REMIGIJUS LAPINSKAS

Lecture Notes

PRACTICAL ECONOMETRICS. II. TIME SERIES ANALYSIS

PRAKTINĖ EKONOMETRIJA. II. LAIKINĖS SEKOS

remigijus.lapinskas@mif.vu.lt

Vilnius 2013

Contents

0. Introduction

0.1. Preface

A time series (or a random process in discrete time) is a sequence of observations that are arranged according to the time of their outcome. The annual crop yield of sugar-beets and their price per ton for example is recorded in agriculture. The newspapers' business sections report daily stock prices, weekly interest rates, monthly rates of unemployment and annual turnovers. Meteorology records hourly wind speeds, daily maximum and minimum temperatures and annual rainfall. Geophysics is continuously observing the shaking or trembling of the earth in order to predict possibly impending earthquakes. An electroencephalogram traces brain waves made by an electroencephalograph in order to detect a cerebral disease, an electrocardiogram traces heart waves. The social sciences survey annual death and birth rates, the number of accidents in the home and various forms of criminal activities. Parameters in a manufacturing process are permanently monitored in order to carry out an on-line inspection in quality assurance.

There are, obviously, numerous reasons to record and to analyze the data of a time series. Among these is the wish to gain a better understanding of the data generating mechanism, the prediction of future values or the optimal control of a system. The characteristic property of a time series is the fact that the data are not generated independently, their dispersion varies in time, they are often governed by a trend and they have cyclic components. Statistical procedures that suppose independent and identically distributed data are, therefore, excluded from the analysis of time series. This requires proper methods that are summarized under time series analysis.

0.2. Statistical Data and their Models

This section introduces common types of economic data and describes some basic models associated with their use.

- **Cross-sectional data**

Some researchers often work with data that is characterized by individual **units**. These units might refer to companies, people or countries. For instance, a researcher investigating theories relating to portfolio allocation might collect data on the return earned on the stocks of many different companies. With such **cross-sectional** data, the method of ordering the data (for example, by alphabet or size) usually does not matter.

Typically, for the cross-sectional data the notations X_i, Y_i, and like are used to indicate an observation on variables X, Y etc for the ith individual. Observations in a cross-sectional data set run from unit $i = 1$ to N. By convention, N indicates the number of cross-sectional units (e.g., the number of companies surveyed). For instance, a researcher might collect data on the share price of $N = 100$ companies at a certain point in time. In this case, Y_1 will be

equal to the share price of the first company, Y_2 the share price of the second company, and so on.

In many cases, a researcher is interested in establishing a relationship between two or more cross-sectional variables. Figure 1.1 is a graph of data on $Y =$ executive compensation (i.e., the salary paid to the chief executive, expressed in millions of dollars) for 70 companies, along with data on the $X =$ profits of the companies (i.e., profit expressed in millions of dollars; the data is available as a file executive.xls). It is commonly thought that there should be a relationship between these two variables, either because more profitable companies can afford to hire better chief executives (and pay them more) or because executive talent leads to higher profits (note that Figure 1.1 confirms this belief). To digitalize this belief one can use either the correlation coefficient (in our case `corr(Comp,Profit)=0,66`) or create a regression model

$$\texttt{Comp}_i = \alpha + \beta \cdot \texttt{Profit}_i + \varepsilon_i .$$

For the sake of brevity and generality, we rewrite this model as

$$Y_i = \alpha + \beta X_i + \varepsilon_i ,$$

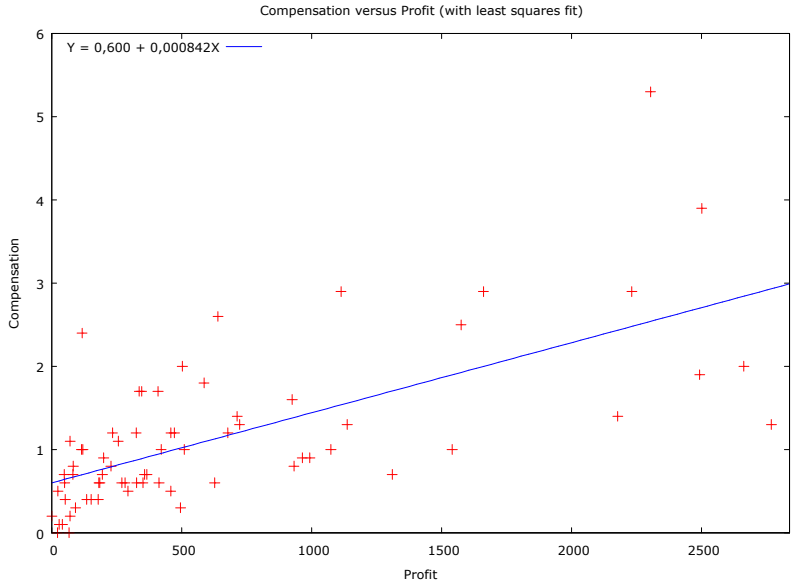


Figure 0.1. *XY*-plot (or *scatter diagram*) of profits against executive compensation

where Y is referred to as the **dependent** variable, X the independent or **explanatory** variable, ε the **error term** of the model, and α and β, **coefficients**. The coefficient β quantifies the X's influence on Y, its **estimate** $b = \hat{\beta}$ (alongside with $a = \hat{\alpha}$) is usually obtained with the help of **ordinary least squares** (OLS) method:

$$(a,b) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} (Y_i - (\alpha + \beta X_i))^2 \qquad (0.1)$$

(in our case $b = 0.00084$, which means that if company's profit increases by 1 (million of dollars), then executive's salary increases on average by 0.00084 (million of dollars)=$840).

```
Model 1: OLS, using observations 1-70
Dependent variable: Comp

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       0,599965      0,112318      5,342    1,15e-06  ***
  Profit      0,000842      0,000117      7,228    5,50e-010 ***

Mean dependent var   1,138571   S.D. dependent var   0,928189
Sum squared resid    33,61789   S.E. of regression   0,703122
R-squared            0,434479   Adjusted R-squared   0,426162
F(1, 68)             52,24308   P-value(F)           5,50e-10
Log-likelihood      -73,65540   Akaike criterion     151,3108
Schwarz criterion    155,8078   Hannan-Quinn         153,0971
```

Recall that here *t*-ratio (or *t*-statistics) is defined as $t = (\hat{\beta} - \beta_0) / std.error$ and is designed to test the hypothesis $H_0 : \beta = \beta_0$. In the above table $\beta_0 = 0$, thus $t = \hat{\beta} / std.error$. If the discrepancy t is "big", more specifically, if *p*-value is less than 0.05, we reject H_0 – this means that X influences *Y*.

Clearly, compensation cannot be explained by the profit alone. In fact, the data set executive.xls contains two more **explanatory** variables: *W*=change in sales and *Z*=change in debt. The extented, **multiple regression**, model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_i + \beta_3 Z_i + \varepsilon_i$$

is usually presented in the form of a regression table:

```
Model 2: OLS, using observations 1-70
Dependent variable: Comp

             coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       0,565674      0,132710      4,262    6,57e-05 ***
  Profit      0,000807743   0,000113558   7,113    1,02e-09 ***
  ChSales    -0,00397477    0,0102140    -0,3891   0,6984
  ChDebt      0,0279660     0,0100031     2,796    0,0068   ***

Mean dependent var   1,138571   S.D. dependent var   0,928189
Sum squared resid    30,00558   S.E. of regression   0,674263
R-squared            0,495245   Adjusted R-squared   0,472302
F(3, 66)             21,58553   P-value(F)           7,45e-10
Log-likelihood      -69,67678   Akaike criterion     147,3536
Schwarz criterion    156,3475   Hannan-Quinn         150,9261

Excluding the constant, p-value was highest for variable 3 (ChSales)
```

The interpretation of regression coefficients is subject to *ceteris paribus* conditions. For instance, $b_j = \hat{\beta}_j$ measures the marginal effect of X_j on *Y*, **holding the other explanatory variables constant**. Thus, the marginal effect of `Profit` remains virtually unchanged.

As the `ChSales` variable is not significant, we rewrite Model 2 in a more parsimonious form:

```
Model 3: OLS, using observations 1-70
Dependent variable: Comp

             coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const      0,536816      0,109359       4,909    6,17e-06  ***
  Profit     0,000814567   0,000111483    7,307    4,27e-010 ***
  ChDebt     0,0269409     0,00958869     2,810    0,0065    ***

Mean dependent var   1,138571   S.D. dependent var   0,928189
Sum squared resid    30,07442   S.E. of regression   0,669979
R-squared            0,494087   Adjusted R-squared   0,478985
F(2, 67)             32,71694   P-value(F)           1,22e-10
Log-likelihood      -69,75699   Akaike criterion     145,5140
Schwarz criterion    152,2595   Hannan-Quinn         148,1934
```

Which of the three models is "best"? The first requirement is "all explanatory variables must be significant" – thus we shall compare only Model 1 and Model 3.

Secondly, most model selection criteria attempt to find the model with the smallest $SSR = \sum_{i=1}^{N} e_i^2$. The criteria we examine fit this general approach; the differences among criteria amount to different penalties for the number of degrees of freedom used in estimating the model (that is, the number of parameters estimated). Because all of the criteria are effectively estimates of the sum of squares of residuals, they have a negative orientation – that is, the smaller the better.

The popular characteristic of the model quality is the coefficient of determination R^2:

$$R^2 = 1 - \frac{SSR}{SST},$$

where $SST = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$ (if, for example, $R^2 = 0.65$, we say that the right-hand variables explain 65% of *Y*'s variability) . The problem with R^2 is that it cannot fall when more explanatory variables are added to a model and therefore `R-squared` of Model 3 (=`0,494087`) could be greater than that of Model 1 (=`0,434479`) just because the Model 3 has more explanatory variables.

There are many possibilities to penalize for extra explanatory variables (*k* is the number of explanatory variables):

- Adjusted R^2: $\bar{R}^2 = 1 - \frac{SSR/(N-k)}{SST/(N-1)}$

- Akaike information criterion : $AIC = \exp(2k/N)\frac{SSR}{N}$

- Schwarz information criterion: $SIC = N^{k/N}\frac{SSR}{N}$

- Hannan-Quinn information criterion $HQC = n\ln\frac{SSR}{n} + 2k\log\log n$

Note that sometimes these criteria (most popular among them are AIC and SIC) give conflicting answers. Always follow the rule: if a few models have the same left-hand variable, the best is with the smallest AIC and/or SIC. In our case, Model 3 is definitely better than Model 1.

- **Time series data**

Financial researchers are often interested in phenomena such as stock prices, interest rates, exchange rates, etc. This data is collected at specific points in time. In all of these examples, the data are ordered by time and are referred to as **time series** data. The underlying phenomenon which we are measuring (e.g., stock prices, interest rates, etc.) is referred to as a **variable**. Time series data can be observed at many **frequencies**. Commonly used frequencies are: **annual** (i.e. a variable is observed every year), **quarterly** (i.e. four times a year), **monthly**, **weekly** or **daily**.

In this course, we will use the notation Y_t to indicate an observation on variable Y (e.g., an exchange rate) at time t. A series of data runs from period $t = 1$ to $t = T$. "T " is used to indicate the total number of time periods covered in a data set. To give an example, if we were to use monthly time series data from January 1947 through October 1996 on the UK pound/US dollar exchange – a period of 598 months – then $t = 1$ would indicate January 1947, $t = 598$ would indicate October 1996 and $T = 598$ the total number of months. Hence, Y_1 would be the pound/dollar exchange rate in January 1947, Y_2 this exchange rate in February 1947, etc. Time series data are typically presented in chronological order.

One objective of analysing economic data is to predict or forecast the future values of economic variables. One approach to do this is to build a more or less structural (for example, regression) econometric model, describing the relationship between the variable of interest with other economic quantities, to estimate this model using a sample of data, and to use it as the basis for forecasting and inference. Although this approach has the advantage of giving economic content to one's predictions, it is not always very useful. For example, it may be possible to adequately model the contemporaneous relationship between unemployment and the inflation rate, but as long as we cannot predict future inflation rates we are also unable to forecast future unemployment.

In the first part of this course (Ch. 1-3) we follow a different route: a pure time series approach. In this approach the current values of an economic variable are related to past values. The emphasis is purely on making use of the information in past values of a variable for forecasting its future. In addition to producing forecasts, time series models also produce the distribution of future values, conditional upon the past, and can thus be used to evaluate the likelihood of certain events. ◀◀

The most interesting results in econometrics during the last 20-30 years were obtained in the intersection of cross-sectional and time series methods. In the second part of this course (Ch. 4-7) we shall get to know different variants of

- **regression models for time series**

Another possibility to combine the two above-mentioned methods is to deal with the so-called

- **panel data**

A data set containing observations on multiple phenomena observed over multiple time periods is called panel data. Panel data aggregates all the individuals, and analyzes them in a period of time. Whereas time series and cross-sectional data are both one-dimensional, panel data sets are two-dimensional.

person	year	income	age	sex
1	2003	1500	27	1
1	2004	1700	28	1
1	2005	2000	29	1
2	2003	2100	41	2
2	2004	2100	42	2
2	2005	2200	43	2

In the above example, a data set with panel structure is shown. Individual characteristics (income, age, sex) are collected for different persons and different years. Two persons (1, 2) are observed over three years (2003, 2004, 2005). Because each person is observed every year, the data set is called a panel.

0.3. Software

There are many statistical software programs. Broadly speaking, they can be divided into commercial (SAS, SPSS, EViews,...) and free (R, GRETL) software; on the other hand, according to the way the procedures are performed, they can be divided into menu-driven and programable (R). The latter two groups nowadays have shown a tendency to combine – for example, EViews, gretl and the commercial S-Plus, all allow to program your steps or perform them from the toolbar. This course is accompanied by computer labs where statistical procedures will be usually performed with GRETL or R (necessary packages are described in http://cran.r-project.org/web/views/TimeSeries.html).

****

The rest of this course is devoted to the analysis of time series data.

1. Time Series: Examples

Time series are analysed to understand the past and to predict the future, enabling managers or policy makers to make properly informed decisions. A time series analysis quantifies the main features in data and the random variation. These reasons, combined with improved computing power, have made time series methods widely applicable in government, industry, and commerce.

1.1 example. Stock returns.

Let P_t be the price of an asset at time t. The one-period (simple) return is

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

Consider monthly returns on Citigroup stock from 1990:01 through 1998:12 (to input the data, open gretl, go to File| Open data| Sample file...| Ramanathan| data9-13| right-click on `cret` and choose Time series plot).

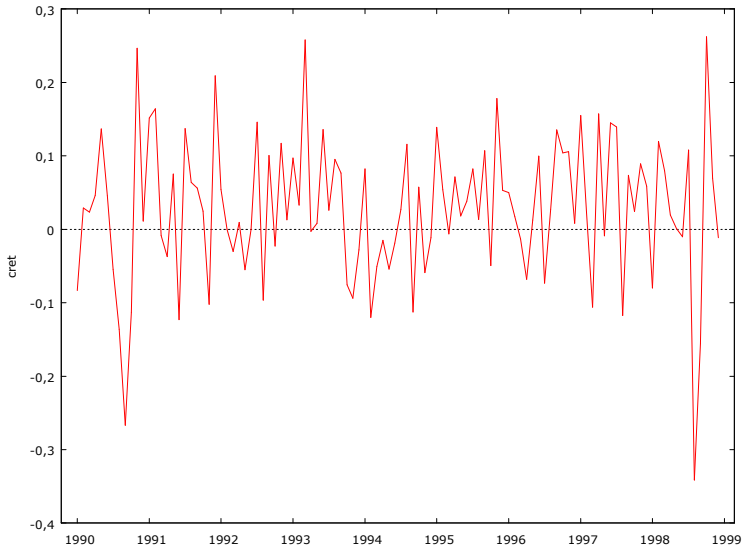


Figure 1.1. Monthly returns on Citigroup stock.

The returns oscillate rather regularly around some constant (which is greater than zero – this means that the returns are generally positive). It is a very simple time series, the best future forecast is probably just this constant. One of the main objectives of this course is to learn how to forecast time series.

1.2 example. Air passenger bookings.

The number of international passenger bookings (in thousands) per month on an airline (Pan Am) in the United States were obtained from the Federal Aviation Administration for the period 1949:1–1960:12 (this classic Box & Jenkins airline data is available as `airlines` file from the fma package of the R program; it can also be found as AP.gdt or AP.txt in the PEdata folder accompanying this course). The company used the data to create a model which was used to predict future demand before ordering new aircraft and training aircrew.

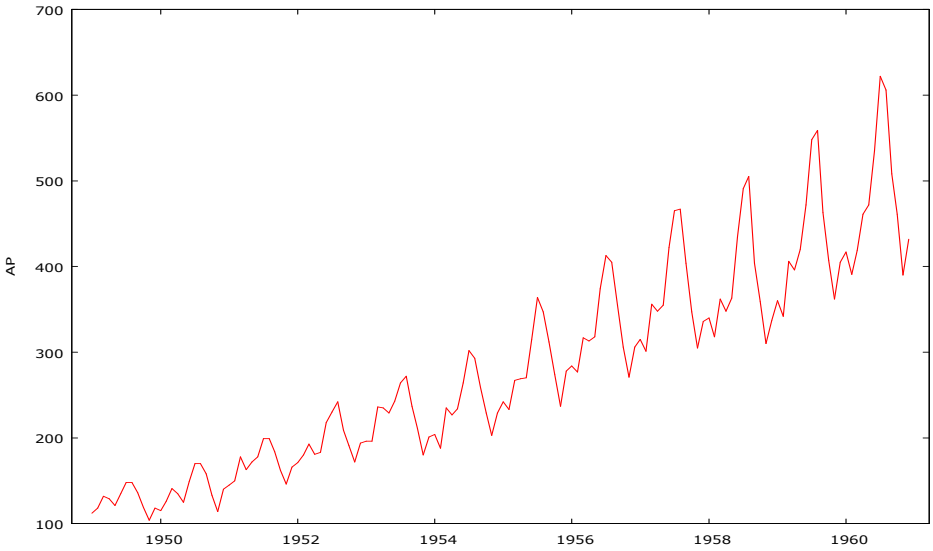


Figure 1.2. International air passenger bookings in the United States for the period 1949-1960.

There are a number of features in the time plot of the air passenger data that are common to many time series. For example, it is apparent that the number of passengers travelling on the airline is increasing with time. In general, a systematic and deterministic change in a time series that does not appear to be periodic is known as a **trend**. The simplest model for a trend is a linear increase or decrease, and this is often an adequate approximation.

A repeating pattern within each year is known as **seasonal** variation, although the term is applied more generally to repeating patterns within any fixed period, such as restaurant bookings on different days of the week. There is clear seasonal variation in the air passenger time series. At the time, bookings were highest during the summer months of June, July, and August and lowest during the autumn month of November and winter month of February.

1.3 example. Sales of shampoo over a three year period

The data available as `shampoo` in the fma package is a monthly sales of shampoo for the period 1:1 – 3:12.

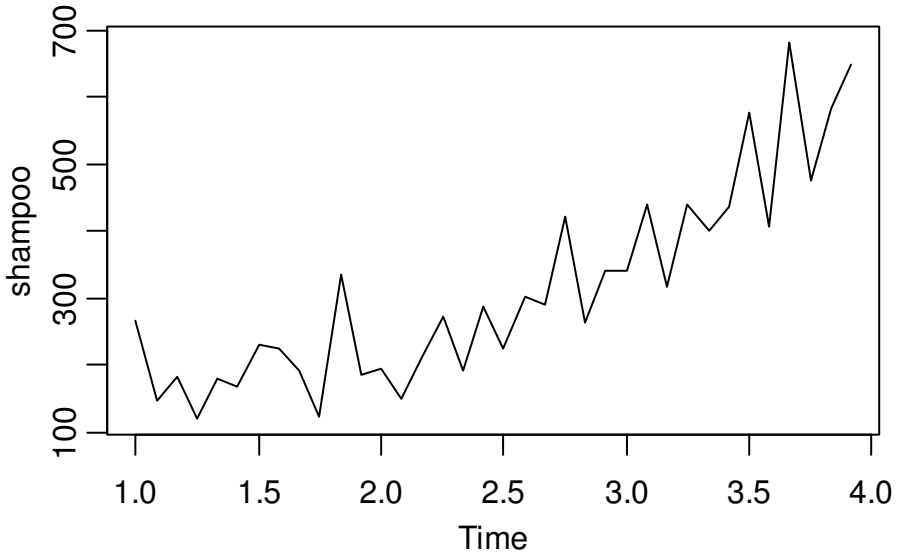


Figure 1.3. Sales of shampoo over a three-year period.

The series has no seasonal component, it is consisting only of a trend and an irregular component.

1.4 example. Quarterly exchange rate: GBP to NZ dollar

The trends and seasonal patterns in the previous two examples were clear from the plots. In addition, reasonable explanations could be put forward for the possible causes of these features. With financial data, exchange rates for example, such marked patterns are less likely to be seen, and different methods of analysis are usually required. A financial series may sometimes show a dramatic change that has a clear cause, such as a war or natural disaster. Day-to-day changes are more difficult to explain because the underlying causes are complex and impossible to isolate, and it will often be unrealistic to assume any deterministic component in the time series model.

The quarterly exchange rates for British pounds sterling to New Zealand dollars for the period 1991:1 to 2000:3 are shown in Fig. 1.4. The data (available as pounds_nz.dat) are mean values taken over quarterly periods of three months.

