitinio ryšio funkcija) ir $\Phi(x;0,\pi^2/3)$ grafikas (abi kreivės sunkiai atskiriamos).

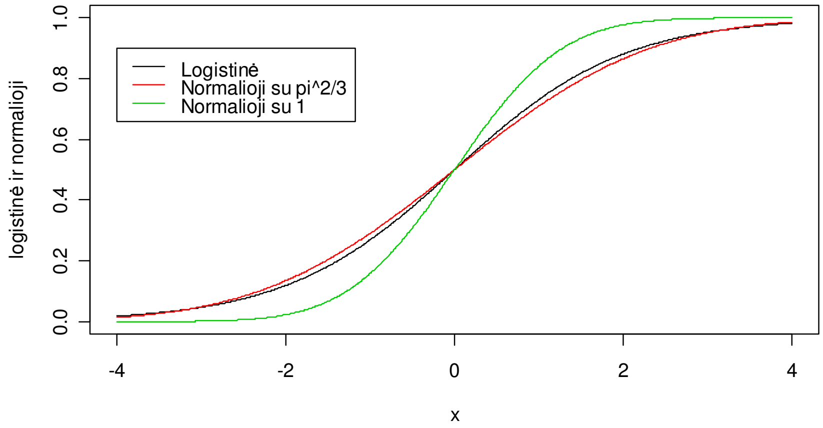


Figure 4.3. Logistinė ir dvi normaliosios (su dispersijomis $\pi^2/3$ ir 1) skirstinio funkcijos

Taigi, norint palyginti logitinį ir probitinį modelius, pastarojo koeficientus reikia padauginti iš $\pi/\sqrt{3}$. Žemiau pateiktoje lentelėje matyti, kad koeficientai prie `alter` mažai skiriasi (o prie laisvojo nario – abu nereikšmingi).

```
c.pro= glm(y ~ alter, family=binomial(link="probit"),data=credit)
```

```
             c.pro.orig c.pro.rescaled  c.log.orig
(Intercept) -0.14245301    -0.25838117 -0.19845642
alter       -0.01085904    -0.01969613 -0.01851229
```

Ar galima patikslinti aptartą modelį, įtraukiant netiesinius `alter` narius? Čia galimi du iš principo skirtingi būdai: i) į modelį įtraukti aukštesnius (pvz., kvadratinį) `alter` narius arba ii) `alter` suskaidyti į grupes. Pradėsime kvadratiniu modeliu.

```
c.log2 = glm(y~alter+I(alter^2),family=binomial(link="logit"),data=credit)
summary(c.log2)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2430239  0.6913629   1.798  0.07219 .
alter       -0.0965881  0.0358010  -2.698  0.00698 **
I(alter^2)   0.0009556  0.0004280   2.233  0.02555 *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1208.3  on 997  degrees of freedom
AIC: 1214.3

Number of Fisher Scoring iterations: 4
```

```
anova(c.log,c.log2)      # Lyginame tiesinį ir kvadratinį modelius
```

```
Analysis of Deviance Table

Model 1: y ~ alter
Model 2: y ~ alter + I(alter^2)
```

```
  Resid. Df Resid. Dev  Df Deviance
1       998     1213.1
2       997     1208.3   1       4.8
```

```
1-pchisq(4.8,1)
[1] 0.02845974     #<0.05
```

Taigi tiesinis modelis atmetamas su maždaug 3% reikšmingumu.

Dabar išbandysime kubinį modelį.

```
c.log3=glm(y~alter+I(alter^2)+I(alter^3),family=binomial(link="logit"),
data=credit)
summary(c.log3)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.092e+00  2.145e+00   1.908   0.0564 .
alter       -3.240e-01  1.664e-01  -1.947   0.0515 .
I(alter^2)   6.583e-03  4.056e-03   1.623   0.1046
I(alter^3)  -4.326e-05  3.115e-05  -1.389   0.1649

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1206.3  on 996  degrees of freedom
AIC: 1214.3

Number of Fisher Scoring iterations: 4
```

```
anova(c.log2,c.log3)       # Lyginame kvadratinį ir kubinį modelius

Analysis of Deviance Table

Model 1: y ~ alter + I(alter^2)
Model 2: y ~ alter + I(alter^2) + I(alter^3)
  Resid. Df Resid. Dev  Df Deviance
1       997     1208.3
2       996     1206.3   1       2.0
```

```
1-pchisq(2,1)

[1] 0.1572992    # >0,05
```