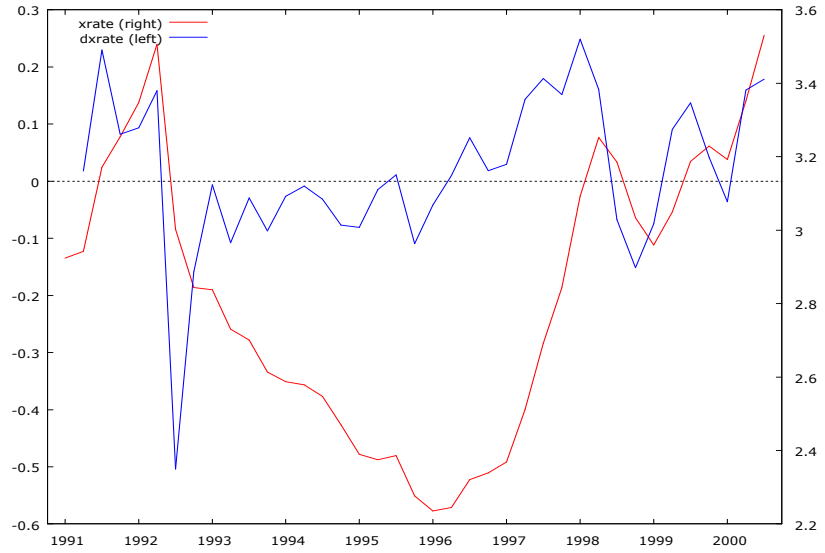


Figure 1.4. Quarterly exchange rates `xrate` for the period 1991–2000 (red); differences `diff(xrate)` (blue) seem to be more regular and, possibly, stationary

The trend seems to change direction at unpredictable times rather than displaying the relatively consistent pattern of the air passenger series. Such trends have been termed *stochastic trends* to emphasise this randomness and to distinguish them from more deterministic trends like those seen in the previous examples. A mathematical model known as a random walk or DS time series can sometimes provide a good fit to data like these and is discussed in Ch. 4.

One of the main applications of the time series theory is prediction. If we reconsider our examples, it seems plausible that we should use the mean to forecast stock returns; to forecast sales of shampoo use its trend, air passenger bookings should be forecasted with its trend plus seasonal component and it is not quite clear at the moment how to forecast stochastic trend of the quarterly exchange rate. We will learn soon how to realize this program.

2. Stationary Time Series

2.1. White Noise - 1

All time series may be divided into two big classes – (covariance or weak) stationary and nonstationary. We shall postpone the exact definitions to subsequent sections, but for the moment - a process which randomly but rather regularly (with more or less constant spread) fluctuates around its constant mean is called stationary. Three examples of such series can be seen in Fig. 2.1.

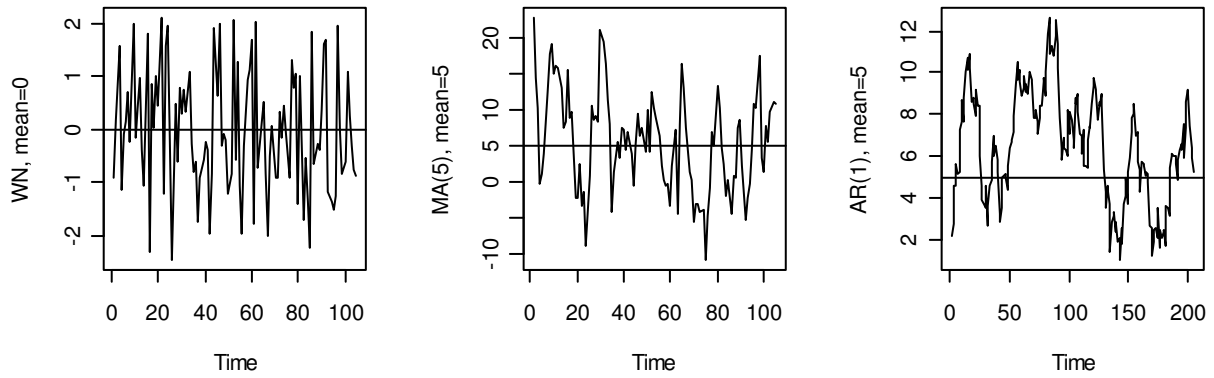


Figure 2.1. Three examples of stationary series; note that the third process (right-most) reverts to its mean much slowlier than the previous two.

In Fig. 2.2 you can see four examples of nonstationary time series.

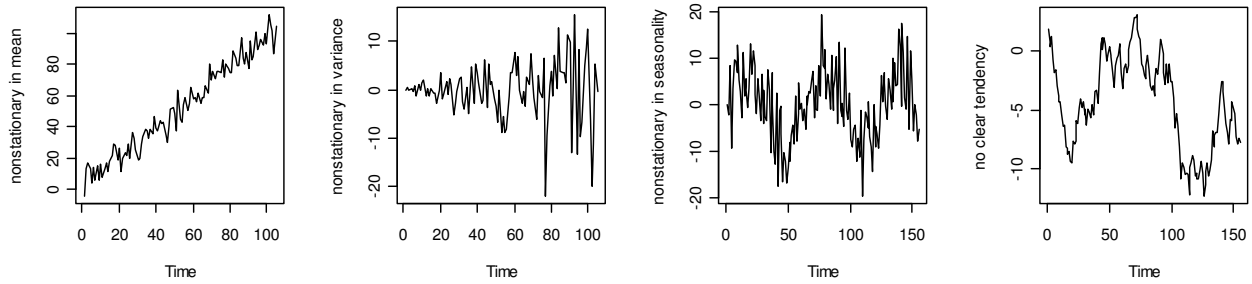


Figure 2.2. All four time series in this figure are not „rather regularly fluctuating around its constant mean"; these time series are not stationary

The simplest **stationary** random process, which is at the same time the main building block of all other stationary series, is the so-called **white noise** – this is a sequence of uncorrelated random variables with zero mean and constant variance (its graph is plotted in Fig. 2.1, left; note that the graph of the stock return, see Fig. 1.1, is quite similar to it). However, how can we know that the other two graphs there are not of the WN? Two functions, ACF (autocorrelation function) and PACF (partial autocorrelation function), come to our rescue: if all the bars (except the zeroth in ACF) are within the blue band, the stationary process is WN (see Fig. 2.3).

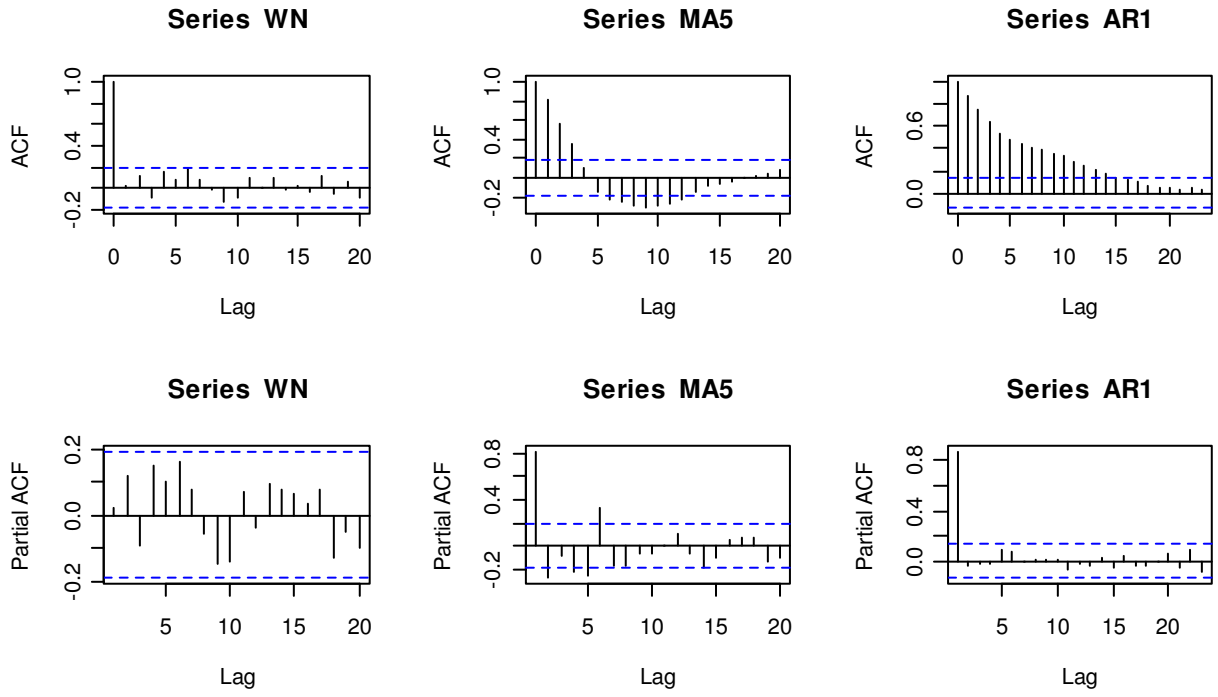


Figure 2.3. The time series WN is a white noise, while the other two are not.

To decide whether the time series is stationary, examine its graph.
To decide whether a stationary time series is white noise, examine its ACF and PACF[1].

Remark. Note that when observing or measuring time series we obtain numbers $y_1, y_2, ..., y_T$. On the other hand, these numbers are just one realization of random variables $Y_1, Y_2, ..., Y_T$. In the sequel, we shall follow a common practice and in both cases use the same upper-case letter notation (it will be clear from the context what do we mean). If we mean random variables, we can speak about the probability distribution of these random variables, their expectations EY_t, variances DY_t etc. Using these probabilistic concepts, we can now give a more precise definition of stationary series: if $EY_t \equiv const$ (i.e., does not depend on t), the process is called mean-stationary; if $DY_t \equiv \sigma^2$, the process is called variance-stationary; if both conditions hold, the process is called (weakly-) stationary[2].

[1] This is only the first step which must be followed by the Ljung-Box test (see p. 2-9).
[2] In fact, the definition of stationarity is still a bit more complicated, see p. 2-3.

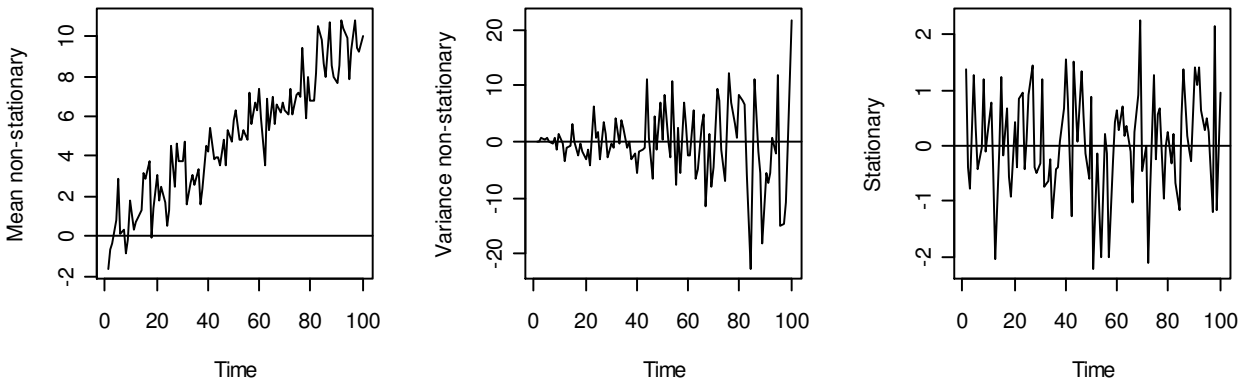


Figure 2.4. One realisation (or path or trajectory) of a mean-nonstationary series (left), variance-nonstationary series (center), and stationary time series (right)

2.2. Covariance Stationary Time Series

In contrast to cross-sectional data where different observations were assumed to be uncorrelated, in time series we require is that there be some dynamics, some persistence, some way in which the present is linked to the past and the future to the present. The known or **historical** data conveys information regarding their future, thus allowing to forecast the future.

If we want to forecast a series, at a minimum we would like its mean and its covariance structure to be stable over time; in that case, we say that the series has covariance stationarity. The first requirement for a series to be covariance stationary is that the mean of the series be stable over time. The mean of the series at time *t* is $EY_t = \mu_t$; if the mean is stable over time, as required by covariance stationarity, then we can write $EY_t \equiv \mu$.

The second requirement for a series to be covariance stationary is that its covariance structure be stable over time. Quantifying stability of the covariance structure is a bit tricky, but tremendously important, and we do it by using the autocovariance function (ACF). The autocovariance at displacement τ is just the covariance between Y_t and $Y_{t-\tau}$. It will, of course, depend on τ, and it may also depend on *t,* so in general we write

$$\gamma(t,\tau) = \mathrm{cov}(Y_t, Y_{t-\tau}) = E(Y_t - \mu)(Y_{t-\tau} - \mu) \qquad (2.1)$$

If the covariance structure is stable over time, as required by covariance stationarity, then the autocovariances depend only on displacement τ, not on time *t,* and we write $\gamma(t,\tau) = \gamma(\tau)$. Note that $\gamma(0) = \mathrm{cov}(Y_t, Y_t) = DY_t (< \infty)$. Autocovariance function is important because, after all, it is the only characteristics which allows us to distinguish between different stationary time series (it provides a basic summary of cyclical dynamics and interdependence in a covariance stationary series.)

A time series Y_t is **stationary** if its mean, variance and covariance do not depend on t: $EY_t \equiv \mu$, $DY_t \equiv \sigma^2$, and $\gamma(t,\tau) \equiv \gamma(\tau)$.

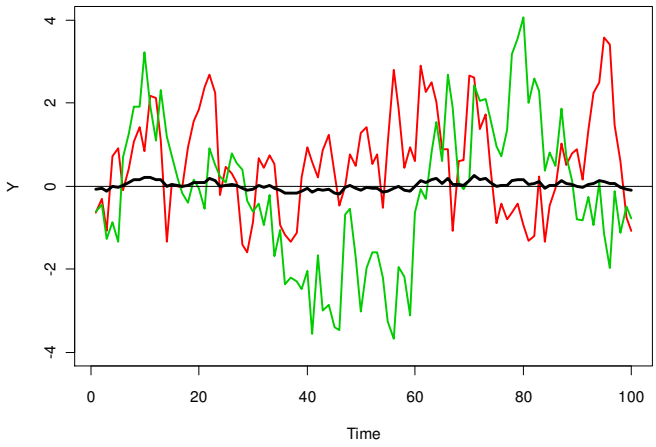


Figure 2.5. Two trajectories (red and green) of stationary process can depart considerably from its mean (=0) but the sample mean of 100 such trajectories (black) is almost equal to 0.

In light of the superior interpretability of correlations compared to covariances, we often work with the correlation between Y_t and $Y_{t-\tau}$. That is, we work with the autocorrelation function $\rho(\tau)$ rather than the autocovariance function $\gamma(\tau)$. The autocorrelation function is obtained by dividing the autocovariance function by the variance:

$$\rho(\tau) = \frac{\text{cov}(Y_t, Y_{t-\tau})}{\sqrt{DY_t DY_{t-\tau}}} = \frac{\gamma(\tau)}{\gamma(0)} \tag{2.2}$$

(note that we always have $\rho(0) = 1, |\rho(\tau)| \leq 1$).

Finally, the partial autocorrelation function (PACF) $p(\tau)$ is sometimes useful ($p(k)$ is just a coefficient on Y_{t-k} in a population linear regression $Y_t = \alpha + \beta_1 Y_{t-1} + ... + \beta_k Y_{t-k} + \varepsilon_t$, i.e., $p(k) = \beta_k$). The partial autocorrelations, in contrast to just autocorrelations, measure the association between Y_t and $Y_{t-\tau}$ after controlling for the effects of $Y_{t-1}, ..., Y_{t-\tau+1}$; that is, they measure the partial correlation between Y_t and $Y_{t-\tau}$.

2.3. White Noise - 2

The fundamental building block of all stationary time series is the so-called white noise[3] process which we define as a sequence of zero mean, constant variance σ^2 and serially uncorrelated[4] r.v.'s (we denote it by $\varepsilon_t \sim (0, \sigma^2)$ or $\varepsilon_t \sim WN(0, \sigma^2)$.) Note that sometimes, instead

[3] White noise is analogous to white light, which is composed of all colors of the spectrum in equal amounts. We can think of white noise as being composed of a wide variety of cycles of differing periodicities in equal amounts.

[4] That is, $\rho(t, \tau) \equiv 0$ for $\tau > 1$ and any t.

of uncorrelatedness, we demand a somewhat stronger property of *independence* (if ε_t is *normally* distributed, then these two definitions coincide; the normal or Gaussian white noise will be denoted by $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$, where i.i.d. means independent identically distributed).

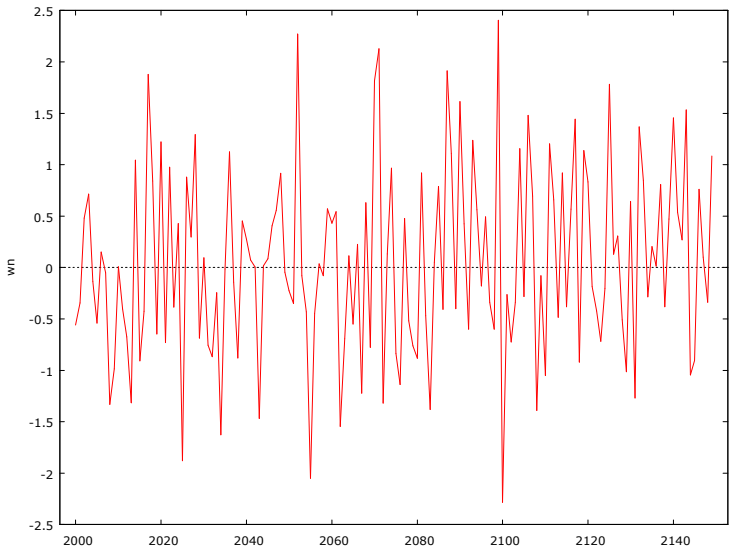


Figure 2.6. Sample path of a simulated[5] Gaussian white noise `wn` of length $T = 150$; there are no patterns of any kind in the series due to the independence over time.

From the definition it follows that $E\varepsilon_t \equiv 0$, $D\varepsilon_t \equiv \sigma^2$. To complete the proof of stationarity, we have to show that ACF $\gamma(t,\tau) = E(\varepsilon_t - E\varepsilon_t)(\varepsilon_{t-\tau} - E\varepsilon_{t-\tau}) = E\varepsilon_t\varepsilon_{t-\tau}$ or, what is the same $\rho(t,\tau)$ $(= \gamma(t,\tau)/\sigma^2)$ does not depend on *t*. Fortunately, it is very easy: for any t,

$$\rho(t,\tau) = \begin{cases} 1, \tau = 0 \\ 0, \tau \neq 0 \end{cases} = \rho(\tau)\,(= p(\tau))$$

(the absence of autocorrelation means that WN "has no memory".)

A bit later we shall show that any stationary process can be expressed in terms of WN. An example in that direction is as follows:

2.1 example. The time series $Y_t = \varepsilon_t + b_1\varepsilon_{t-1}$ is stationary. Indeed,

i) $EY_t = E(\varepsilon_t + b_1\varepsilon_{t-1}) = 0 + b_1 \cdot 0 \equiv 0$, ii) $DY_t \equiv \sigma^2(1+b_1^2)$ and

iii) $\gamma(t,1) = EY_tY_{t-1} = E(\varepsilon_t + b_1\varepsilon_{t-1})(\varepsilon_{t-1} + b_1\varepsilon_{t-2}) \equiv b_1\sigma^2$, $\gamma(t,\tau) = EY_tY_{t-\tau} \equiv 0$, if $\tau > 1$.

The last line means that this process has a very short memory (i.e., if Y_t and $Y_{t+\tau}$ are separated by more that one time period, they are uncorrelated). On the other hand, this series does not make a WN (why?).

[5] GRETL command is `series wn=normal()`; in R we write `wn=rnorm(n)`.

2.4. The Lag Operator

The lag operator and related constructs are the natural language in which time series models are expressed. The lag operator L is very simple. It "operates" on a series by lagging it; thus,

$$LY_t = Y_{t-1}. \tag{2.3}$$

Similarly, $L^2Y_t = L(L(Y_t)) = L(Y_{t-1}) = Y_{t-2}$ and so on: $L^pY_t = Y_{t-p}$. Typically, we operate on a series not with the lag operator but with a polynomial in the lag operator**.** A lag operator polynomial of degree m is just a linear function of powers of L up through the mth power: $B(L) = b_0 + b_1L + b_2L^2 + ... + b_mL^m$. For example, if $B(L) = 1 + 0.9L - 0.6L^2$, then

$$B(L)Y_t = Y_t + 0.9Y_{t-1} - 0.6Y_{t-2}.$$

A well-known operator, the first-difference operator Δ, is actually a first-order polynomial in the lag operator; you can readily verify that $\Delta Y_t = Y_t - Y_{t-1} = (1-L)Y_t$.

Thus far, we have considered only finite-order polynomials in the lag operator; it turns out, however, that infinite-order polynomials are also of great interest. We write the infinite-order lag operator polynomial as

$$B(L) = b_0 + b_1L + b_2L^2 + ... = \sum_{i=0}^{\infty} b_iL^i$$

At first sight, infinite distributed lags may seem esoteric and of limited practical interest because models with infinite distributed lags have infinitely many parameters $(b_0, b_1, b_2, ...)$ and therefore cannot be estimated with a finite sample of data. On the contrary, and surprisingly, it turns out that models involving infinite distributed lags are central to time series modeling and forecasting. Wold's theorem, to which we now turn, establishes that centrality.

2.5. The General Linear Process

What we are looking for now is a model for stationary process. Wold's representation theorem points to the appropriate model.

Wold's Representation Theorem. Let $\{ Y_t \}$ be any zero-mean covariance stationary process. Then we can write it as

$$Y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i\varepsilon_{t-i},\ \ \varepsilon_t \sim WN(0, \sigma^2),$$

where $b_0 = 1$ and $\sum_{i=0}^{\infty} b_i^2 < \infty$. On the other hand, any process of the above form is stationary. ◀

2.2 example. Take $b_1 = b_2 = ... = 0$ - this corresponds to white noise. Thus, once again, WN is a stationary process.

2.3 example. The time series of 2.1 example is stationary

2.4 example. Take $b_1 = \varphi$, $b_2 = \varphi^2$, $b_3 = \varphi^3, \ldots$ – since $1 + \varphi + \varphi^2 + \ldots = 1/(1-\varphi) < \infty$ if $|\varphi| < 1$, the process $Y_t = \varepsilon_t + \varphi\,\varepsilon_{t-1} + \varphi^2 \varepsilon_{t-2} + \ldots$ is stationary in this case.

In our statement of Wold's theorem, we assumed a zero mean. That may seem restrictive, but it is not. Rather, whenever you see Y_t, just read $Y_t - \mu$ so that the process is expressed in deviations from its mean. The deviation from the mean has a zero mean, by construction. Working with zero-mean processes therefore involves no loss of generality while facilitating notational economy.

As we have seen, the Wold representation points to the crucial importance of models with infinite distributed lags. Infinite distributed lag models, in turn, are stated in terms of infinite polynomials in the lag operator, which are therefore very important as well. Infinite distributed lag models are not of immediate practical use, however, because they contain infinitely many parameters, which certainly inhibits practical application. Fortunately, infinite polynomials in the lag operator need not contain infinitely many free parameters. The infinite polynomial $B(L)$ may, for example, be a function of of one parameter φ as in 2.4 example or a ratio of finite-order (and perhaps very low-order) polynomials. Such polynomials are called rational polynomials, and distributed (weighted) lags constructed from them are called rational distributed lags. Suppose, for example, that $B(L) = \Theta(L)/\Phi(L)$, where the numerator polynomial is of degree q, $\Theta(L) = \sum_{i=0}^{q} \theta_i L^i$ and the denominator polynomial is of degree p, $\Phi(L) = \sum_{i=0}^{p} \varphi_i L^i$. There are not infinitely many free parameters in the $B(L)$ polynomial; instead, there are only $p + q$ parameters. If p and q are small, say 0, 1, or 2, then what seems like a hopeless task – estimation of $B(L)$ – may actually be easy.

More realistically, suppose that $B(L)$ is not exactly rational, but is approximately rational. Then we can approximate the Wold representation by using a rational distributed lag. Rational distributed lags produce models of cycles that economize on parameters (they are parsimonious), while nevertheless providing accurate approximations to the Wold representation. The popular ARMA and ARIMA forecasting models, which we will study shortly, are simply rational approximations to the Wold representation.

2.6. Estimation and Inference for the Mean, Autocorrelation, and Partial Autocorrelation Functions

Now suppose we have a sample of data of a stationary time series, but we know neither the true model that generated the data (i.e., the polinomial $B(L)$) nor the mean, autocorrelation function, or partial autocorrelation function associated with that true model. Instead, we want to use the data to *estimate* the mean, autocorrelation function, and partial autocorrelation function, which we might then use to help us learn about the underlying dynamics, and to decide upon a suitable model or set of models to fit to the data.

- **Sample Mean**

The mean of a stationary series is (usually unknown nonrandom) number $\mu = EY_t$. A fundamental principle of estimation, called the analog principle, suggests that we develop estima-

tors by replacing expectations with sample averages. Thus, our estimator of the population mean, given a sample of size *T*, is the sample mean[6] $\bar{Y} = \sum_{t=1}^{T} Y_t / T$. Typically, we are not directly interested in the estimate of the mean, but it is needed for estimation of the autocorrelation function.

- **Sample Autocorrelations**

The autocorrelation at displacement or *lag* τ for the covariance stationary series $\{Y_t\}$ is

$$\rho(\tau) = \frac{E(Y_t - \mu)(Y_{t-\tau} - \mu)}{E(Y_t - \mu)^2}.$$

Application of the analog principle yields a natural estimator of $\rho(\tau)$:

$$\hat{\rho}(\tau) = \frac{\sum_{t=\tau+1}^{T} [(Y_t - \bar{Y})(Y_{t-\tau} - \bar{Y})] / T}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2 / T}.$$

This estimator, viewed as a function of τ, is called the sample autocorrelation function (sample ACF). It is often of interest to assess whether a series is reasonably approximated as white noise, which is to say whether all of its autocorrelations are zero in population. A key result, which we simply assert, is that if a series is white noise, then the sample autocorrelations $\hat{\rho}(\tau),\ \tau = 1,..., K,$ in large samples are independent and all have the $\mathcal{N}(0, (1/\sqrt{T})^2)$ distribution. Thus, if the series is white noise, approximately 95% of the sample autocorrelations should fall in the interval $\pm 2/\sqrt{T}$. In practice, when we plot the sample autocorrelations for a sample of data, we typically include the "two standard error bands" which are useful for making informal graphical assessments of whether and how the series deviates from white noise. Note that exactly the same claim holds for PACF: if a series is white noise, approximately 95% of the sample partial autocorrelations should fall in the interval $\pm 2/\sqrt{T}$. As with the sample autocorrelations, we typically plot the sample partial autocorrelations along with their two-standard-error bands.

The two-standard-error bands, although very useful, provide 95% bounds for only the sample autocorrelations taken one at a time. Ultimately, we are often interested in whether a series is white noise, that is, whether all its autocorrelations are jointly 0, that is, all $\rho(i),\ i = 1, 2,..., K,$ equal zero. Clearly, because of a limited sample size we can take only a finite number of ρ's. It can be shown that the **Ljung-Box statistic** $Q = T(T+2) \cdot \sum_{\tau=1}^{K} \hat{\rho}^2(\tau) / (T - \tau)$**,** under the null hypothesis that *Y* is white noise, is approximately distributed as a χ^2_K random variable. To test the zero $H_0 : \rho(1) = 0,\ \rho(2) = 0,...,\ \rho(K) = 0$, we have to calculate the $p-$value *p-value* $= P(\chi^2_K > q)$: if *p-value* < 0.05, we reject H_0 and assume that *Y* is not white noise.

[6] If we treat Y_t as random, $\bar{Y}$ is random, otherwise it is just a real nonrandom number.

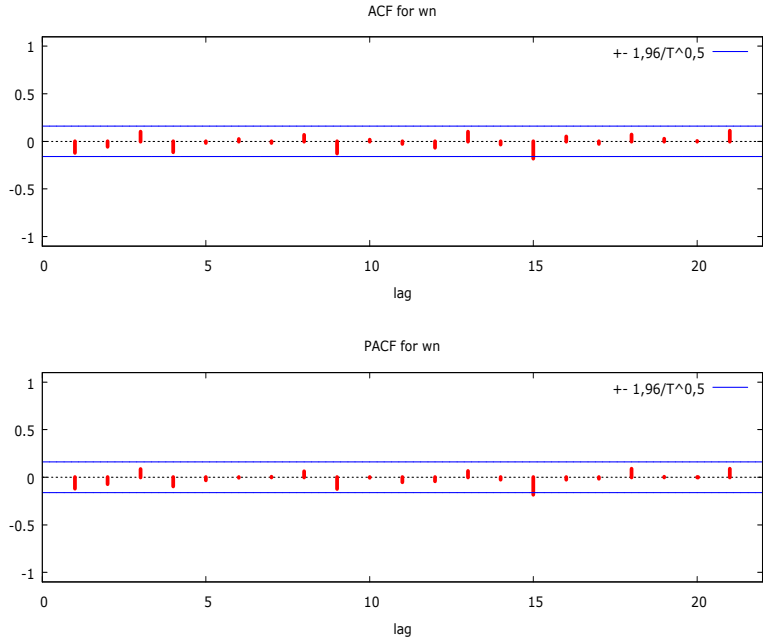


Figure 2.7. ACF and PACF of the generated `wn` of Fig. 2.1; (almost) all $\hat{\rho}(\tau)$ and $\hat{p}(\tau)$ are within "blue band", i.e., there is no ground to reject the white noise hypothesis.

Selection of K is done to balance competing criteria. On the one hand, we don't want K to be too small because, after all, we are trying to do a joint test on a large part of the autocorrelation function. On the other hand, as K grows relative to T, the quality of the distributional approximations we have invoked deteriorates. In practice, focusing on K in the neighborhood of $\sqrt{T}$ is often reasonable[7]. ◀

To illustrate the ideas we have introduced, we examine a quarterly, seasonally adjusted index of Canadian employment (available as caemp.txt) for 1962:1-1995:4 (see Fig.2.8). The series displays no trend, and, of course, it displays no seasonality because it is seasonally adjusted (we shall discuss seasonality in the next chapter). It does, however, appear highly serially correlated (it evolves in a slow, persistent fashion – high in business cycle booms and low in recessions).

The sample autocorrelations are very large relative to their standard errors and display slow one-sided decay. The sample partial autocorrelations, in contrast, are large relative to their standard errors at first (particularly for the 1-quarter displacement) but are statistically negligible beyond displacement 2. It is clear that employment has a strong cyclical component; the diagnostics in Fig. 2.8 rejects the white noise hypothesis. Moreover, the sample autocorrelation and partial autocorrelation functions have particular shapes – the autocorrelation function displays slow one-sided damping, whereas the partial autocorrelation function cuts off at a displacement of 2. You might guess that such patterns, which summarize the dynamics in the series, might be useful for suggesting candidate forecasting models. Such is indeed the case, as we shall see in the next sections.

[7] The "blue lines" are for quick estimation of possible outliers. Always test your assumption of WN with Ljung-Box test not for one but for a set of different K 's (in GRETL and R (function `tsdiag` in the forecast package) this is done automatically).

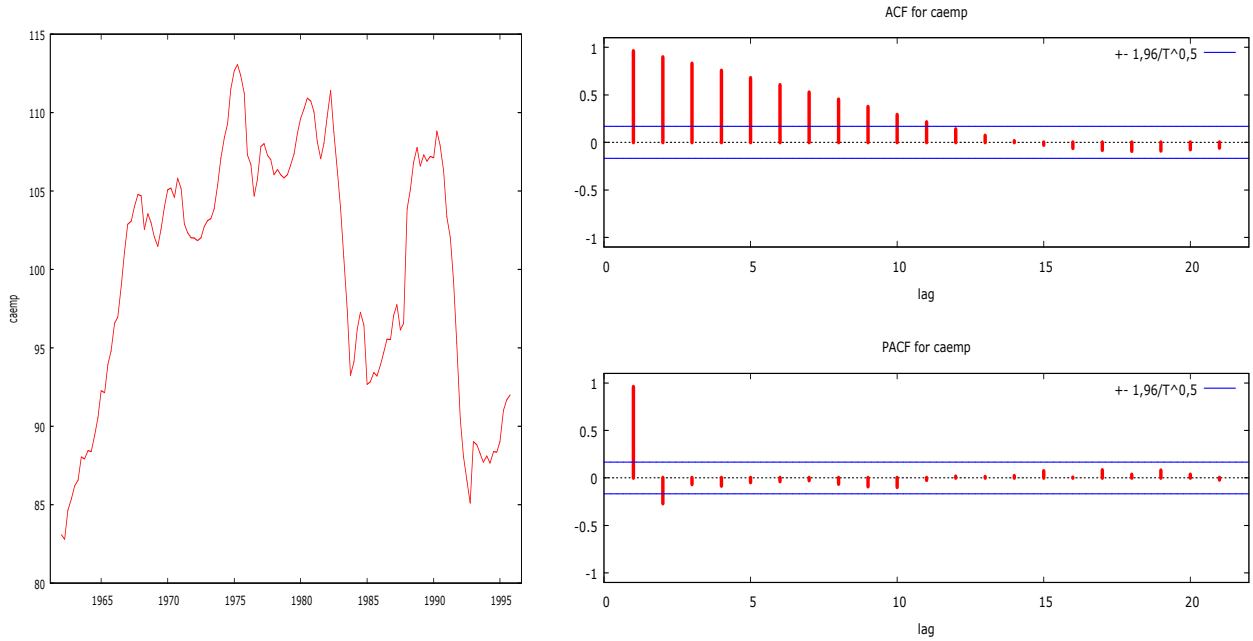


Figure 2.8. Canadian employment index (left; clearly, `caemp` cannot depart too far from its mean, therefore we treat it as stationary) and ACF and PACF graphs (right)

To end up with `caemp`, we shall once again test the WN hypothesis, this time by using the Ljung-Box statistic. After opening GRETL and importing caemp.txt, direct to Model| Time Series| ARIMA…, fill all boxes with 0, and press OK. In the table, choose Graphs| Residual correlogram. You will get the following table:

LAG	ACF		PACF		Q-stat.	[p-value]
1	0,9585	***	0,9585	***	127,7312	[0,000]
2	0,8971	***	-0,2669	***	240,4517	[0,000]
3	0,8285	***	-0,0671		337,3193	[0,000]
4	0,7545	***	-0,0819		418,2578	[0,000]
5	0,6782	***	-0,0451		484,1585	[0,000]
6	0,6018	***	-0,0364		536,4449	[0,000]
7	0,5270	***	-0,0256		576,8441	[0,000]
8	0,4518	***	-0,0627		606,7683	[0,000]
9	0,3736	***	-0,0904		627,3982	[0,000]
10	0,2915	***	-0,0969		640,0584	[0,000]
11	0,2110	**	-0,0218		646,7420	[0,000]
12	0,1369		0,0153		649,5790	[0,000]
...						
19	-0,0868		0,0790		654,5980	[0,000]
20	-0,0745		0,0350		655,4961	[0,000]

In the first line, `Q-stat.(= 127,7312)` corresponds to $m=1$, *p-value* [0.000] is definitely lesss than 0.05, thus we reject the hypothesis $H_0:\rho(1)=0$. In the second line, `Q-stat.(= 240,4517)` corresponds to $m=2$, *p-value* [0.000] is lesss than 0.05, thus we reject the hypothesis $H_0:\rho(1)=0,\rho(2)=0$ and so on. Once again we see that `caemp` is definitely not a white noise.

Now we shall present a few more examples of stationary processes.

2.7. Moving-Average (MA) Models

The finite-order moving-average processes are natural and obvious approximations to the Wold representation, which is an infinite-order moving-average process. Finite-order moving-average processes also have direct motivation: the fact that all variation in time series, one way or another, is driven by shocks of various sorts suggests the possibility of modeling time series directly as distributed lags of current and past shocks, that is, as moving-average processes.

• The MA(1) Process

The first-order moving-average or MA(1) process is

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1+\theta L)\varepsilon_t,\ -\infty < \theta < \infty,\ \varepsilon_t \sim WN(0,\sigma^2). \qquad (2.4)$$

The defining characteristic of the MA process in general, and the MA(1) process in particular, is that the current value of the observed series is expressed as a function of current and lagged unobservable shocks ε_t. Note that whatever is θ, MA(1) is always a stationary process with

$$EY_t = E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) \equiv 0, \qquad DY_t = D\varepsilon_t + \theta^2 D\varepsilon_{t-1} \equiv \sigma^2(1+\theta^2),$$

and

$$\rho(\tau) = \begin{cases} 1, & \tau = 0 \\ \theta/(1+\theta^2), & \tau = 1 \\ 0, & \text{otherwise} \end{cases}.$$

The key feature here is the sharp cutoff in the autocorrelation functions - all the autocorrelations are zero beyond displacement 1, the order of the MA process (see Fig. 2.9, left).

It is easy to find another expression of MA(1). As it follows from (2.4), $\varepsilon_t = Y_t/(1+\theta L)$ or, recalling the formula of geometric series, if $|\theta| < 1$,

$$\varepsilon_t = (1-\theta L+\theta^2L^2-\theta^3L^3+...)\,Y_t = Y_t - \theta Y_{t-1} + \theta^2 Y_{t-2} - \theta^3 Y_{t-3} + ...,$$

that is Y_t can also be expressed as infinite series of its own lags:

$$Y_t = \theta Y_{t-1} - \theta^2 Y_{t-2} + \theta^3 Y_{t-3} - ... + \varepsilon_t. \qquad (2.5)$$

As it follows from this expression, the partial autocorrelation function will decay gradually to zero (we have already mentioned in section 2.1 that the partial autocorrelations are just the coefficients on the last included lag in a sequence of progressively higher order autoregressive approximations.) If $\theta < 0$, then the pattern of decay will be one-sided; otherwise, the decay will be one of damped oscillation (see Fig. 2.9, right).

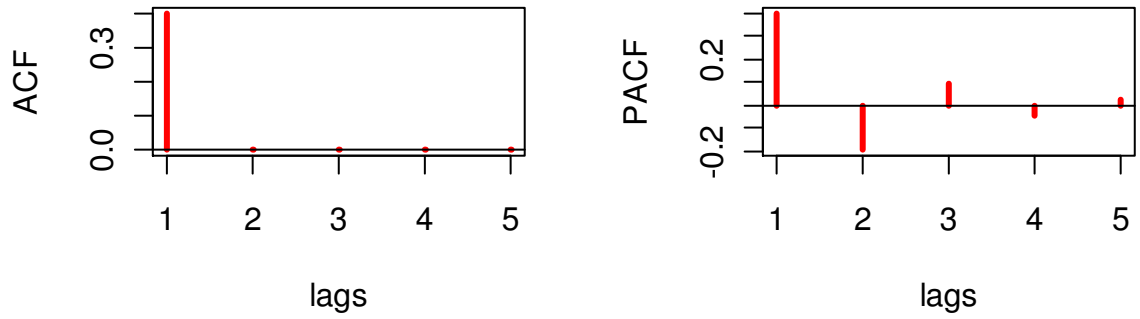


Figure 2.9. The theoretical correlogramm of the MA(1) process with $\theta = 0.5$.

- **The MA(q) Process**

Now we shall consider the general finite-order moving-average process of order *q,* or MA(q) process for short,

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} = \Theta(L)\varepsilon_t, \; -\infty < \theta_i < \infty, \; \varepsilon_t \sim WN(0, \sigma^2)$$

where

$$\Theta(L) = 1 + \theta_1 L + ... + \theta_q L^q \qquad (2.6)$$

is a *q*th-order lag operator polynomial. The MA(q) process is a natural generalization of the MA(1) process. By allowing for more lags of the shock on the right side of the equation, the MA(q) process can capture richer dynamic patterns, which we can potentially exploit for improved forecasting. The MA(1) process is, of course, a special case of the MA(q), corresponding to q=1.

The properties of the MA(q) processes parallel those of the MA(1) process in all respects, so in what follows we refrain from grinding through the mathematical derivations. Just as the MA(1) process was covariance stationary for any value of its parameters, so too is the finite-order MA(q) process. The potentially longer memory of the MA(q) process emerges clearly in its autocorrelation function. In the MA(1) case, all autocorrelations beyond displacement 1 are 0; in the MA(q) case, all autocorrelations beyond displacement *q* are 0. This autocorrelation cutoff is a distinctive property of moving-average processes. The partial autocorrelation function of the MA(q) process, in contrast, decays gradually, in accordance with the infinite autoregressive representation.

Note that if we have a finite sample of a MA(q), its sample correlogram should not differ much from the theoretical one. This will help us to detect that the time series under consideration is a moving average process.

Again, as in MA(1) case, if θ_i 's satisfy certain conditions, Y_t may be written as infinite convergent series $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + ... + \varepsilon_t$ (such a MA(q) process is called invertible).

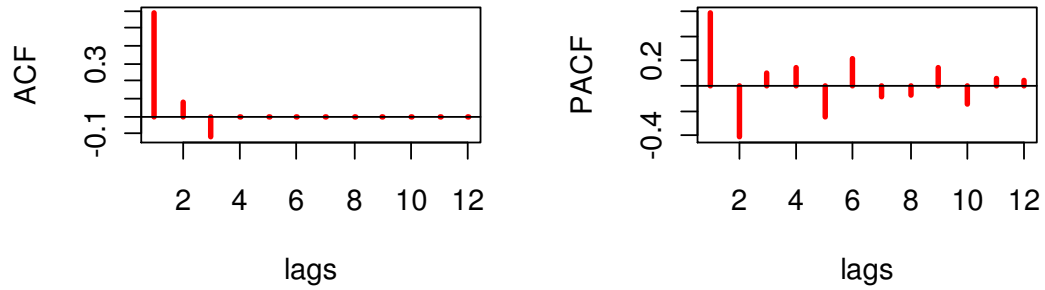


Figure 2.10. The theoretical correlogramm of the MA(3) process with $\theta_1 = 1.2,\ \theta_2 = 0.65,\ \theta_3 = -0.35$ (note that its ACF cuts off at $t = 3$ and PACF decays gradually).