Matome, kad kubinis modelis nepagerina modelio tikslumo (žr. taip pat AIC koeficientus), todėl sustosime prie kvadratinio. Atkreipsime dėmesį, kad šį kartą defolto tikimybė jau nėra monotoniška `alter` funkcija (žr. 4.4 pav., kairėje), kas visai natūralu.

```
plot(alter,predict(c.log2,type="response"),ylab="P(y=1|alter)")
```

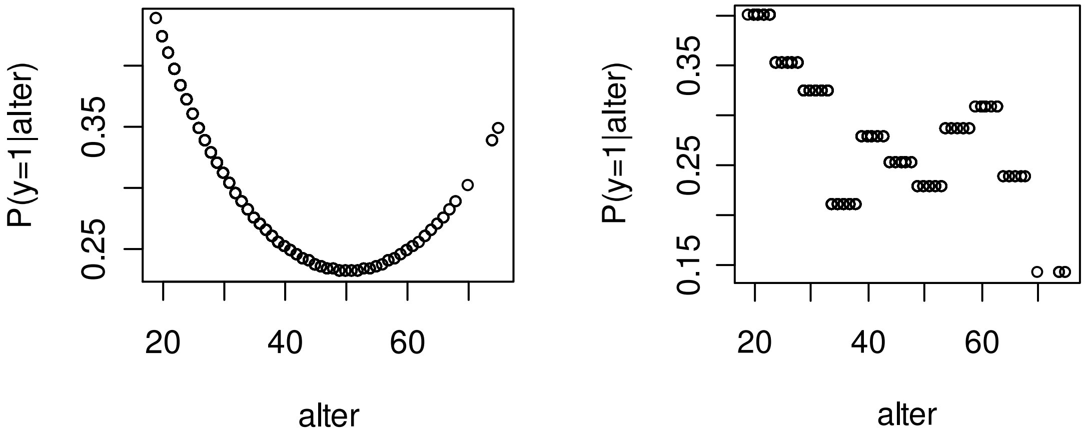


Figure 4.4. Dviejų modelių – `c.log2` ir `c.cut` (žr. žemiau) atsakų grafikai

Kitas būdas įtraukti netiesiškumą yra suskaidyti `alter` į grupes. Pasirinksime tokias grupes: (18, 23], (23,28],...,(68,75] (visos jos, išskyrus paskutinę, vienodo ilgio). Pirmas intervalas bus bazinis, todėl vertinsime tik likusių dešimties intervalų koeficientus.

```
alter.cut=cut(alter,c(seq(18,68,by=5),75))
c.cut=glm(y~alter.cut,family=binomial(link="logit"),data=credit)
summary(c.cut)
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.4055     0.1992  -2.035 0.041809 *
alter.cut(23,28]  -0.2029     0.2429  -0.836 0.403381
alter.cut(28,33]  -0.3292     0.2545  -1.294 0.195828
alter.cut(33,38]  -0.9144     0.2755  -3.319 0.000903 ***
alter.cut(38,43]  -0.5447     0.2958  -1.842 0.065544 .
alter.cut(43,48]  -0.6763     0.3265  -2.071 0.038340 *
alter.cut(48,53]  -0.8076     0.3970  -2.034 0.041943 *
alter.cut(53,58]  -0.5108     0.4239  -1.205 0.228168
alter.cut(58,63]  -0.4055     0.4693  -0.864 0.387595
alter.cut(63,68]  -0.7577     0.5497  -1.378 0.168100
alter.cut(68,75]  -1.3863     1.0983  -1.262 0.206886

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1203.2  on 989  degrees of freedom
AIC: 1225.2

Number of Fisher Scoring iterations: 4
```

```
plot(alter,predict(c.cut,type="response"),ylab="P(y=1|alter)")
```

Pažymėsime, kad `c.cut` deviacija yra pati mažiausia iš kol kas nagrinėtų (tai galima paaiškinti ir tuo, kad šis modelis yra pats lanksčiausias (turi daugiausiai parametrų)). Antra vertus, AIC koeficientas, kuris atsižvelgia ne tik į paklaidų didumą, bet ir į parametrų skaičių, šį kartą pats didžiausias. Nežiūrint to, skaidymas į grupes dažnai padeda atskleisti netiesinius efektus.