2.8. Autoregressive (AR) Models

The autoregressive process is also a natural approximation to the Wold representation. We have seen, in fact, that under certain conditions a moving average process has an autoregressive representation (see (2.5)) so an autoregressive process is in a sense the same as a moving-average process.

- **The AR(1) Process**

The first-order autoregressive process, AR(1) for short, is

$$Y_t = \varphi Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN(0, \sigma^2)$$

or, in lag operator form, $(1-\varphi L)Y_t = \varepsilon_t$ or[8] $Y_t = \dfrac{1}{1-\varphi L}\varepsilon_t$.

Note that the fluctuations (of the simulated AR(1) process) in the $\varphi = 0.9$ case (see Fig. 2.11) appear much more persistent that those of the AR(1) with parameter $\varphi = 0.1$. The former case contrasts sharply with the MA(1) process, which has a very short memory regardless of parameter value. Thus the AR(1) model is capable of capturing much more persistent dynamics than is the MA(1) model.

Also note the special interpretation of the errors or disturbances or shocks ε_t in the time series theory: in contrast to the regression theory where they were understood as the summary of all unobserved $X's$, now they are treated as economic effects which have developed in period t.

[8] We have still to explain how to understand $1/(1-\varphi L)$.

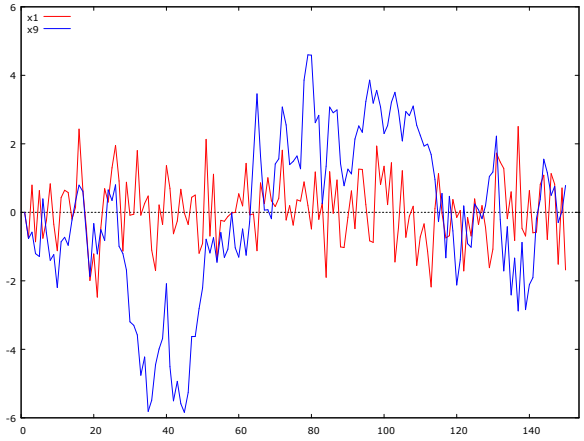


Figure 2.11. Two simulated[9] AR(1) processes - red with $\varphi = 0.1$ (it is very like a WN) and blue with $\varphi = 0.9$ (the much more persistent one)

Recall that a finite-order moving-average process is always covariance stationary, but that certain conditions must be satisfied for AR(1) to be stationary. The AR(1) process $Y_t = \varphi Y_{t-1} + \varepsilon_t$ can be rewritten as

$$Y_t = \frac{1}{1-\varphi L}\varepsilon_t = (1+\varphi L+\varphi^2 L^2+...)\,\varepsilon_t = \varepsilon_t + \varphi\,\varepsilon_{t-1} + \varphi^2\,\varepsilon_{t-2} + \dots .$$

This Wold's moving-average representation for Y is convergent if $|\varphi| < 1$; thus

AR(1) is stationary if $|\varphi| < 1$

Equivalently, the condition for covariance stationarity is that the root z_1 of the autoregressive lag operator polynomial (i.e., $1-\varphi z = 0 \sim z_1 = 1/\varphi$) be greater than 1 in absolute value. This seems to be a complicated and redundant variant of the previous condition, but we shall see shortly that a similar condition on the roots is important in AR(p) case.

From the moving-average representation of the covariance stationary AR(1) process, we can compute the mean and variance:

$$EY_t = ... = 0\,,\ DY_t = ... = \frac{\sigma_\varepsilon^2}{1-\varphi^2}\,.$$

An alternative procedure to find the mean is as follows: when $|\varphi| < 1$, the process is stationary, i.e., $EY_t \equiv m$; therefore, $EY_t = \varphi EY_{t-1} + E\varepsilon_t$ implies $m = \varphi m + 0$ or $m = 0$. This methods allows us to easily estimate the mean of the following generalized AR(1) process: if $Y_t = \alpha + \varphi Y_{t-1} + \varepsilon_t$, then $m = \alpha/(1-\varphi)$.

[9] In GRETL this can be done with the following commands:

```
series e = normal()
series x = 0
x = 0.9*x(-1) + e
```

It is easy to show that the correlogram of AR(1) is in a sense symmetric to that of MA(1) – the ACF function decays exponentially, namely, $\rho(\tau)=\varphi^{\tau}$, $\tau=0,1,2,...$ and PACF cuts off abruptly, specifically,

$$p(\tau)=\begin{cases}\varphi, & \tau=1\\ 0, & \tau>1\end{cases}.$$

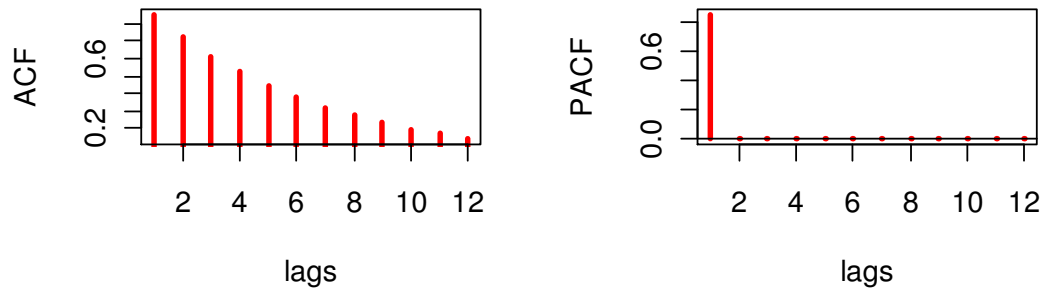


Figure 2.12. Theoretical AC function (left) and PAC function (right) for the AR(1) process with $\varphi = 0.85$.

- **The AR(p) Process**

The general *p*th order autoregressive process, or AR(p) process for short, is

$$Y_t=\varphi_1 Y_{t-1}+\varphi_2 Y_{t-2}+...+\varphi_p Y_{t-p}+\varepsilon_t,\ \varepsilon_t\sim WN(0,\sigma^2).$$

In lag operator form, we write

$$\Phi(L)Y_t=(1-\varphi_1 L-\varphi_2 L^2-...-\varphi_p L^p)Y_t=\varepsilon_t.$$

Similar to the AR(1) case, the AR(p) process is covariance stationary if and only if all the roots z_i of the autoregressive lag operator polynomial $\Phi(z)$ are outside the complex unit circle[10] (note this condition does not have any relationship with the value of φ_1):

$$1-\varphi_1 z-\varphi_2 z^2-...-\varphi_p z^p=0 \Rightarrow |z_i|>1.$$

AR(p) is stationary if all the roots $|z_i|>1$

In the covariance stationary case, we can write the process in the convergent infinite moving-average form $Y_t=\left(1/\Phi(L)\right)\varepsilon_t$. The autocorrelation function for the general AR(p) process, as

[10] For a quick check of stationarity, use the following rule: if $\sum_{i=1}^{p}\varphi_i\geq 1$, the process isn't stationary.

with that of the AR(1) process, decays gradually when the lag increases. Finally, the AR(p) partial autocorrelation function has a sharp cutoff at displacement p for the same reason that the AR(1) partial autocorrelation function has a sharp cutoff at displacement 1.

2.5 example. Consider the AR(2) process $Y_t = 1.5Y_{t-1} - 0.9Y_{t-2} + \varepsilon_t$. The corresponding lag operator polynomial is $1 - 1.5L + 0.9L^2$ with two complex conjugate roots, $z_{1,2} = 0.83 \pm 0.58\text{i}$, $|z_{1,2}| = 1.0253 > 1$ (thus, the process is stationary). It can be shown that the autocorrelation function for an AR(2) process is $\rho(0) = 1$, $\rho(1) = \varphi_1 / (1 - \varphi_2)$ and

$$\rho(\tau) = \varphi_1 \rho(\tau - 1) + \varphi_2 \rho(\tau - 2), \tau = 2, 3, ...$$

By using this formula, we can evaluate the autocorrelation function for the process at hand. We plot the correlogram of AR(2) process in Fig. 2.13. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

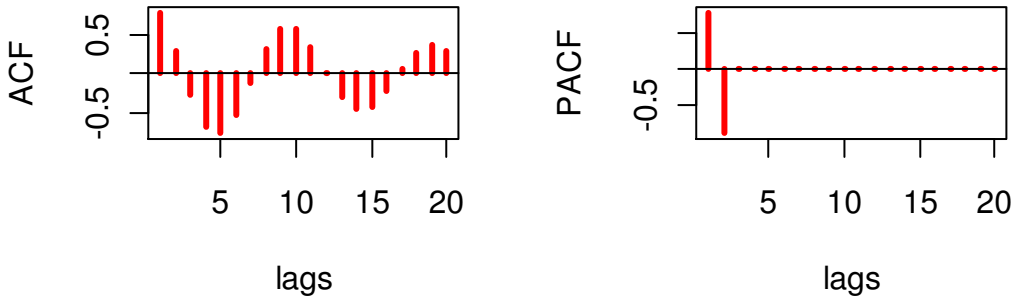


Figure 2.13. Theoretical ACF of AR(2) process with complex roots

The AR(p) is an obvious generalization of the AR(1) strategy for approximating the Wold representation. The (infinite) moving-average representation associated with the AR(p) process depends on p parameters only: $Y_t = (1 / \Phi(L))\varepsilon_t$ and this fact simplifies the estimation procedure.

Thus, a stationary AR process can be rewritten as a MA process (in general, of infinite order). However, in some cases the AR form of a stationary process is preferred to that of MA. The necessary definition says that the MA process is called invertible if it can be expressed as an AR process (in general, of infinite order). It appears that the MA(q) process

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} = \Theta(L)\varepsilon_t, \; -\infty < \theta_i < \infty, \; \varepsilon_t \sim WN(0, \sigma^2)$$

is invertible if all the roots of the polynomial $\Theta(x) = 1 + \theta_1 x + ... + \theta_q x^q$ lie outside the unit circle. For example, the MA(1) process $Y_t = \varepsilon_t - \varepsilon_{t-1}$ (differences of WN) is not invertible (why?).

2.9. Autoregressive Moving-Average (ARMA) Models

Autoregressive and moving-average models are often combined in attempts to obtain better and more parsimonious approximations to the Wold representation, yielding the autoregressive moving-average process, **ARMA(p,q) process** for short. As with moving-average and autoregressive processes, ARMA processes also have direct motivation. First, if the random shock that drives an autoregressive process is itself a moving-average process, then it can be shown that we obtain an ARMA process. Second, ARMA processes can arise from aggregation. For example, sums of AR processes, or sums of AR and MA processes, can be shown to be ARMA processes. Finally, AR processes observed subject to measurement error also turn out to be ARMA processes.

The simplest ARMA process that is not a pure autoregression or pure moving-average process is the ARMA(1,1), given by

$$Y_t = \varphi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1},\ \varepsilon_t \sim WN(0, \sigma^2)$$

or, in lag operator form,

$$(1 - \varphi L) Y_t = (1 + \theta L) \varepsilon_t$$

where $|\varphi| < 1$ is required for stationarity and $|\theta| < 1$ is required for invertibility. If the covariance stationarity condition is satisfied, then we have the moving-average representation

$$Y_t = \frac{(1 + \theta L)}{(1 - \varphi L)} \varepsilon_t \ \left(= \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + ... \right)$$

which is an infinite distributed lag of current and past innovations. Similarly, we can rewrite it in the infinite autoregressive form

$$\left(Y_t + a_1 Y_{t-1} + a_2 Y_{t-2} + ... = \right) \frac{1 - \varphi L}{1 + \theta L} Y_t = \varepsilon_t .$$

The ARMA(p,q) process is a natural generalization of the ARMA(1,1) that allows for multiple moving-average and autoregressive lags. We write

$$Y_t = \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q,\ \varepsilon_t \sim WN(0, \sigma^2)$$

or

$$\Phi(L)\, Y_t = \Theta(L)\, \varepsilon_t .$$

If all of the roots of $\Phi(L)$ are outside the unit circle, then the process is stationary and has convergent infinite moving-average representation $Y_t = (\Theta(L) / \Phi(L))\, \varepsilon_t$. If all roots of $\Theta(L)$ are outside the unit circle, then the process can be expressed as the convergent infinite autoregression $(\Phi(L) / \Theta(L))\, Y_t = \varepsilon_t$.

ARMA models, by allowing for both moving average and autoregressive components, often provide accurate approximations to the Wold representation that nevertheless have just a few parameters. That is, ARMA models are often both highly accurate and highly parsimonious. In a particular situation, for example, it might take an AR(5) to get the same approximation accuracy as could be obtained with an ARMA(1,1), but the AR(5) has five parameters to be estimated, whereas the ARMA(1,1) has only two.

The rule to determine the number of AR and MA terms:
• AR(p) – ACF declines, PACF = 0 if k > p
• MA(q) – ACF = 0 if k > q, PACF declines
• ARMA(p,q) – both ACF and PACF decline

2.10. Specifying and Estimating Models with GRETL

In Fig. 2.11, we plotted a sample of AR(1) process $Y_t = (\varphi Y_{t-1} + \varepsilon_t =) 0.9 Y_{t-1} + \varepsilon_t$. How can we restore or **estimate** the coefficient φ assuming that we have „forgotten" the way the trajectory was created? The first step is to make a guess about the type of the process, i.e., MA, AR, or ARMA? To do this, draw a correlogram of the time series (Fig. 2.14).

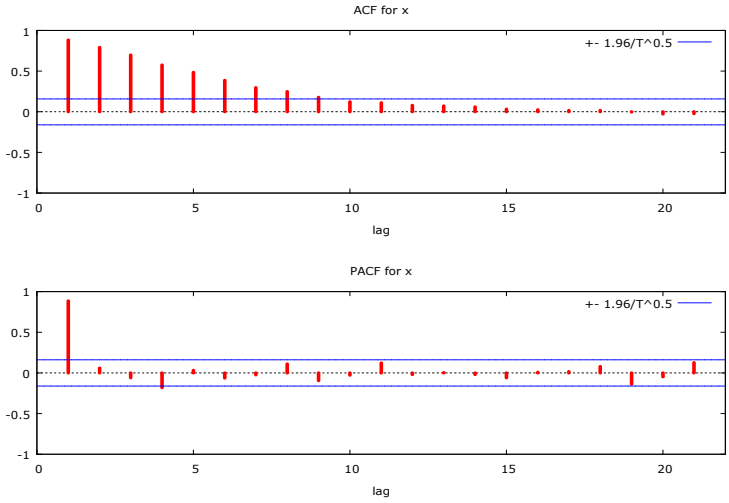


Figure 2.14. The correlogram of the time series x of Fig. 2.11

The ACF is declining, PACF is cut off at 1, thus most probably this is an AR(1) process. As a second step, in order to estimate its coefficients in GRETLl, go to script window and type `arma 1 0 ; x`

```
Model 3: ARMA, using observations 1-150
Dependent variable: x

             coefficient   std. error      z         p-value
  ------------------------------------------------------------
  const       0.0262209    0.554593      0.04728   0.9623
  phi_1       0.886945     0.0366913    24.17      4.26e-129 ***
```

The intercept is insignificant, thus we repeat the calculation without it:

```
arma 1 0 ; x -nc
series res = $uhat    # save residuals ε̂_t
corrgm res 12         # find their Ljung-Box p-values
```

```
Model 4: ARMA, using observations 1-150
Dependent variable: x

             coefficient   std. error      z      p-value
  ---------------------------------------------------------
  phi_1        0.886872     0.0366704    24.18   3.20e-129 ***
```

```
Autocorrelation function for res

  LAG      ACF           PACF          Q-stat.  [p-value]

    1  -0.0387       -0.0387           0.2297  [0.632]
    2   0.0696        0.0682           0.9752  [0.614]
    3   0.1361  *     0.1421  *        3.8471  [0.278]
    4  -0.0738       -0.0689           4.6966  [0.320]
    5   0.0770        0.0530           5.6300  [0.344]
    6  -0.0156       -0.0202           5.6688  [0.461]
    7  -0.1358  *    -0.1315           8.6111  [0.282]
    8   0.0878        0.0624           9.8493  [0.276]
    9  -0.0449       -0.0078          10.1756  [0.336]
   10  -0.1430  *    -0.1352  *       13.5054  [0.197]
   11   0.0663        0.0338          14.2265  [0.221]
   12  -0.0594       -0.0034          14.8091  [0.252]
```

The model is quite satisfactory (the coefficient is significant, residuals make a WN) and the estimated $\hat{\varphi}$ (=0.89) is very close to the true value of 0.9.

2.1 exercise. Create a new data set 200 observation long (go to File| New data| ...| set Frequency Other 1). Generate AR(2) process $Y_t = 4 + 0.5Y_{t-1} - 0.3Y_{t-2} + \varepsilon_t$, $\varepsilon_t \sim (0, 2^2)$ (see Footnote 7). Is it a stationary process (use the graph of Y_t)? What is its theorethical and sample mean? Now "forget" the formula and draw the correlogram of the time series. What is your guess about the type of the time series? To estimate its parameters, go to Model| Time Series| ARIMA… Are they close to 4, 0.5, and -0.3? Once the model is created, draw its 15 steps forecast with Analysis| Forecasts… ◀◀

2.2 exercise. Assume that Y_t is a stationary AR(1) process $Y_t = \varphi Y_{t-1} + \varepsilon_t$. What are the formulas to estimate φ? Recall that according to (3.4), PE.I – Lecture Notes,

$$\hat{\varphi}^{OLS} = \widehat{\mathrm{cov}}(Y_t, Y_{t-1}) / \widehat{\mathrm{var}}(Y_{t-1}) = \hat{\rho}(1).$$

On the other hand, multiply both sides of $Y_t = \varphi Y_{t-1} + \varepsilon_t$ by Y_{t-1} and take its expectation - you will get the so-called Yule-Walker equation $\gamma(1) = \varphi\,\gamma(0)$ and, by the analogue principle, the same estimator of φ.

Now, use the same Yule-Walker procedure for $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varepsilon_t$ and write the formulas for $\hat{\varphi}_1$ and $\hat{\varphi}_2$. ◀◀

Generally, the stationary AR(p) process $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, can be treated in exactly the same way as a classical regression problem where Y_t is the dependent variable and $Y_{t-1}, ..., Y_{t-p}$ are the independent variables. That is to say, the OLS estimates of the coefficients φ_i, as well as the joint limiting distribution of these estimates, are the same as if the equation was treated as a classical regression problem

Now let us switch to a more realistic example. In Fig. 2.8, we examined the correlogram for the quarterly, 1962:1-1995:4, index of Canadian employment (available as caemp.txt) and saw that the sample autocorrelations damp slowly and the sample partial autocorrelations cut off, just the opposite of what is expected for a moving average. Thus, the correlogram indicates that a finite-order moving-average process would not provide a good approximation to employment dynamics. Nevertheless, nothing stops us from fitting moving average models, so let us fit them and use AIC and SIC to guide model selection.

The next step would be to estimate MA(q) models, q=1,2,3,4. Both AIC and SIC suggest that the MA(4) is best. The results of the MA(4) estimation, although better than lower-order MAs, are nevertheless poor. For example, the correlogram and Ljung-Box statistic of `caempRES` show that residuals are far from being white noise:

Table 3.1: MA(q) models

q	AIC	SIC
1	795,86	804,60
2	686,97	698,62
3	624,40	638,97
4	579.26	596.74

```
arma 0 4 ; caemp      # generate a model
fcast caemp04         # generate a forecast series
series caempRES=$uhat # generate a residual series and test it for WN
```

Let us now consider alternative approximations, such as autoregressions. The AR(p) processes can be conveniently estimated by ordinary least squares regressions or `arma` function.

```
ols caemp 0 caemp(-1) caemp(-2) caemp(-3) caemp(-4)    # the 1st method
arma 4 0; caemp # Note that these two estimates differ (different methods)
arma 3 0; caemp # estimate AR(3)
arma 2 0; caemp
arma 1 0; caemp
```

Table 3.2: AR(p) models

p	AIC	SIC
1	525,93	534,67
2	493,57	505,22
3	494,95	509,51
4	496,88	514,35

The 4th and 3rd order lags in AR(4) model are insignificant and the 3rd order lag in AR(3) is insignificant as well; in the AR(2) model all variables are significant and its AIC and SIC are less than in other models. On the other hand, `arma 2 0; caemp` residuals make a nearly perfect WN (check).

```
? arma 2 0; caemp

Model 2: ARMA, using observations 1962:1-1995:4 (T = 136)
Estimated using Kalman filter (exact ML)
Dependent variable: caemp
```

```
            coefficient   std. error   t-ratio    p-value
  -------------------------------------------------------------
  const       97,4773      4,40286      22,14     1,32e-108 ***
  phi_1        1,45048     0,0748910    19,37     1,44e-083 ***
  phi_2       -0,476183    0,0761685    -6,252    4,06e-010 ***

Mean dependent var   100,2198   S.D. dependent var   7,997169
Mean of innovations  0,002213   S.D. of innovations  1,421887
Log-likelihood      -242,7859   Akaike criterion     493,5718
Schwarz criterion    505,2224   Hannan-Quinn         498,3063
```

We could stop here but consider for completeness the ARMA(p,q) approximation to the Wold representation. The model `arma 3 1 ; caemp` is almost as good as `arma 2 0; caemp` (the AIC of the latter model is the smallest among its neighbours and equals 496.02) but it has more parameters than AR(2) model. Following the KISS (Keep It Sophisticatedly Simple) principle and also comparing AIC's, we choose AR(2) model.