Iki šiol nagrinėjome defolto tikimybės priklausomybę tik nuo `alter`. Dabar į modelį įtrauksime ir paskolos dydį `hoehe`. Nagrinėsime keturis logitinius modelius.

- Tiesinis be sąveikos

```
logit(P(y=1|alter,hoehe))=c(1)+c(2)alter+c(3)hoehe:
```

```
c.alt.ho=glm(y~alter+hoehe,family=binomial,data=credit)
```

- Tiesinis su sąveika

```
logit(P(y=1|alter,hoehe))=c(1)+c(2)alter+c(3)hoehe+c(4)alter·hoehe:
```

```
c.alt.ho.int=glm(y~alter*hoehe,family=binomial,data=credit)
```

- Kvadratinis be sąveikos

```
logit(P(y=1|alter,hoehe))=c(1)+c(2)alter+c(3)alter^2+c(4)hoehe+c(5)hoehe^2:
```

```
c.alt2.ho2=glm(y~alter+I(alter^2)+hoehe+I(hoehe^2),family=binomial,data=cre
dit)
```

- Kvadratinis su sąveika

```
logit(P(y=1|alter,hoehe))=c(1)+c(2)alter+c(3)alter^2+c(4)hoehe+c(5)hoehe^2+

+c(6)alter·hoehe:
```

```
c.alt2.ho2.int=glm(y~alter*hoehe+I(alter^2)+I(hoehe^2),family=binomial,data
= credit)
```

Nesunku įsitikinti, kad geriausi[3] modeliai yra du paskutiniai:

```
summary(c.alt2.ho2)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.181e+00  7.085e-01   1.668  0.09539 .
alter       -1.012e-01  3.658e-02  -2.768  0.00564 **
I(alter^2)   9.856e-04  4.378e-04   2.251  0.02436 *
hoehe       -7.289e-06  7.455e-05  -0.098  0.92211
I(hoehe^2)   1.048e-08  5.979e-09   1.753  0.07958 .

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1180.2  on 995  degrees of freedom
AIC: 1190.2
```

```
summary(c.alt2.ho2.int)
```

[3] Jų AIC ir deviacijos mažiausios.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.486e+00  7.393e-01   2.010  0.04438 *
alter       -1.083e-01  3.715e-02  -2.915  0.00355 **
hoehe       -1.178e-04  1.054e-04  -1.117  0.26385
I(alter^2)   9.322e-04  4.441e-04   2.099  0.03579 *
I(hoehe^2)   9.513e-09  5.972e-09   1.593  0.11119
alter:hoehe  3.372e-06  2.172e-06   1.552  0.12055

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1177.7  on 994  degrees of freedom
AIC: 1189.7
```

Kiek geresnis yra pats paskutinis, tačiau LR (=Likelihood Ratio) testas teigia, kad jis nėra reikšmingai geresnis už modelį be sąveikos:

```
anova(c.alt2.ho2,c.alt2.ho2.int)

Analysis of Deviance Table

Model 1: y ~ alter + I(alter^2) + hoehe + I(hoehe^2)
Model 2: y ~ alter * hoehe + I(alter^2) + I(hoehe^2)
  Resid. Df Resid. Dev  Df Deviance
1       995    1180.22
2       994    1177.71   1     2.51