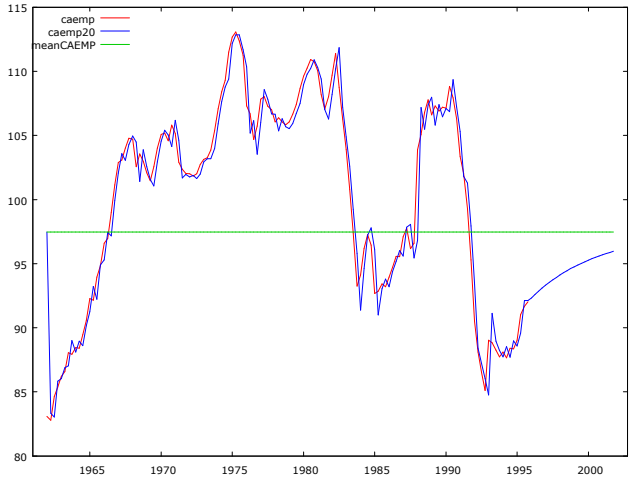


Figure 2.15. caemp and its AR(2) forecast for 1962:1 – 2001:4

2.11. Forecasting

So far we thought of the information set as containing the available past history of the series, $\Omega_T = \{Y_T, Y_{T-1}, \ldots\}$, where for theoretical purposes we imagine history as having begun in the infinite past (note that $\{Y_T, Y_{T-1}, \ldots\}$ contains the same information as $\{\varepsilon_T, \varepsilon_{T-1}, \ldots\}$). Based upon that information set, we want to find the optimal forecast of Y at some future time $T+h$ (we shall denote the forecast through $Y_{T+h,T}, h \geq 1$). If Y_t is a stationary process, then the forecast tends to the process mean as h increases, therefore, the forecast is interesting only for several „small" values of h.

- **Forecasting MA(q) process**

Our forecast method is always the same: write out the process for the future time period, $T+h$, and project it on what is known at time T, when the forecast is made. Consider, for example, an MA(1) process $Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon \sim WN(0, \sigma_\varepsilon^2)$. We have

$$Y_{T+1} = \mu + \varepsilon_{T+1} + \theta\varepsilon_T, \quad Y_{T+1,T} = \mu + 0 + \theta\varepsilon_T$$
$$Y_{T+2} = \mu + \varepsilon_{T+2} + \theta\varepsilon_{T+1}, \; Y_{T+2,T} = \mu + 0 + 0$$
$$Y_{T+h,T} \equiv \mu, h \geq 2$$

The forecast quickly approaches the (sample) mean of the process and starting with $h = \mathrm{q}+1$ coincides with it. When h increases, the accuracy of the forecast diminishes (i.e., the width of the confidence band increases) up to the moment $h = \mathrm{q}+1$ whereupon it becomes constant.

2.6 example. Most financial studies involve returns, instead of prices, of assets. There are two main reasons for using returns. First, for average investors, return of an asset is a complete and scale-free summary of the investment opportunity. Second, return series are easier to handle than price series because the former have more attractive statistical properties (returns are often stationary). There are, however, several definitions of an asset return. Let P_t be be the price of an asset at time t. A **simple return** is $R_t = (P_t - P_{t-1})/P_{t-1}$. If $p_t = \ln(P_t)$, then $r_t = p_t - p_{t-1}$ is **log return**. If these returns are small (which is usually the case), they are almost identical.

Figure 2.16 shows the time plot of monthly simple returns of the CRSP equal-weighted index from January 1932 to December 2003 (the data is available as m-ew.txt). One can see that the series has significant ACF at lags 1, 3, and 9. There are some marginally significant ACFs at higher lags, but we do not consider them here. Based on the sample ACF, the following MA(9) model

$$(R_t =)Y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_3\varepsilon_{t-3} + \theta_9\varepsilon_{t-9}$$

is identified for the series. Note that, unlike the sample PACF, the sample ACF provides information on the nonzero MA lags of the model.

```
arma 0 9 ; crsp                  # all 9 MA terms; some are insignificant
arma 0 {1 3 9} ; crsp            # only 1st, 3rd, and 9th MA terms
addobs 12                        # we shall forecast crsp for 12 months
fcast crspMA
```

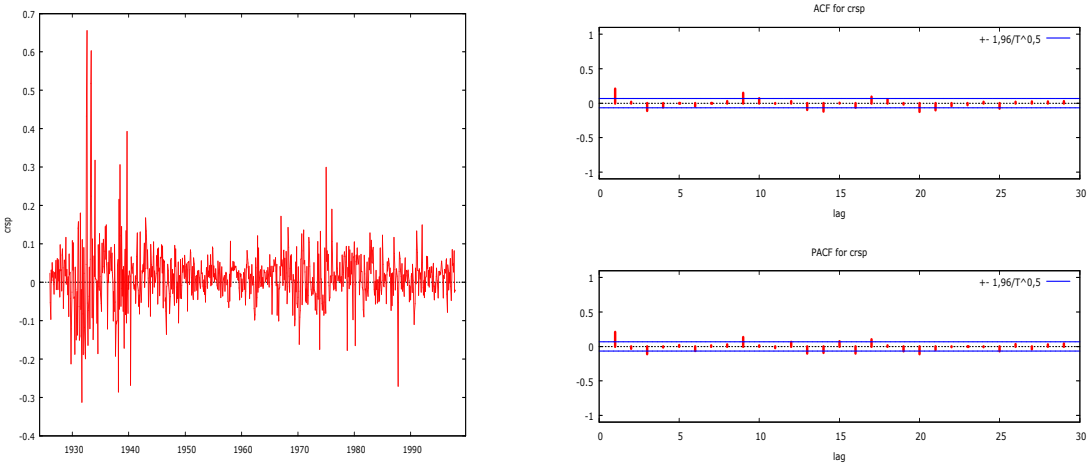


Figure 2.16. Monthly simple returns and sample ACF of CRSP equal-weighted index

It is easy to check that the second model has the smallest AIC and SIC among similar models. We forecast `crsp` for the coming year. To enhance the plot visibility, go to Sample| Set range.. and choose Start: 1996:1. In Fig. 3.9 one can see that in 9 months forecast sets to the mean value and does not change anymore.

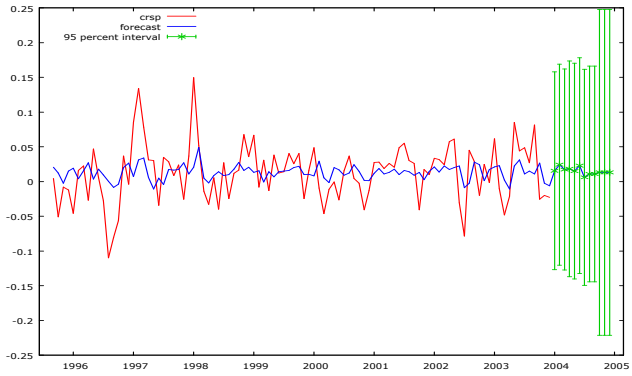


Figure 2.17. `crsp` (red) and its forecast (blue) (the graphs were plotted through the menu bar)

- **Forecasting AR(p) process**

The forecast exponentially fast tends to the (sample) mean of the process but never reaches it. When h increases, the accuracy of the forecast diminishes (i.e., the width of the confidence band increases) but never reaches the limit. For example, in AR(1) case, we write out the process for time $T+1$, $Y_{T+1} = \varphi Y_T + \varepsilon_{T+1}$, and project the right-hand side on time T information set: $Y_{T+1,T} = \varphi Y_T$. To get 2-step-ahead forecast, write $Y_{T+2} = \varphi Y_{T+1} + \varepsilon_{T+2}$ and project it on the same information set: $Y_{T+2,T} = \varphi Y_{T+1,T} = \varphi^2 Y_T$. Similarly, $Y_{T+h,T} = \varphi^h Y_T$ for any h, thus the forecast tends to zero (or to the mean, in general case) exponentially fast.

2.7 example. Now we put our forecasting technology to work to produce point and interval forecasts for Canadian employment. Recall that the best moving-average model was an MA(4), whereas the best autoregressive model, as well as the best ARMA model and the best model overall, was an AR(2). We model `caemp` restricting our data to 1962:1 – 1993:4, forecast it to the next two years and compare with the real 1994:1 – 1995:4 data.

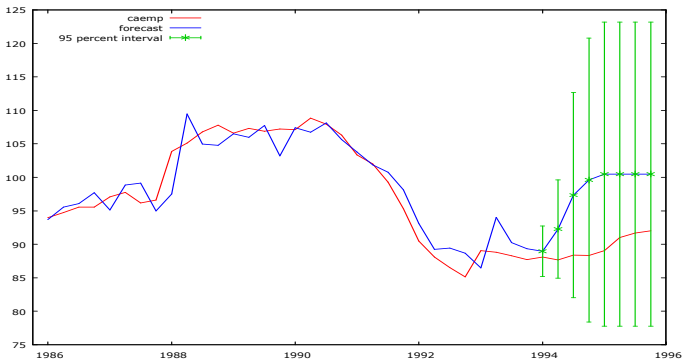


Figure 2.18. Employment history and forecast together with interval extrapolation and realization - MA(4) model (blue line is for forecast, green lines for 95% confidence interval)

Fig. 2.18 shows the 8-quarter-ahead forecast with MA(4) and realization. It proves that MA(4) model is unsatisfactory. The actual employment series stays well below its mean over the forecast period, whereas the forecast rises quickly back to the mean.

Now consider forecasting with the AR(2) model. Fig. 2.19 shows the 8-quarter-ahead extrapolation forecast, which reverts to the unconditional mean much slowlier quickly, which seems natural given the high persistence of employment. The 8-quarter-ahead point forecast is, in fact, still well below the mean. Similarly, the 95% error bands grow gradually and have not approached their long-horizon values by eight quarters out.

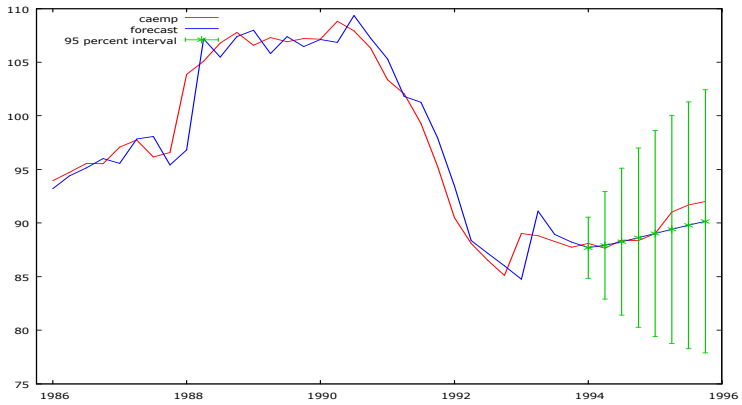


Figure 2.19. Employment history (red) and AR(2) forecast (blue) together with the interval extrapolation (the green lines are for 95% confidence interval)

The forecast $Y_{T+h,T}$ of MA(q) process in (h =) q steps reaches its average and then does not change anymore.
The forecasts of AR(p) and ARMA(p,q) tend to the average, but never reach it. The speed of convergence depends on the coefficients.

2.12. Financial Volatility and the ARCH Models

Consider a variable Y_t evolving over time so that it grows annually at rate r (usually, r is a small number, $|r|<0.05$): $Y_t=(1+r)Y_{t-1}=(1+r)^2Y_{t-2}=...=(1+r)^tY_0=e^{\log(1+r)\cdot t}Y_0=e^{Rt}Y_0$ (thus, the values of Y_t lie on an exponent; this might roughly describe the growth of GDP, capital stock, or (stock) price level). The model is too simplistic to describe a real grow, so we have to introduce an economic shock component, for example, as $Y_t=(1+r+\varepsilon_t)\cdot Y_{t-1}=\prod_{s=1}^{t}(1+r+\varepsilon_s)\cdot Y_0$, $\varepsilon_t\sim WN(0,\sigma^2)$. The values of Y_t are again close to exponent and $EY_t=e^{Rt}Y_0$ (thus Y_t is nonstationary; their differences $\Delta Y_t=Y_t-Y_{t-1}\approx e^{R(t-1)}\cdot(e^R-1)$ are nonstationary, too (why?)). However, take the logarithms and apply the well known equality $\log(1+x)\approx x$ (it follows from the Taylor's expansion and holds for small x's): $\tilde{Y}_t=\log Y_t=\log Y_0+\sum_{s=1}^{t}\log(1+r+\varepsilon_s)\approx\log Y_0+rt+\sum_{s=1}^{t}\varepsilon_s$. The graph of $\tilde{Y}_t$ now is close to a straight line (thus, $\tilde{Y}_t$ is nonstationary) but its differences (or the log differences of Y_t)

$\Delta\tilde{Y}_t = r + \varepsilon_t$ are stationary (why?). What is even more important, these differences have a clear economic interpretation: it is the series of (logarithmic) returns (see 2.6 example), i.e., the annual growth of Y_t. It is well known that stock or bond returns, exchange rates and some other financial series Y_t (to simplify notation, we shall write Y_t instead of $\Delta\tilde{Y}_t$) are almost unpredictable. i.e., are (satisfactory) described as white noise $Y_t = r + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$ (in other words, stock returns are on average r but otherwise unpredictable from the past values).

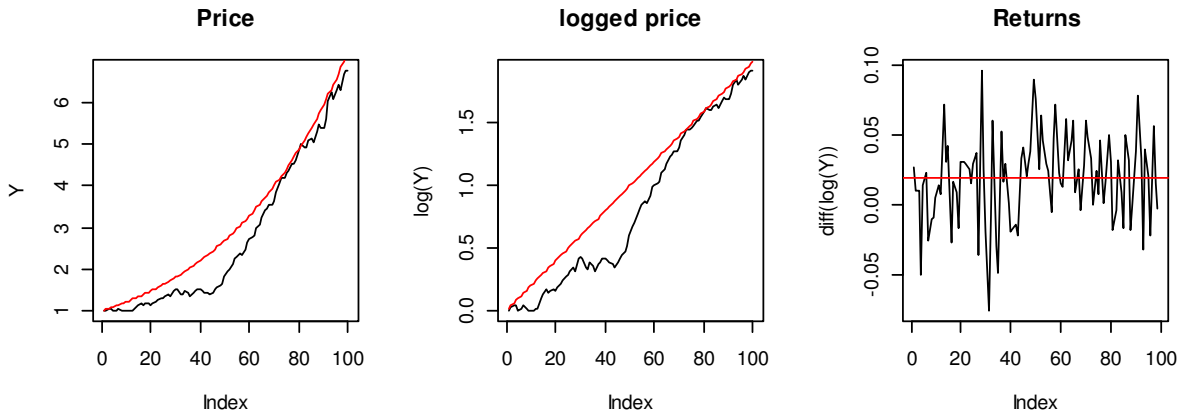


Figure 2.20. Price, logged price, and (logarithmic) returns

In this section, we shall try to create a model of the residuals $e_t = \hat{\varepsilon}_t$, i.e., centered returns $Y_t - \overline{Y} = Y_t - \hat{r}$ of **real** stocks possessing some interesting empirical properties, for example:

- the high-volatility[11] events tend to cluster in time[12] (it is called the *persistency* or inertia of volatility)
- Y_t is uncorrelated with its lags, but Y_t^2 is correlated with $Y_{t-1}^2, Y_{t-2}^2, \ldots$
- Y_t is heavy-tailed, i.e., the right tail of its density decreases slowlier than that of the Gaussian density (this means that Y_t (for example, returns) take big values more often than the Gaussian r.v.)

The core idea of the ARCH (Auto Regressive Conditional Heteroscedastic) model is to effectively describe the dependence of volatility on recent returns.We begin with a "naive" approach to these matters.

2.8 example. The data file stock.xls contains weekly data for the logged stock prices `lStock` (T=208). Though the differences, i.e., returns `d_lstock` ($= \Delta \log P_t$) do not pass WN tests (check), for the meantime we assume that the WN assumption is true. In Fig. 2.20,

[11] For a time, we define the volatility (or variability or changeability) of ε_t as e_t^2 (it is close in a sense to the variance of ε_t).

[12] "Many market strategists expect that the market will remain volatile between now and mid-January" or "Stock markets reacted nervously to...".

we can see that volatility[13] (blue line) has indeed the property of persistency – in the vicinity of $t = 100$ the volatility is high for some time whereas it is quite small in stable times.

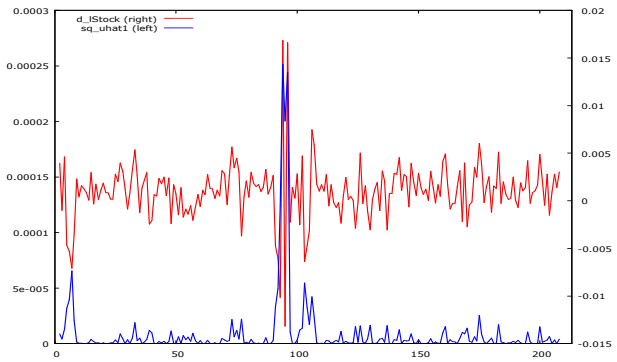


Figure 2.21. The log differences (red) seem to be stationary whereas volatility u_t^2 (= `sq_uhat1`) (blue) has a clustering property

We have already learned that AR process is able to model the phenomenon of persistency which, in our case, may be called „clustering in volatility". Consider, for instance, an AR(1) model of volatility: $u_t^2 = \alpha + \varphi u_{t-1}^2 + w_t$, $w_t \sim WN$:

```
OLS, using observations 3-208 (T = 206)
Dependent variable: sq_uhat1

              coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const        2.40841e-06   1.48446e-06    1.622    0.1063
  sq_uhat1_1   0.737005      0.0473286     15.57     1.51e-036 ***
```

The model[14] $u_t^2 = 0.0000024 + 0.737\, u_{t-1}^2 + w_t$ might be of great interest to an investor wishing to purchase this stock. Suppose an investor has just observed that $u_{t-1}^2 = 0$. In other words, the stock price changed by its average amount in period $t-1$. The investor is interested in predicting volatility in period t in order to judge the likely risk involved in purchasing the stock. Since the error is unpredictable, the investor ignores it (i.e., it is just as likely to be positive as negative). However, the predicted volatility in period t is 2.4e-06. On the other hand, had he observed $u_{t-1}^2 = 0.0001$, he would have predicted volatility in period t to be 0.0000024 + 0.0000737 = 7.6e-05, i.e., 30 times bigger. This kind of information can be incorporated into financial models of investor behavior. ◀◀

So far, we have analysed empirical properties of stock.xls returns. Now we shall discuss in more detail a concept of WN which, as you remember, is defined as a sequence of uncorrelated or independent random variables $\{\varepsilon_t\}$ with zero mean and constant variance σ^2. Thus far, we have not cared much about this "or" but now it becomes important (the sequence of uncor-

[13] That is, squared residuals (`sq_uhat1` =) $\hat{u}_t^2$ of the model $d\_lstock_t = \alpha + u_t$.

[14] Thus we assume that ε_t are not correlated (the series is a WN) but ε_t^2 (to fit the empirical facts) are and the dependence is described by AR(1).

related r.v.'s is called a *weak* WN[15] whereas the sequence of independent r.v.'s a *strong* WN; for normal, or Gaussian, process ε_t both definitions are equivalent). Note that if ε_t is a strong WN then ε_t^2 (and any other function of ε_t) is too.

2.9 example. If ε_t is a strong WN, then

- conditional mean $E(\varepsilon_t \mid \Omega_{t-1}) \equiv 0$
- conditional variance $\mathrm{var}(\varepsilon_t \mid \Omega_{t-1}) = E((\varepsilon_t - E(\varepsilon_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2 \mid \Omega_{t-1}) \equiv \sigma^2$.

Here $\Omega_s = \mathcal{F}(\varepsilon_s, \varepsilon_{s-1}, ...)$ is the set containing all the information on the past of the process; the proof of these features follows from the properties of conditional mean (see PE.I, Lecture Notes, p. 3-2).

2.10 example. If ε_t is a strong WN and $Y_t = \varphi Y_{t-1} + \varepsilon_t$ is a stationary AR(1) process, then

- conditional mean $E(Y_t \mid \Omega_{t-1}) = \varphi Y_{t-1} \not\equiv const$
- conditional variance $\mathrm{var}(Y_t \mid \Omega_{t-1}) = E((\varphi Y_{t-1} + \varepsilon_t - \varphi Y_{t-1})^2 \mid \Omega_{t-1}) \equiv \sigma^2$.

In general, as it follows from the Wold representation theorem, if $\{\varepsilon_t\}$ makes a strong WN, then $\mathrm{var}(Y_t \mid \Omega_{t-1}) = \mathrm{var}(B(L)\varepsilon_t \mid \Omega_{t-1}) \equiv const$. ◀◀

Unfortunately (or fortunately), if $\{\varepsilon_t\}$ makes only a weak WN, the conditional variance $\mathrm{var}(Y_t \mid \Omega_{t-1})$ is not necessarily constant. Now we shall present a model of weak WN process (its variance is constant) such that its conditional variance or *volatility* may change in time[16]. The simplest way to model the above presented clustering phenomenon is to use the ARCH(1) model which describes the (centered) returns ε_t as

$$\begin{cases} \varepsilon_t = \sigma_t w_t \\ \sigma_t^2 = E(\varepsilon_t^2 \mid \Omega_{t-1}) = \omega + \alpha \varepsilon_{t-1}^2 \end{cases} \tag{2.7}$$

where the *innovations* w_t are $(0,1)$ - Gaussian (or Student's or similar symmetric) i.i.d.r.v.'s (that is, w_t is a strong WN), $\omega, \alpha > 0$. It is easy to show that ε_t is again a weak WN, i.e., $E\varepsilon_t \equiv 0, D\varepsilon_t \equiv \omega/(1-\alpha)$ and $\mathrm{cov}(\varepsilon_{t+h}, \varepsilon_t) = 0$ for all $t \geq 0$ and $|h| \geq 1$. In other words, its variance is constant whereas its conditional variance σ_t^2 is not![17] Indeed, the conditional distribution of ε_t given ε_{t-1} is Gaussian with changing variance: $\varepsilon_t \mid \varepsilon_{t-1} \sim N(0, \omega + \alpha\varepsilon_{t-1}^2)$. Note that every trajectory of ARCH process seems to be non-regular but when you take the whole infinite ensamble, the mean value and variance at any time t is the same (i.e., an ARCH pro-

[15] Correlogram and the Ljung-Box test are used to test whether a stationary process is a weak WN.
[16] This allows to model the persistency of volatility.
[17] Thus ε_t is (unconditionally) homoskedastic, but conditionally heteroskedastic. Since σ_t^2 is the one-period ahead forecast variance based on past information, it is called the *conditional* variance.

cess is stationary). On the other hand, the scale σ_t depends on ε_{t-1}, therefore, the broker knowing the history of our process till today can predict tomorrow's variability or volatility of the process (this is important in some cases, for example, in capital asset pricing model (CAPM) and the Black–Scholes option price formula where the volatility of the price of the underlying asset plays a crucial role). ARCH processes are also important in analyzing returns (the main problem in the analysis of these type of financial data is to forecast the volatility of future returns).

The ARCH(1) process can be generalized: the process

$$\begin{cases} \varepsilon_t = \sigma_t w_t \\ \sigma_t^2 = E(\varepsilon_t^2 \mid \Omega_{t-1}) = \omega + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_q \varepsilon_{t-q}^2 \end{cases}$$

is called an ARCH(q) process and the process

$$\begin{cases} \varepsilon_t = \sigma_t w_t \\ \sigma_t^2 = E(\varepsilon_t^2 \mid \Omega_{t-1}) = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \end{cases}$$

a GARCH(p,q) process (the most popular GARCH(1,1) process is just a parsimonious way to rewrite a high order ARCH(q) process; both ARCH and GARCH processes are (weak) WN processes with a special structure of their conditional variance.)

Quite often, the process we want to investigate for the ARCH effects is stationary but not a WN. For example, if $Y_t = \alpha + \varphi Y_{t-1} + \varepsilon_t$, we can first remove the AR(1) part and then analyze the residuals e_t for the presence of, say, ARCH(1) effect. However, in order to properly estimate all the coefficients it is better to carry out the procedure in one step (respective model is called an AR(1) - ARCH(1) model[18]).

- Let ε_t be a weak WN$(0, \sigma^2)$ and consider the model $Y_t = r + \varepsilon_t$ or $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ or $Y_t = \alpha + \varphi Y_{t-1} + \varepsilon_t$ or similar
- To test whether the WN shocks ε_t make an ARCH process, plot a graph of e_t^2 – if ε_t is an ARCH process, this graph must show a clustering property
- To further test whether the shocks ε_t form an ARCH process, test them for normality (the hypothesis must be rejected)
- To further test whether the shocks ε_t form an ARCH process, draw the correlogram of e_t – the correlogram must indicate WN, but that of e_t^2 must not (and it should be similar to the correlogram of AR(p) process)
- To formally test whether the shocks ε_t form an ARCH(q), use the auxiliary regression $e_t^2 = \alpha + \alpha_1 e_{t-1}^2 + ... + \alpha_q e_{t-q}^2 + w_t$; if you reject the hypothesis $H_0 : \alpha_1 = ... = \alpha_q = 0$, ε_t is ARCH(q)

[18] We say that the mean equation is described by AR(1) and variance equation by ARCH(1).

- Instead of using ARCH(q) with big q, often a more parsimonious[19] description of ε_t is given by GARCH(1,1)
- In order to show that the selected ARCH(q) or GARCH(1,1) model is "good", test whether residuals $\hat{w}_t = \hat{\varepsilon}_t / \hat{\sigma}_t$ and $\hat{w}_t^2$ make WN (as they are expected to).

2.8 example (cont.) The traditional OLS method to estimate the parameters ω and α of ARCH(1) does not work because of complicated nature of (2.7); however, GRETL and R have relevant procedures to do the job. Once you have imported stock.xls, create $R_t = \Delta \log P_t$ = `d_lstock`, go to Model| Time series| GARCH...| choose `d_lstock` as dependent variable and GARCH p: 0, ARCH q: 1| OK.

```
Model 1: GARCH, using observations 2-208 (T = 207)
Dependent variable: d_lStock
Standard errors based on Hessian

             coefficient   std. error      z      p-value
  ---------------------------------------------------------
  const       0.00104847   0.000113235   9.259   2.06e-020 ***

  alpha(0)    2.40025e-06  3.90431e-07   6.148   7.86e-010 ***
  alpha(1)    0.659880     0.157144      4.199   2.68e-05  ***
```

Thus, the estimate of the model

$$\begin{cases} d\_lstock_t = r + \varepsilon_t \\ \varepsilon_t \,|\, \Omega_{t-1} \sim N(0, \sigma_t^2) \\ \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{cases} \quad \text{is given by} \quad \begin{cases} \widehat{d\_lstock} = \hat{r} = 0.001 \\ \widehat{\sigma_t^2} = 2.400e-06 + 0.660 e_{t-1}^2 \end{cases}$$

and consists of two parts – the (conditional) mean equation is just a constant 0.001 while the second line presents the ARCH(1) equation for (conditional variance or) volatility. Note that this equation is very close to the "naive" one on p. 2-25. ◀◀

Now it is right time to recall that `d_lstock` is not a WN (see its correlogram). To find the mean equation, go to Model| Time series| ARIMA... and verify that residuals of the ARIMA(6,0,0) model for `d_lstock` form a WN. Save the residuals and create ARCH(1) model for them. Alternatively, go directly to Model| Time series| GARCH..., insert `d_lstock` as dependent variable , go to the lags... box and, after choosing 1 to 6 lags, estimate ARCH(1) model:

```
GARCH, using observations 8-208 (T = 201)
Dependent variable: d_lStock

              coefficient    std. error       z      p-value
  ----------------------------------------------------------
  const         0.00106556   0.000178667    5.964   2.46e-09 ***
  d_lStock_1   -0.0982519    0.0701929     -1.400   0.1616
  d_lStock_2    0.0608469    0.0514942      1.182   0.2374
```

[19] Usually GARCH(p,q) models with small p and q nicely capture persistence of high volatility (volatility clusters), whereas ARCH(q) specifications usually need high orders.

```
d_lStock_3   -0.0122428     0.0464596     -0.2635    0.7922
d_lStock_4   -0.0605324     0.0419533     -1.443     0.1491
d_lStock_5   -0.0201123     0.0422073     -0.4765    0.6337
d_lStock_6    0.0885649     0.0373857      2.369     0.0178   **

alpha(0)      2.25813e-06   4.08361e-07    5.530     3.21e-08 ***
alpha(1)      0.656985      0.180902       3.632     0.0003   ***
```

The upper part of the table stands for the AR(6) mean equation while the last two lines for volatility.

As alternative, we can use a more straightforward approach: go to Model| Time series| ARIMA...| choose `d_lstock` as Dependent variable and ARMA(6,0) model| OK, then in the model's window go to 1) Tests| Autocorrelation (you will get „no autocorrelation" in the residuals) and 2) Tests| ARCH (you will find that ARCH effect is definitely present, thus create an AR(6)-ARCH(1) model). ◀◀

The following example explains how to apply the GARCH procedure.

2.11 example. The data file nyse.txt contains monthly data `StockPrice` from 1952:1 through 1996:1 on a major stock price index provided by the New York Stock Exchange.

```
ols ld_StockPrice 0              # regress log differences on constant only
series sigma = $sigma            # Standard Error of Regression (=0.04)
```

In the Model 1 window, go to Save| Residuals and then Save| Squared residuals; (centered) returns `uhat1` constitute a WN, but the corr elogram and Ljung-Box table shows that `usq1` do not (and what about the normality of `uhat1`?). Thus, we suspect that `uhat1` may be an ARCH process.

To test the hypothesis $H_0: \alpha_1 = ... = \alpha_q = 0$ which means no ARCH in $\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_q \varepsilon_{t-q}^2$, choose the proper AR(q) model of the auxiliary regression $e_t^2 = \alpha + \alpha_1 e_{t-1}^2 + ... + \alpha_q e_{t-q}^2 + w_t$ (proper means minimum AIC and WN residuals w_t); now, to test H_0, use the F - test (or the LM test: save R^2 of the auxiliary regression and test whether $LM = TR^2$ is greater than 5% critical value of χ_q^2).

To perform the above procedure, in the Model 1 window go to Tests| ARCH, where we start with lag order 12 and end with 9 (now the top term `alpha(9)` is significant): at the bottom of the table you will find the lines

```
Null hypothesis: no ARCH effect is present
  Test statistic: LM = 23.4373
  with p-value = P(Chi-Square(9) > 23.4373) = 0.00528544
```

which mean that we must reject no ARCH hypothesis (thus, `ld_StockPrice` is ARCH(9) process; note that the second best is ARCH(3) model).

The ARCH(9) model depends on 10 parameters. A common approach is to use a more parsimoniuos model GARCH(1,1) (in fact, there is no way to test for its accepability):

```
garch 1 1; ld_StockPrice 0                  # to create GARCH(1,1) model
```

```
series ht_garch11 = $h                          # conditional variance
series st_err11 = sqrt(ht_garch11)              # conditional standard error
```

The same result may be achieved from the pop-up menu: in GRETL window, go to Model| Time series| GARCH...| and choose `ld_StockPrice` as Dependent variable (this will create your Model 2).
To get an impression of the accuracy of the models, copy the following lines to GRETL script window:

```
garch 0 3; ld_StockPrice 0                      # to create ARCH(3)
series ht_arch3 = $h
series st_err03 = sqrt(ht_arch3)                # conditional standard error
```

Then, to appreciate difference across models, plot a graph of unconditional standard error `sigma` as well as two conditional variants (for only 1985:01 through 1996:01, see Fig 2.21,left):

```
# gnuplot StockPrice ld_StockPrice --time-series --with-lines --output=display
smpl 1985:01  1996:01
gnuplot sigma st_err11 st_err03 --time-series --with-lines --output=display
```

To forecast Model 2, go to its window, choose Analysis| Forecasts...| and add 11 observations (see Fig. 2.21, right). As expected, the forecast of the WN process `ld_StockPrice` is just its mean while the conditional standard errors tend to the unconditional `sigma`.

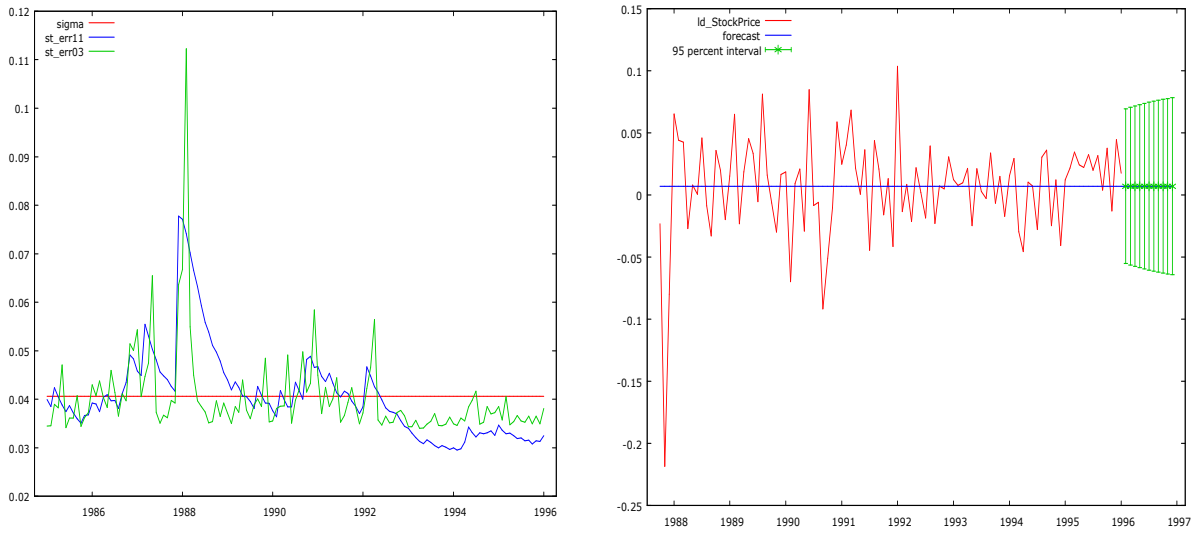


Figure 2.22. The left graph compares three standard errors of `ld_StockPrice`; two conditional variants are clearly past-dependent; the GARCH(1,1) 11-months forecast (right). Note that the extreme value of `ld_StockPrice` at around 1988 implies maximum value of `st_err11` at the same time.

When we say "describe a stationary process Y_t", we mean to express it as a sum of its structural part (like AR or MA) Z_t and a zero mean white noise w_t: $Y_t = Z_t + w_t$. In general, we must then test whether w_t is a strong WN, i.e., to test for the ARCH effects

3. Trend Stationary Time Series[1]

Many forecasting methods are based on the concept that the underlying pattern, provided it exists, can be distinguished from randomness by smoothing or averaging past values. The purpose of smoothing is to eliminate randomness so the pattern can be projected into the future and used as the forecast.

The general mathematical representation of the decomposition approach is as follows:

$$Y_t = f(Tr_t, S_t, \varepsilon_t) \qquad (3.1)$$

where Y_t is the time series value (actual data) at period t
Tr_t a deterministic trend or general movement component
S_t a deterministic seasonal component
ε_t the remaining stationary component.

The exact functional form of (3.1) depends on the decomposition method actually used. A common approach is to assume equation (3.1) has the additive form

$$Y_t = Tr_t + S_t + \varepsilon_t .$$

That is, the trend, seasonal and irregular component are simply added together to give the observed series.

Alternatively, the multiplicative decomposition has the form

$$Y_t = Tr_t \cdot S_t \cdot \varepsilon_t .$$

That is, the trend, seasonal and irregular component are multiplied together to give the observed series.

In both, additive and multiplicative, cases
the series Y_t is called a trend stationary (TS) series

This definition means that after removing a deterministing part from a TS series what remains is a **stationary** series. Thus, if our historical data end at $t = T$ and the process under consideration is additive, we can forecast the deterministic part simply by taking $\widehat{Tr}_{T+h} + \hat{S}_{T+h}$ provided[2] we know the analytic expression for both trend and seasonal part (recall that we have already learned how to obtain the forecast $\varepsilon_{T,T+h}$). Note that not all time series are TS. Actually, most economical series are described by another, DS, model (see Ch.4).

An additive model is appropriate if the magnitude of the seasonal fluctuations does not vary with the level of the series. But if the seasonal fluctuations increase and decrease proportional-

[1] Procesai, kurių nuokrypiai nuo trendo stacionarūs.
[2] Most smoothing methods (for example, moving averages) do not give that expression.

ly with increases and decreases in the level of the series, then a multiplicative model is appropriate. Multiplicative decomposition is more prevalent with economic series because most seasonal economic series do have seasonal variations which increases with the level of the series.

Rather than choosing either an additive or multiplicative decomposition, we could use a transformation – very often the transformed series can be modeled additively, when the original data are not additive. Logarithms, in particular, turn a multiplicative relationship into an additive relationship: if $Y_t = Tr_t \cdot S_t \cdot \varepsilon_t$ then

$$\log Y_t = \log Tr_t + \log S_t + \log \varepsilon_t .$$

So we can fit a multiplicative relationship by fitting a more convenient additive relationship to the logarithms of the data and then to move back to original series by exponentiating.

There are various ways to estimate the trend Tr_t at time t, but a relatively simple procedure, which is available in R (for example, with `decompose`) and does not assume any specific form of Tr_t is to calculate a moving average centered on t. A moving average is an average of a specified number of time series values around each value of t in the time series, with the exception of the first few and last few terms. If the time series has no seasonal component, in order to estimate the trend we can take any odd number, for example, if $l = 3$ and we estimate an additive model,

$$\widehat{Tr_t} = (Y_{t-1} + Y_t + Y_{t+1})/3 \quad \text{(two-sided averaging)}$$
$$\widehat{Tr_t} = (Y_{t-2} + Y_{t-1} + Y_t)/3 \quad \text{(one-sided averaging)}$$

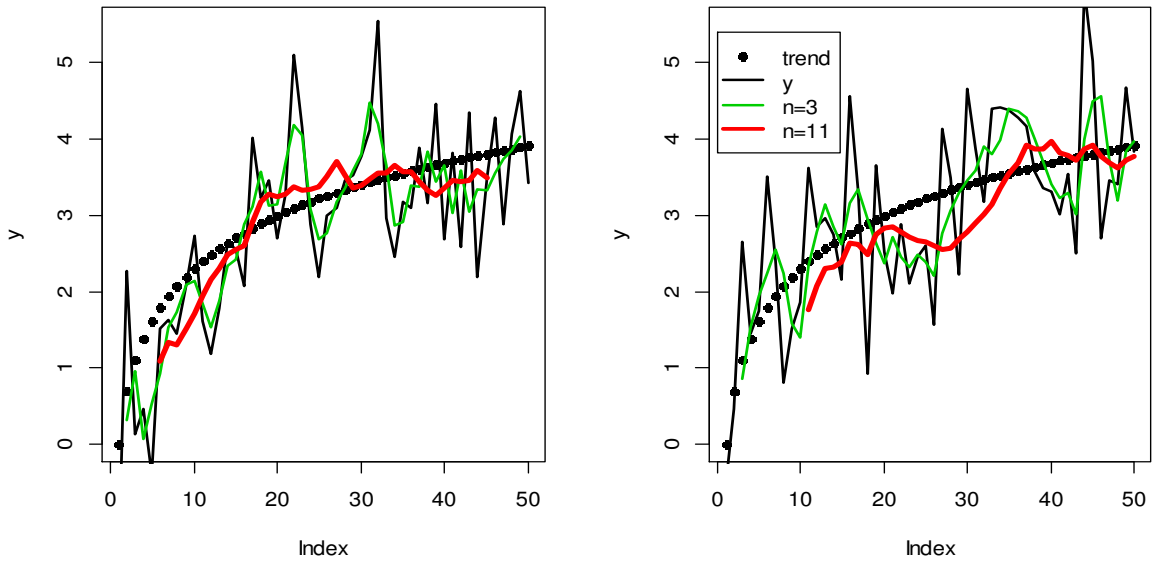


Figure 3.1. (Artificial data) Two-sided averaging (left) and one-sided averaging (right; in both cases the estimates loose some end points); the 11-points averaging (red) produces a smoother estimate but in any case they do not allow us to forecast the time series.

If the time series contains a seasonal component and we want to average it out, the length of the moving average must be equal to the frequency (thus, for monthly series, we take $l = 12$). However, there is a slight snag. Suppose our time series begins at January (t = 1) and we average January up to December (t = 12). This average corresponds to a time t = 6.5, between June and July. When we come to estimate seasonal effects, we need a moving average at integer times. This can be achieved by averaging the average of January up to December and the average of February (t = 2) up to January (t = 13). This average of two moving averages corresponds to t = 7, and the process is called centring. Thus the trend at time t can be estimated by the centered moving average

$$\widehat{Tr}_t = \frac{\left(Y_{t-6} + \ldots + Y_{t+5}\right)/12 + \left(Y_{t-5} + \ldots + Y_{t+6}\right)/12}{2} = \frac{(1/2)Y_{t-6} + Y_{t-5} + \ldots + Y_{t+5} + (1/2)Y_{t+6}}{12}$$

where t = 7,...,n − 6. By using the seasonal frequency for the coefficients in the moving average, the procedure generalises for any seasonal frequency (e.g., quarterly series), provided the condition that the coefficients sum to unity is still met.

An estimate of the monthly additive effect S_t at time t can be obtained by subtracting $\widehat{Tr}_t$:

$$\hat{S}_t = Y_t - \widehat{Tr}_t$$

By averaging these estimates of the monthly effects for each month (January, February etc), we obtain a single estimate of the effect for each month

It is common to present economic indicators, such as unemployment percentages, as seasonally adjusted series. This highlights any trend that might otherwise be masked by seasonal variation attributable, for instance, to the end of the academic year, when school and university leavers are seeking work. If the seasonal effect is additive, a seasonally adjusted series is given by $Y_t - \hat{S}_t$.