1-pchisq(2.51,1)
[1] 0.1131259
```

Kitaip sakant, iš visų aptartųjų modelių vertėtų rinktis `c.alt2.ho2`. Antra vertus, mes dar nenagrinėjome modelių su visais prognoziniais kintamaisiais. Vartosime šiuos kintamuosius (ranginių kintamųjų skales apibrėžė patyrę banko specialistai: žemas rangas – blogai, aukštas rangas - gerai):

`moral` kliento patikimumas (nustatomas pagal tai, kaip grąžino ankstesnius kreditus) - 0- mažas, ..., 4 – labai didelis

`beszeit` 1 – bedarbis, ..., 5 – toje pačioje vietoje dirba ne mažiau kaip 7 metus (šį kintamąjį perkoduosime: kintamasis `dirba` bus =0, jei klientas nedirba ir =1, jei turi darbą)

`laufzeit` kredito trukmė mėnesiais (kuo ilgesnė, tuo geriau) – šį kintamąjį sudiskretinsime, t.y., jį paversime faktoriaus lygiais

`laufzeit<=9` jei kreditas išduotas ne daugiau kaip 9 mėnesiams (bazinė grupė)

`laufzeit(9,12]` trukmė tarp 9 ir 12 mėnesių

`laufzeit(12,18]` trukmė tarp 12 ir 18 mėnesių

`laufzeit(18,24]` trukmė tarp 18 ir 24 mėnesių

`laufzeit>=24` trukmė ilgesnė nei 24 mėnesiai

`sparkont` santaupos: 1 – santaupų neturi, 5 – didelės santaupos

`verw` paskolos tikslas: 0 – kiti tikslai, 1 – naujas automobilis, 2 – naudotas automobilis, 3 – baldai, ..., 10 – verslas `verm` kliento vertingiausias turtas: 1 – nėra arba nežinomas, 2 – automobilis, 3 – gyvybės draudimas, 4 – namas ar žemės sklypas (šį kintamąjį perkoduosime: kintamasis `namas` bus lygus 1, jei - `verm`=4 ir =0 kitais atvejais)

Pirmiausiai įvesime papildomus kintamuosius.

```
dirba=ifelse(beszeit==1,0,1)
d.laufzeit=cut(laufzeit,c(0,9,12,18,24,80))
namas=ifelse(verm==4,1,0)
```

Nagrinėsime du modelius.

- Modelis su visais kintamaisiais

```
c.visi = glm(y ~ alter*hoehe + I(alter^2) + I(hoehe^2) + moral + dirba +
d.laufzeit + sparkont + verw + namas,family = binomial,data = credit)

summary(c.visi)
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        2.560e+00  8.558e-01   2.991 0.002782 ** 
alter             -8.950e-02  4.005e-02  -2.235 0.025426 *  
hoehe             -2.984e-04  1.267e-04  -2.355 0.018506 *  
I(alter^2)         8.389e-04  4.828e-04   1.738 0.082295 .  
I(hoehe^2)         2.219e-08  7.450e-09   2.978 0.002898 ** 
moral             -4.558e-01  7.433e-02  -6.133 8.65e-10 ***
dirba             -2.794e-01  3.084e-01  -0.906 0.365016    
d.laufzeit(9,12]   5.735e-01  2.896e-01   1.980 0.047699 *  
d.laufzeit(12,18]  9.212e-01  2.975e-01   3.097 0.001956 ** 
d.laufzeit(18,24]  1.043e+00  2.986e-01   3.493 0.000478 ***
d.laufzeit(24,80]  1.625e+00  3.326e-01   4.887 1.02e-06 ***
sparkont          -3.060e-01  5.465e-02  -5.600 2.15e-08 ***
verw              -3.489e-02  2.834e-02  -1.231 0.218324    
namas              6.501e-01  2.140e-01   3.038 0.002384 ** 
alter:hoehe        1.460e-06  2.355e-06   0.620 0.535384    

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1063.1  on 985  degrees of freedom
AIC: 1093.1
```

- Iš modelio pašalinsime tris mažiausiai reikšmingus narius: `dirba`, `verw` ir sąveikos narį `alter:hoehe`.

```
c.visi.fin = glm(y ~ alter + hoehe + I(alter^2) + I(hoehe^2) + moral +
d.laufzeit + sparkont + namas,family = binomial,data = credit)

summary(c.visi.fin)
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        2.119e+00  7.928e-01   2.673 0.007526 ** 
alter             -9.097e-02  3.950e-02  -2.303 0.021296 *  
hoehe             -2.398e-04  9.852e-05  -2.434 0.014953 *  
```

```
I(alter^2)          9.203e-04  4.740e-04   1.942 0.052173 .
I(hoehe^2)          2.215e-08  7.473e-09   2.964 0.003040 **
moral              -4.505e-01  7.368e-02  -6.115 9.66e-10 ***
d.laufzeit(9,12]    5.752e-01  2.890e-01   1.990 0.046545 *
d.laufzeit(12,18]   9.239e-01  2.965e-01   3.116 0.001834 **
d.laufzeit(18,24]   1.020e+00  2.978e-01   3.427 0.000611 ***
d.laufzeit(24,80]   1.560e+00  3.296e-01   4.733 2.21e-06 ***
sparkont           -3.039e-01  5.446e-02  -5.580 2.40e-08 ***
namas               6.854e-01  2.111e-01   3.247 0.001168 **