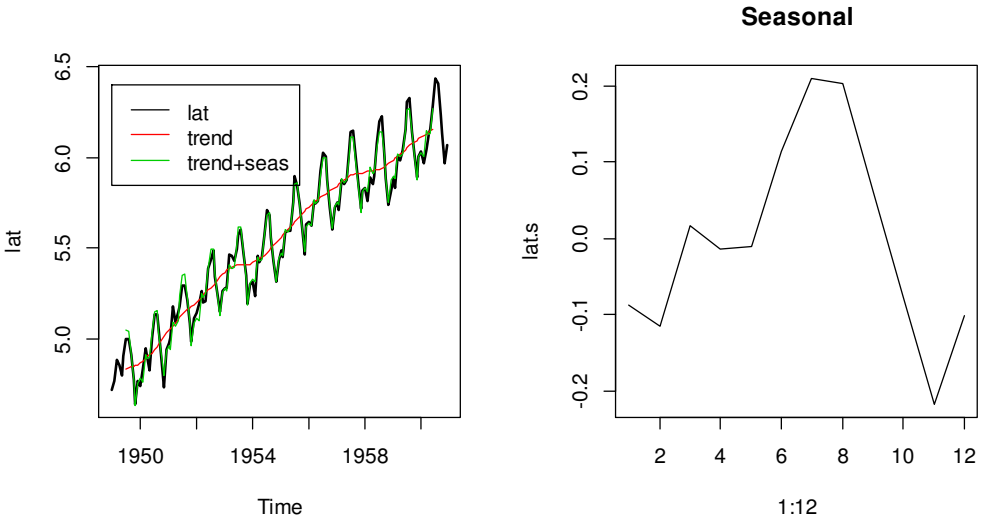


Figure 3.2. Logarithms of the `airpass` time series, `lat`, from the `fma` package and its trend (left); seasonal component (right)

The above described moving average procedure usually quite succesfully describes the times series in question, however, it does allows to forecast it.

To decide upon the mathematical form of a trend, one must first draw the plot of the time series. If the behavior of the series is rather „regular“, one can choose a parametric trend – usually it is a low order polynomial in t, exponential, inverse or similar functions. The most popular method to estimate the coefficients of the chosen function is OLS. On the other hand, if it is not quite clear which concrete function to use, one can look for some implicit function described not by a formula but by a certain computational algorithm. A useful variant of this procedure is presented in Sec. 3.2. In any case, the smoothing method is acceptable if the residuals $\hat{\varepsilon}_t = Y_t - \widehat{Tr}_t - \hat{S}_t$ constitute a stationary process. If we have a few competing trends, the best one can be chosen by AIC, BIC or similar criterions. An alternative approach is to create models for all but some T_0 end points and then to choose the model whose forecast fits original data best (to select the model, one can use such characteristics as

Root Mean Square Error $\quad \mathrm{RMSE} = \sqrt{(1/T_0)\sum_{t=T-T_0}^{T} \hat{\varepsilon}_t^2}$

Mean Absolute Percentage Error $\quad \mathrm{MAPE} = (1/T_0)\sum_{t=T-T_0}^{T} \left|\hat{\varepsilon}_t / Y_t\right| \times 100$

and like).

3.1. The Global Method of Decomposition and Forecasting - OLS

The well known OLS method estimates the coefficients of, say, quadratic trend $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$ by minimizing $RSS(\beta_0, \beta_1, \beta_2) = \sum_{t=1}^{T} (Y_t - (\beta_0 + \beta_1 t + \beta_2 t^2))^2$. Note that if the value of the last Y_T for whatever reason deviates much from the trend, this may considerably change the estimates $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\beta}_2$ and, therefore, the fitted value of the first $\hat{Y}_1$. This is why we term the method global. One local method which little alters the estimate of Y_1, following a change in a remote Y_T, will be examined in the next section.

3.1 example. We shall examine the number of international passenger bookings (in thousands) per month on an airline (PanAm) in the United States for the period 1949:1–1960:12 (the data is available as AP.txt in the PEdata folder). We shall create three models:

Model 1 $\quad AP_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$

Model2 $\quad AP_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma_1 dm1_t + ... + \gamma_{11} dm11_t + \varepsilon_t$

Model3 $\quad \log AP_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma_1 dm1_t + ... + \gamma_{11} dm11_t + \varepsilon_t$

where $t = 1,...,144$ is for the time trend, $dm1$ is a dummy variable for the first month (January), $dm2$ is for February etc. Recall that in order to avoid the dummy variable trap (see PE.I, Lecture Notes, Sec. 4.5), we have to exclude one dummy variable (for example, $dm12$) from

our regression models (γ_1 is the extra number of passengers in January compared with the base month of December, γ_2 is the extra number of passengers in February etc).

After importing AP, to create the trend variable ($t =$) `time`, go to Add| Time trend, then select `time` and, to create t^2, go to Add| Squares of selected variables. To create Model 1, go to Model| Ordinary Least Squares, select `AP` as dependent and `time` and `sq_time` as independent variables:

```
^AP = 112 + 1.64*time + 0.00701*sq_time
     (11.4)(0.362)      (0.00242)
```

(below the coefficients, one can see the standard errors; clearly, both `time` and `sq_time` are significant).

To visualize this model, in the Model 1 window, go to Graphs| Fitted, actual plot| Against time (see Fig. 3.3, left). To capture seasonal variations, in the GRETL window, go to Add| Periodic dummies, then to Model| Ordinary Least Squares, select `AP` as dependent, `time`, `sq_time`, and `dm1`, ..., `dm11` as independent variables, then in the Model 2 window, go to Graphs| etc.

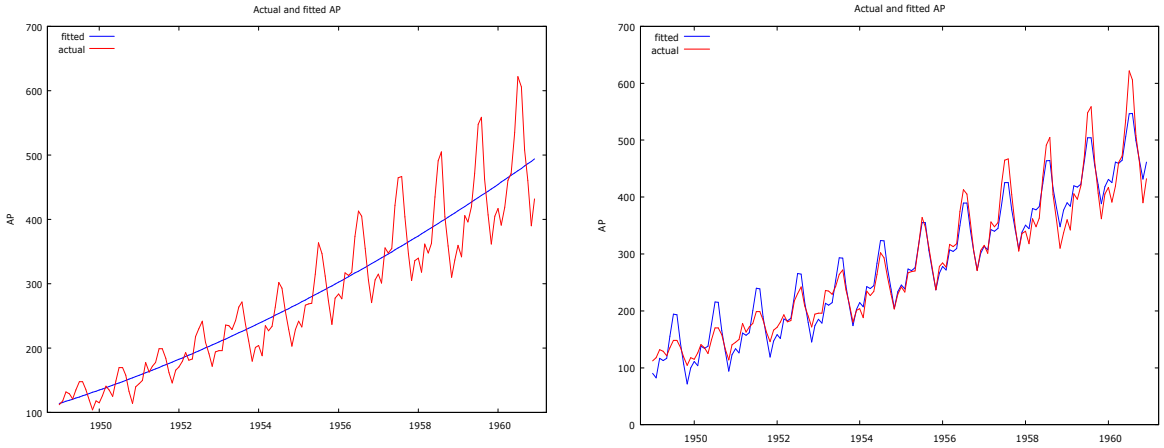


Figure 3.3. `AP` (red) and fitted values (blue) from Model 1 (left) and Model 2 (right)

Additive Model 2 describes seasonality in a wrong way (the `AP`'s seasonal fluctuations increase together with level but in our fit they do not). To correct for multiplicativity, we make Model 3 for logarithms (but first, you have to create logarithms: select `AP` and go to Add| Logs of selected variables; a new variable `l_AP` will be created).

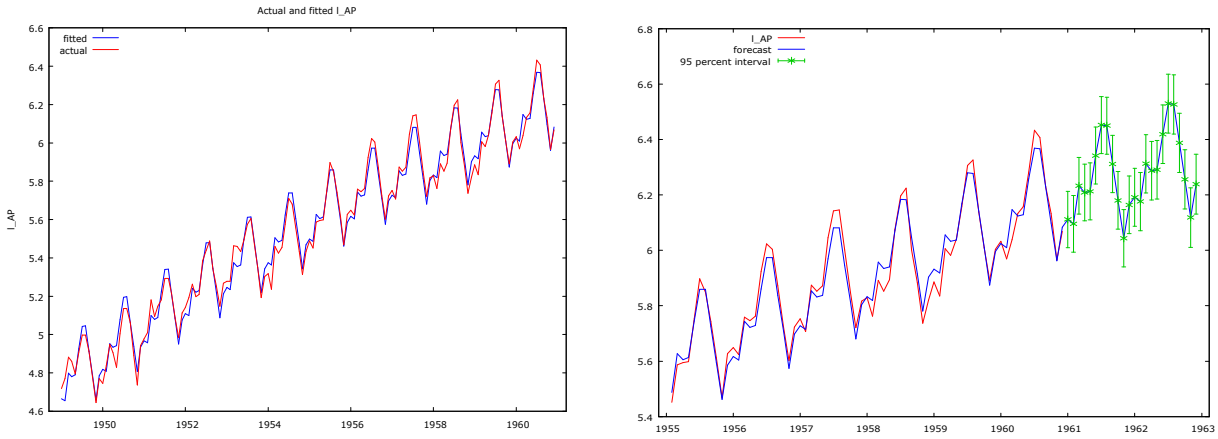
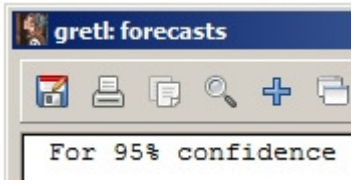


Figure 3.4. l_AP (red) and fitted values (blue) from Model 3 (left) and forecast made by Model 3 (right)

Now the fit is much better (see Fig. 3.4, left); to use this final Model 3 for the 24-months-ahead forecast of l_AP , go, in GRETL window, to Data| Add observations| 24, then, in the Model 3 window, go to Analysis| Forecasts| OK (see Fig. 3.4, right).

To plot the forecast not for l_AP but for the original AP, note that this graph is accompanied by the forecast table

```
For 95% confidence intervals, t(130, 0.025) = 1.978

                 l_AP     prediction     std. error        95% interval

 1955:02      5.451038      5.487067
 1955:03      5.587249      5.627105
 1955:04      5.594711      5.605647
 ..............................................
 1960:11      5.966147      5.961961
 1960:12      6.068426      6.083059
 1961:01                    6.111356      0.051549     6.009372 - 6.213339
 1961:02                    6.096019      0.051622     5.993890 - 6.198147
 ..............................................
 1962:11                    6.118544      0.054246     6.011225 - 6.225864
 1962:12                    6.238611      0.054409     6.130969 - 6.346253
```



To save prediction, click on the blue cross on the top and save it as l_AP_fore. To get back to AP, go first to Save| Standard error of the regression (as sigma_3) and then, in GRETL window, go to Add| Define new variable| and apply the formula AP_fore = exp(l_AP_fore + sigma_3^2/2). Now, plot both AP and AP_fore (see Fig. 3.5):

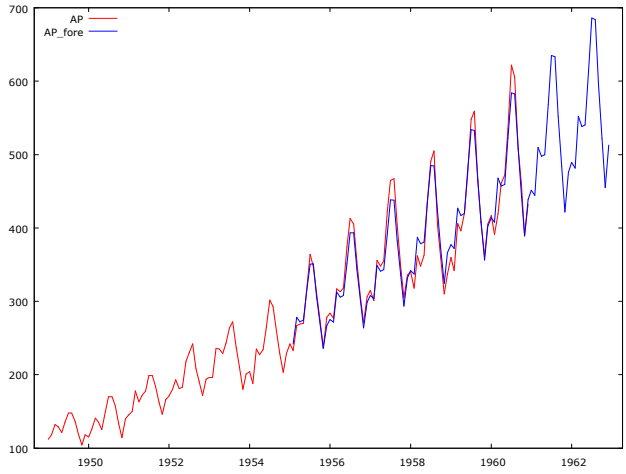


Figure 3.5. `AP`, its Model 3 fit and 24-months-ahead forecast `AP_fore`.

3.1 exercise. Do the residuals of Model 3 form a stationary process? White noise?

3.2. One Local Method of Decomposition and Forecasting

Here we shall present a short introduction to the exponential smoothing. Exponential smoothing is a technique that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts.

Simple Exponential Smoothing

As is common, we state the exponential smoothing procedure as an algorithm for converting the observed series[3], Y_t, $t=1,...,T$ into a smoothed series, $\hat{Y}_t$, $t=1,...,T$, and forecasts, $\hat{Y}_{T+h,T}$:

1. Initialize at $t=1$: $\hat{Y}_1 = Y_1$
2. Update: $\hat{Y}_t = \alpha Y_t + (1-\alpha)\hat{Y}_{t-1}$, $t=2,...,T$.
3. Forecast: $\hat{Y}_{T+h,T} = \hat{Y}_T$, $h=1,2,\ldots$.

We call $\hat{Y}_t$ the estimate of the *level* at time *t*. The smoothing parameter α is in the unit interval, $\alpha \in [0,1]$. The smaller α is, the smoother the estimated level. As α approaches 0, the smoothed series approaches constancy, and as α approaches 1, the smoothed series approaches point-by-point interpolation. Typically, the more observations we have per unit of calendar time, the more smoothing we need; thus we would smooth weekly data (52 observations per year) more than quarterly data (4 observations per year). There is no substitute, how-

[3] Common practice is to apply the exponential smoothing "always" though, in fact, it is applicable only when Y_t is ARIMA(0,1,1). Namely this ARIMA structure allows us to forecast the time series.

ever, for a trial-and-error approach involving a variety of values of the smoothing parameter (on the other hand, check `ets` function from the forecast package where α is chosen on an AIC basis).

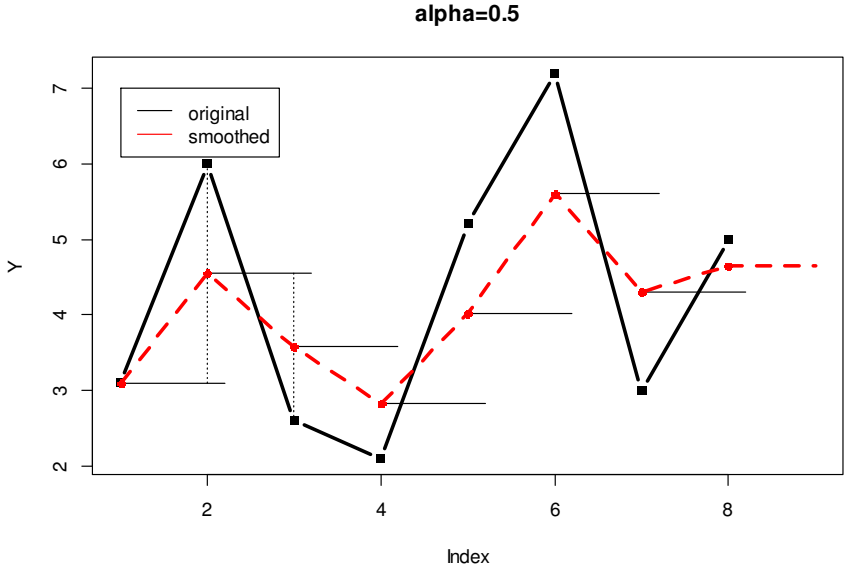


Figure 3.6. The smoothed curve $\hat{Y}_t$ (red) is obtained as the average of the tth value of the original series Y_t, $t = 1,...,8$ (black) and the previous value of the smoothed series (the forecast of Y_t is presented by the red horizontal broken line on the right)

It is not immediately obvious that the algorithm we just described presents a one-sided moving average with exponentially declining weights. To sort it out, start with the basic recursion, $\hat{Y}_t = \alpha Y_t + (1-\alpha)\hat{Y}_{t-1}$ and substitute recursively backward for $\hat{Y}_{t-1}$ which finally yields $\hat{Y}_t = \sum_{j=0}^{t-1} w_j Y_{t-j}$ where $w_j = \alpha(1-\alpha)^j$ (plot w_j).

Holt's Linear Method

Now imagine that we have not only a slowly evolving local level, but also a trend with a slowly evolving local slope. Then the optimal smoothing algorithm is as follows:

1. Initialize at $t = 2$: $\hat{Y}_2 = Y_2$, $F_2 = Y_2 - Y_1$
2. Update: $\hat{Y}_t = \alpha Y_t + (1-\alpha)(\hat{Y}_{t-1} + F_{t-1})$, $0 < \alpha < 1$
   $F_t = \beta(\hat{Y}_t - \hat{Y}_{t-1}) + (1-\beta)F_{t-1}$, $0 < \beta < 1$, $t = 3,...,T$.
3. Forecast: $\hat{Y}_{T+h,T} = \hat{Y}_T + hF_T$

where $\hat{Y}_t$ is the estimated, or smoothed, level at time t, and F_t is the estimated slope at time t. The parameter α controls smoothing of the level, and β controls smoothing of the slope. The h-step-ahead forecast simply takes the estimated level at time T and augments it with h times the estimated slope at time T.

Holt-Winters' Trend and Seasonality Method

If the data has no trend or seasonal patterns, then simple exponential smoothing is appropriate. If the data exhibit a linear trend, then Holt's linear method is appropriate. But if the data are seasonal, these methods on their own cannot handle the problem well. A method known as *Holt-Winters method* is based on three smoothing equations—one for the level, one for trend, and one for seasonality. It is similar to Holt's linear method, with one additional equation for dealing with seasonality. The formulas are a bit involved, so we skip them. ◀◀

The `ets` function from `forecast` package in R presents a fully automated procedure (the best model is selected according to its AIC) based on the exponential moving average filter. We shall smooth both the logged and original data:

```
library(forecast)
library(fma)
data(airpass)
par(mfrow=c(1,2))
ap.log <- ets(log(airpass))
plot(forecast(ap.log),include=36)
ap <- ets(airpass)
plot(forecast(ap),include=36)
```

The graphs of both forecasts are given in Fig. 3.7.

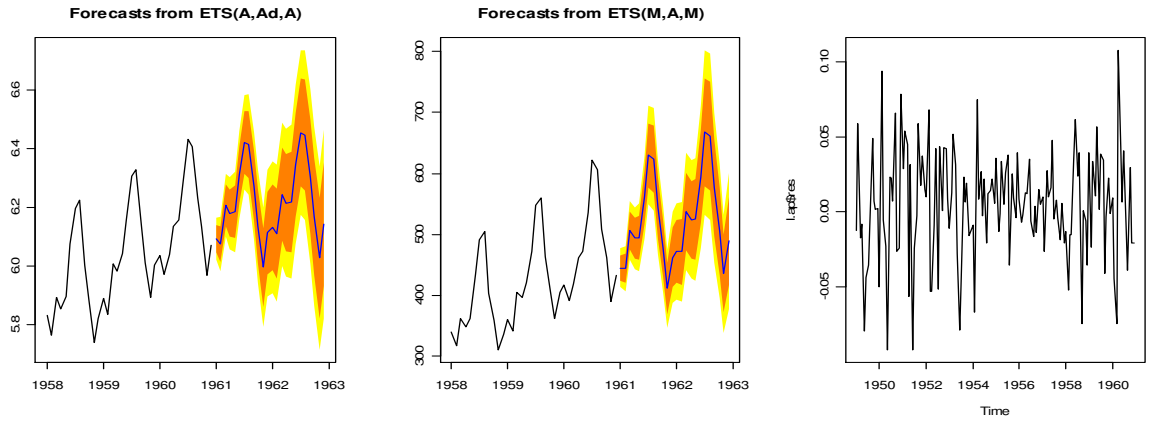


Figure 3.7. Holt-Winters' forecasts of `log(airpass)` (left) and `airpass` (center); residuals of the first model make a stationary process (right)

Recall that this decomposition is valid only if the irregular part (residuals) of our model make a stationary process. In the logged case, the residuals are plotted on the right graph in Fig. 3.2, they seem to form a stationary process.

The h-step-ahead forecast to an additive TS time series $Y_t = Tr_t + S_t + \varepsilon_t$, $t = 1,...,T$, is given by $Y_{T+h,T} = \widehat{Tr}_{T+h} + \hat{S}_{T+h}$, provided ε_t is a WN. If the residuals $\hat{\varepsilon}_t$ constitute a more complicated stationary process (AR, MA, or ARMA), the forecast should take into account their structure (see Ch. 2).

There are many more R functions for decomposition and/or smoothing: `StructTS`, `decompose`, `stl`, `tsSmooth`, `ets`, `ma`, `ses`, `lowess` ... etc (search with the lines `help.`

`search("smoothing")` or `RsiteSearch ("smoothing")` or similar). However, most of them do not allow to forecast the series under consideration.

3.2 example. Below, we have plotted two simulated TS series and their common deterministic trend (black) as well as two estimates of the trend obtained with the `decompose` function (red). This function splits a time series into seasonal, trend and irregular components using moving averages[4]. Note that despite the fact that that the estimates of the trend are close to the true trend, they do not allow to forecast the series.