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1065.9  on 988  degrees of freedom
AIC: 1089.9
```

Galutinis modelis (pagal `Residual deviance`, `AIC` ir koeficientų reikšmingumą) neabejotinai geriausias. Pažymėsime taip pat, kad beveik visi koeficientai turi „teisingus“ ženklus. Pvz., defolto tikimybė sumažėja (nes koeficiento ženklas neigiamas), jei kliento `moral` didesnis arba jo santaupos `sparkont` didesnės. `d.laufzeit` koeficientai (taigi ir defolto tikimybė) didėja kartu su paskolos trukme, kas irgi suprantama. Keistokas koeficiento prie `namas` ženklas – išeitų, kad namo turėjimas padidina defolto tikimybę (antra vertus, tai galima paaiškinti papildomais finansiniais įsipareigojimais).

4.2 exercise. Import to GRETL the cross-sectional data set Davis.txt:

```
Nr  male weight height
1      1     77    182
2      0     58    161
3      0     53    161
4      1     68    177
5      0     59    157
6      1     76    170
........................................
```

Can you, using height and, probably, weight to predict person‘s gender? Use three model – linear, logit with height only, and logit with height and weight. Decorate your report with some graphs. Which model is best? Here is some help:

```
davis.glm1=glm(male~height,family=binomial(link="logit"),data=davis)
summary(davis.glm1)
attach(davis)
plot(height,male)
points(height,predict(davis.glm1,type="response"),col=2)
```

4.3 exercise. Import to GRETL or R the data set CPS5_n.txt where

ED	education (in years, 13 groups from 6 to 18)
SO	region of residence (coded 1 if South, 0 otherwise)
BL	(coded 1 if nonwhite and non-Hispanic, 0 otherwise)
HP	(coded 1 if Hispanic, 0 otherwise)

FE	gender (coded 1 if female, 0 otherwise)
MS	marital status (coded 1 if married, 0 otherwise)
EX	potential labor market experience (in years)
UN	union status (coded 1 if in union job, 0 otherwise)
WG	hourly wage (in dollars)

1) Estimate the OLS model $\texttt{WG} = \beta_0 + \beta_1 \texttt{ED} + \varepsilon$. Plot respective scatter diagram and the regression line.

2) We could improve the model but we take another stand. Call a person poor or "economically disadvantaged" (Y=1) if his or her salary is less than \$5 per hour (they constitute roughly 1/5 of the whole population). Create a logit model $P(\texttt{Y} = 1) = \Lambda(\beta_0 + \beta_1 \texttt{ED})$ and compare the curve with the conditional expectation of Y in every group of ED. How much one extra year of education diminishes the probability of getting to the „poor" group? Compare the classification tables of the OLS and logit models.

3) Create a logit model with the ED, FE, and EX variables on the rhs. What is the estimated probability that a male with 10 years of education and 12 years of experience will find himself in a little paid group? Verify that the same probability for females is much bigger. Assume that a male studied two years more. How, on average, this probability will change?

4) Estimate the probability to be a union member (UN=1) (include EX, FE, EX, EX^2 to your model). What number would you report if asked for an estimate of how much the probability of being in a union job falls per year of additional education? Using the „predict $\texttt{UN} = 1$ if $\widehat{\texttt{UN}} > 1/2$, predict UN=0 otherwise" rule, it turns out that the estimated logit model correctly predicted union status for fully 82% of the individuals in the sample. Are you impressed? Hint. What proportion of the individuals are union members?

References

[A] Adkins Lee C., Using gretl for Principles of Econometrics, http://www.learneconometrics.com/gretl/index.html

[ČM1] Čekanavičius V., Murauskas G. Statistika ir jos taikymai, TEV, Vilnius, 2001,I.

[ČM2] Čekanavičius V., Murauskas G. Statistika ir jos taikymai, TEV, Vilnius, 2002,II.

[K] Kufel T., Ekonometria. Rozwiązywanie problemow z wykorzystaniem programu GRETL, Warszawa : Wydawnictwo Naukowe PWN, 2011