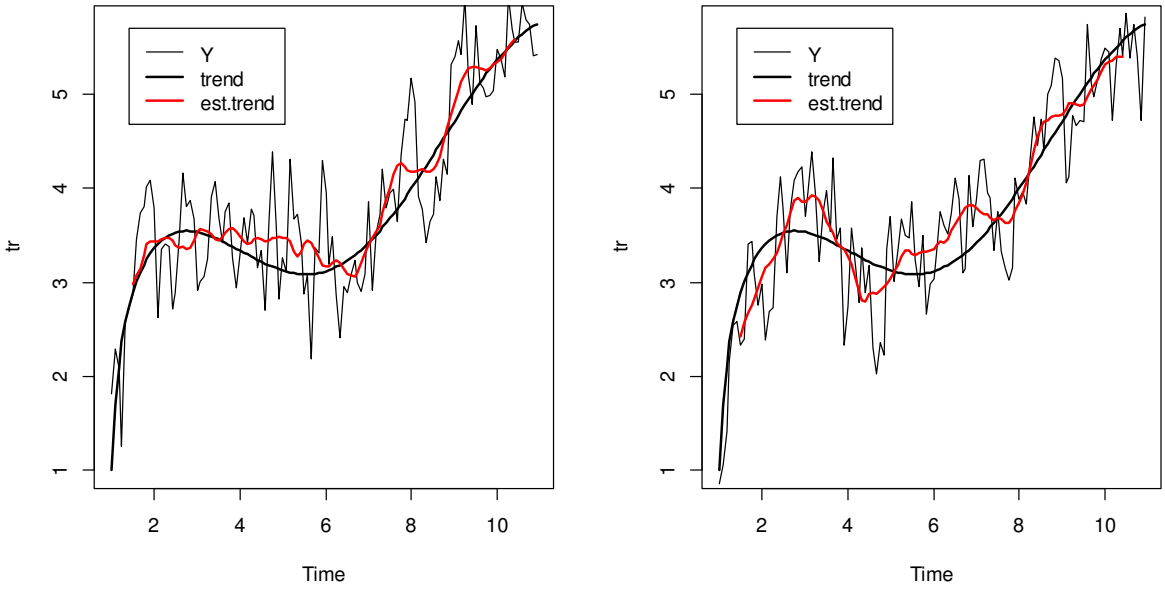


Figure 3.8. Two simulated time series with the same trend (black); estimated trends (red) are close to the true one.

Revision questions

.....

[4] This moving average procedure has little in common with the MA time series.

4. Difference Stationary Time Series[1]

For a long time each economic time series used to be decomposed into a deterministic trend and a stationary process, i.e., it was assumed that all time series are TS series. However, later it appeared that most economic series belong to another class of **difference stationary** series.

- If Y_t is stationary or its residuals ε_t in the decomposition $Y_t = Tr_t + S_t + \varepsilon_t$ are stationary, then Y_t is called Trend Stationary (or TS) series;
- If Y_t is not a TS series but its differences $(\Delta Y)_t = Y_t - Y_{t-1}$ are stationary, then it is called Difference Stationary (or DS) series.

4.1 example. Recall that the AR(1) process $Y_t = \varphi Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, is stationary if $|\varphi| < 1$ and, therefore, under this condition, it is a TS series. However, if $\varphi = +1$, the process is no longer stationary. Indeed, $Y_t = 1 \cdot Y_{t-1} + \varepsilon_t = (Y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \ldots = Y_0 + \varepsilon_1 + \ldots + \varepsilon_t$ (here Y_0 is nonrandom initial value), therefore $EY_t \equiv Y_0$, but $\operatorname{var} Y_t = \sigma_\varepsilon^2 t \neq const$. On the other hand, $\Delta Y_t = \varepsilon_t$ is stationary, thus *a random walk* $Y_t = Y_0 + \varepsilon_1 + \ldots + \varepsilon_t$ (cf. Computer Labs, p. 1-3) is a DS process. ◀◀

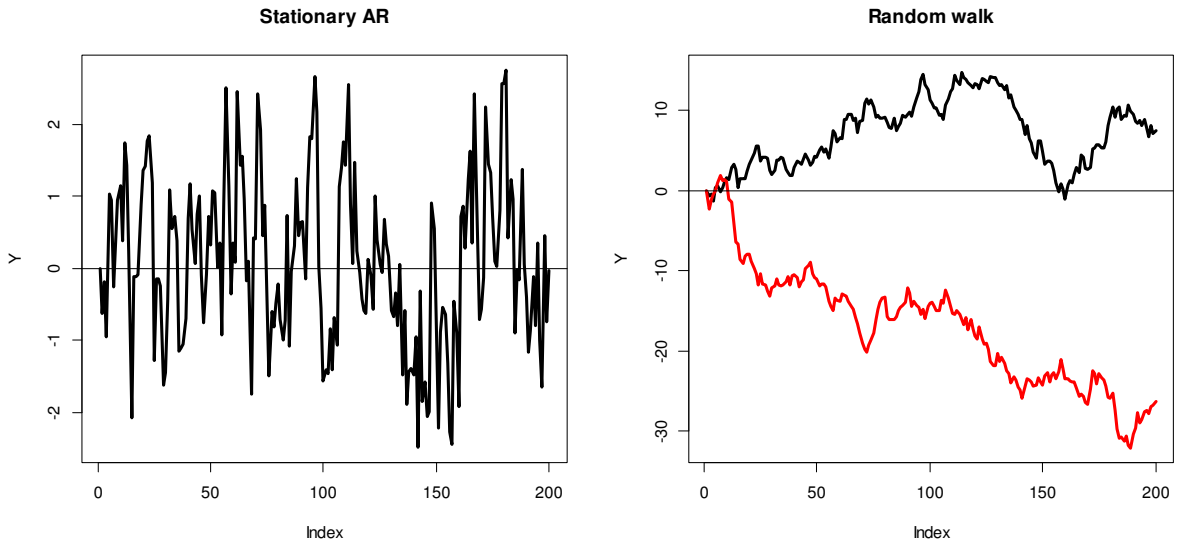


Figure 4.1. One trajectory of a stationary AR(1) process with $\varphi = 0.6$ (left) and two trajectories of random walk (right; note long excursions up and down and no mean reverting tendency)

[1] Procesai, kurių skirtumai stacionarūs.

4.1 exercise. Let $|\varphi|<1$. Is the AR(1) proces $Y_t = \alpha + \varphi Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, a TS process? Is the process $Y_t = \alpha_0 + \alpha_1 t + \varphi Y_{t-1} + \varepsilon_t$ (or $Y_t - \beta_0 - \beta_1 t = \varphi(Y_{t-1} - \beta_0 - \beta_1(t-1)) + \varepsilon_t$) a TS process? What are its trajectories like? Is $Y_t = \alpha + Y_{t-1} + \varepsilon_t$ $(= Y_0 + \alpha t + \varepsilon_1 + ... + \varepsilon_t)$ a DS process (it is called a *random walk with a drift* α)? Sketch its trajectory. ◀◀

It will be useful to have three expresions for the TS series $Y_t = \alpha_0 + \alpha_1 t + \varphi Y_{t-1} + \varepsilon_t$, $|\varphi|<1$. This series can be rewritten as $Y_t - \beta_0 - \beta_1 t = \varphi(Y_{t-1} - \beta_0 - \beta_1(t-1)) + \varepsilon_t$ where $\beta_0 = (\alpha_0(1-\varphi) - \alpha_1\varphi)/(1-\varphi)^2$ and $\beta_1 = \alpha_1/(1-\varphi)$; thus, the deviations of Y_t from the straight line $\beta_0 + \beta_1 t$ are stationary, i.e., Y_t is TS process. The process can also be expressed as

$$\begin{cases} Y_t = \beta_0 + \beta_1 t + u_t \\ u_t = \varphi u_{t-1} + \varepsilon_t \end{cases}$$

(solve the first equation for u_t and substitute it to the second line). In words - Y_t is a straight line plus stationary AR(1) shocks, i.e., it is again a TS process (sketch some paths).

Another way to define DS series is to say that it contains a **stochastic** or random **trend**. Indeed, the above given TS series can be expressed as $Y_t = Tr_t + u_t$ (here Tr_t is a deterministic function and u_t is stationary) whereas the random walk $Y_t = (Y_0 + \alpha t + \varepsilon_1 + ... + \varepsilon_{t-1}) + \varepsilon_t$ has a random "trend" (namely, the expression in parenthesis).

4.2 example. Stock prices on the NYSE

The data file nyse.txt contains monthly data from 1952:1 through 1996:1 on a major stock price index `sp` (=StockPrice) provided by the New York Stock Exchange (NYSE). The price index is a value-weighted one. Figure 4.2 is a time series plots of the index $\texttt{sp}_t$, its logarithm $\texttt{l\_sp}_t = \log \texttt{sp}_t$, and differences of logarithms $\texttt{d\_l\_sp}_t = \Delta\texttt{l\_sp}_t = \texttt{l\_sp}_t - \texttt{l\_sp}_{t-1}$ which are roughly the monthly percentage **change** of $\texttt{sp}_t$. It is common to take the natural logarithm of the time series if it seems to be growing over time. If a series `sp` is growing at a roughly constant rate (and lies on an exponent), then the time series plot of $\texttt{l\_sp} = \log \texttt{sp}$ will approximate a straight line (see Fig. 4.2, top right).

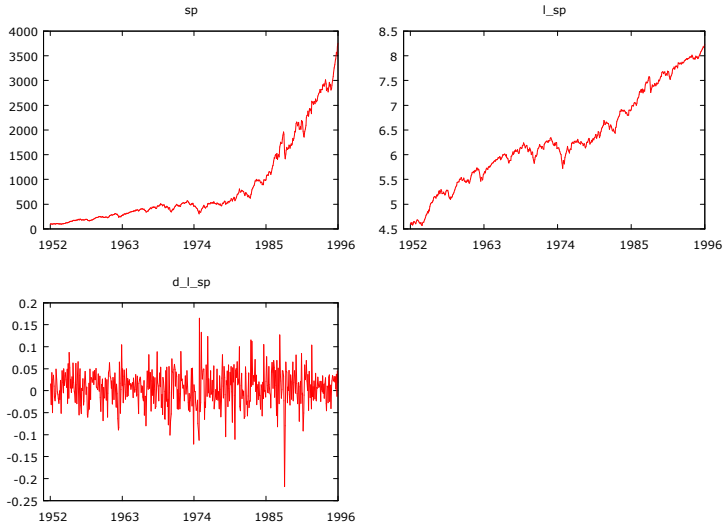
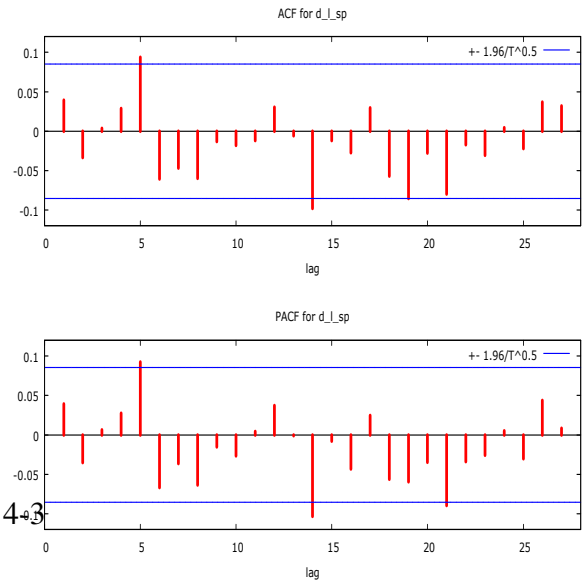


Figure 4.2. NYSE stock price index (montly data from 1952 through 1995) (top, left), logarithm of the index (top, right), and differences of logarithms (bottom, left)

Note that the behavior of `d_l_sp` looks very different from `l_sp`: the trend behavior has disappeared completely. The figure indicates that the change in the stock price each month tends to be small, although there is considerable variability to this growth rate over time. In some of the months, the NYSE stock price index increased by over 5%. In October 1987, it fell by over 20%.

The common property of time series data is the existence of correlation across observations. The stock price index today, for example, is highly correlated with its value last month, i.e., the variable "stock price" is correlated with the variable "stock price lagged one period". In fact, if we calculate the correlation between the stock price and lagged stock price we obtain 0.999044. Yet, if we calculate the correlation between the **change** in the stock price index and the change in the stock price index lagged once, we obtain 0.039. These findings make intuitive sense. Stock markets change only slowly over time; even in bear markets they rarely fall by more than a few percent per month. Consequently, this month's stock price tends to be quite similar to last month's and both are highly correlated. Yet, the return on the stock market is more erratic. This month's and last month's **return** can be quite different, as reflected in the near-zero correlation between them.



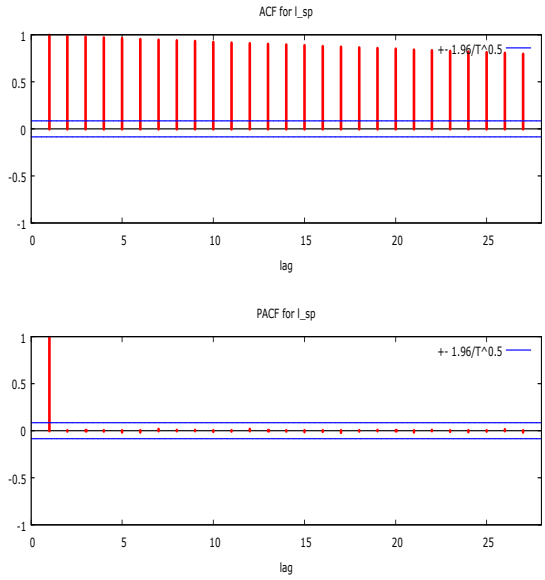


Figure 4.3. Correlogram of `l_sp` (left; resembles the one of AR(1)) and `d_l_sp` (right; resembles WN)

This pattern is common to many financial time series Y: the series itself has autocorrelations near one, but the change in the series has autocorrelations that are much smaller (often near zero). Therefore, if you knew past values of the stock price index, you could make a very good estimate of what the stock price index was this month. However, knowing past values of stock returns will not help you predict the return this month. Y is a **nonstationary** series while ΔY is unpredictable **stationary** white noise.

```
Model 1: ARMA, using observations 1952:01-1996:01 (T = 529)
Dependent variable: l_sp
             coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const       6,41022      0,0411527     155,8    0,0000  ***
  phi_1       0,999715     0,00192066    520,5    0,0000  ***
```

Figure 4.3 suggests and the printout of Model 1: $Y_t = \alpha + \varphi_1 Y_{t-1} + \varepsilon_t$ (see above) confirms that φ is close to (and probably equals) 1, that is, Y_t $(= \alpha + 1 \cdot Y_{t-1} + \varepsilon_t)$ is a random walk.

4.1. Unit Roots

Now we shall describe close relationship between random walks and processes with unit roots.

Definition. We say that an AR(1) process $Y_t = \alpha + \varphi Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, has a unit root if $\varphi = 1$.

Recall that if $|\varphi| < 1$, the process is stationary. The unit root process $Y_t = \alpha + Y_{t-1} + \varepsilon_t$ (it can be rewritten as $\Delta Y_t = \alpha + \varepsilon_t$) is not stationary, it is also referred to as the **random walk** model. The coefficient α is called a **drift**. If Y has a unit root, then ΔY will be stationary. For this reason,

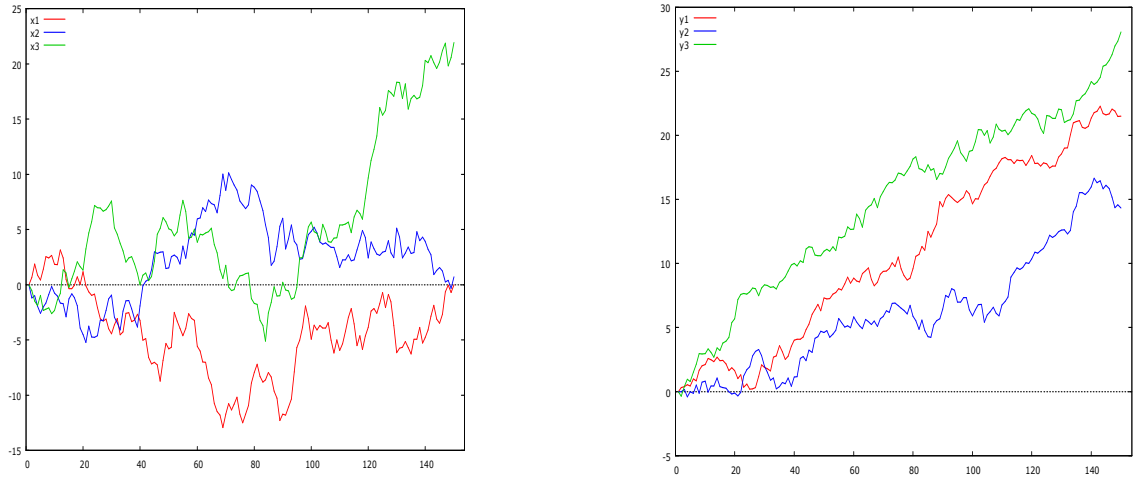


Figure 4.4. Three paths of simulated unit root process (or random walk) with $\alpha = 0$ (left) and three paths of simulated unit root process (or random walk) with a drift $\alpha = 0.1$ (right)

series with unit roots are often referred to as **difference stationary (DS)** series. A series which becomes stationary after first differencing is said to be **integrated of order one** and denoted $I(1)$. If ΔY_t is described by a stationary ARMA(*p, q*) model, we say that Y_t is described by an autoregressive *integrated* moving average (ARIMA) model of order *p*, 1, *q*, or in short an ARIMA(*p*, 1, *q*) model. The symbol $I(0)$ is used to denote a stationary series (for more details see 4.3. Appendix).

The final point can be seen most clearly by subtracting Y_{t-1} from both sides of the equation in the AR(1) model $Y_t = \alpha + \varphi Y_{t-1} + \varepsilon_t$, yielding:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t$$

where $\rho = \varphi - 1$. Note that, if $\varphi = 1$, then $\rho = 0$ and the previous equation can be written solely in terms of ΔY_t, implying that ΔY_t fluctuates randomly around α. For future reference, note that we can test for $\rho = 0$ to see if a series has a unit root. Furthermore, a time series will be stationary if $-1 < \varphi < 1$ which is equivalent to $-2 < \rho < 0$. We will refer to this as the **stationarity condition**.

The AR(1) model can be interpreted as a simple regression model where last period's Y_{t-1} is the explanatory variable. However, it is possible that more lags of *Y* should be included as explanatory variables in order to make the remainder white noise. This can be done by extending the AR(1) model to the autoregressive of order *p*, AR(*p*), model:

$$Y_t = \alpha + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t,\ \varepsilon_t \sim WN(0, \sigma^2), \qquad (4.1)$$

for $t = p+1, ..., T$. In discussing unit root behavior it is convenient to subtract Y_{t-1} from both sides of the previous equation. With some rearranging we obtain

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t \qquad (4.2)$$

where the coefficients in this regression, $\rho, \gamma_1, ..., \gamma_{p-1}$ are simple functions of $\varphi_1, ..., \varphi_p$. For instance, $\rho = \varphi_1 + ... + \varphi_p - 1$. Note that this is identical to the AR(p) model, but is just written differently. Hence we refer to both previous equations as AR(p) models.

The key points here to note are that the above equation is still in the form of a regression model **and $\rho = 0$ implies that the AR(p) time series Y contains a unit root; on the contrary, if $-2 < \rho < 0$, then the series is stationary**[2]. Looking at the previous equation with $\rho = 0$ clarifies an important way of thinking about unit root series which we have highlighted previously: if a time series contains a unit root then a regression model involving only ΔY is appropriate (i.e., if $\rho = 0$, then the term Y_{t-1} will drop out of the equation and only terms involving ΔY or its lags appear in the regression). It is common jargon to say that "if a unit root is present, then the data can be differenced to induce stationarity".

- The AR(p) process has a (first order) unit root if 1) it is not stationary, but 2) its differences are.
- In AR(1) case this is equivallent to $\varphi = +1$ or $\rho = 0$.
- In AR(p) case this is equivallent to $\varphi_1 + ... + \varphi_p = 1$ (why?) or $\rho = 0$.

As we will discuss in the next chapter, with the exception of a case called cointegration, we do not want to include unit root variables in regression models. This suggests that, if a unit root in Y is present, we will want to difference it and use ΔY. In order to do so, we must know first if Y has a unit root. In the past, we have emphasized that drifting unit root series exhibit trend behavior (see Fig. 4.4, right). Does this mean that we can simply examine time series plots of Y for such trending to determine if it indeed has a unit root? The answer is no. To explain why, let us examine three models:

1. Random walk with a drift : $Y_t = \delta + Y_{t-1} + \varepsilon_t = Y_0 + \delta t + \varepsilon_1 + ... + \varepsilon_t, \varepsilon_t \sim WN$ (DS process with a stochastic trend)
2. The process with a linear trend and WN disturbances: $Y_t = \alpha + \delta t + \varepsilon_t, \varepsilon_t \sim WN$ (TS process with a deterministic trend)
3. The process with a linear trend and AR(1) disturbances:

$$\begin{cases} Y_t = \alpha + \delta t + u_t \\ u_t = \varphi u_{t-1} + \varepsilon_t, |\varphi| < 1, \varepsilon_t \sim WN \end{cases}$$

(TS process with a deterministic trend).

[2] The formal definition of a unit root is as follows: we say that the AR(p) process has (one) unit root if its inverse characteristic polynomial $A_p(z) = 1 - \varphi_1 z - ... - \varphi_p z^p$ has one root equal to +1 (this condition is equivalent to $\rho = 0$ (why?)) while all the other roots are greater in modulus than 1 (see p.4-14).

Note that the third model can be rewritten as $Y_t = \tilde{\alpha} + \varphi Y_{t-1} + \tilde{\delta} t + \varepsilon_t, \varepsilon_t \sim WN$, which can still be generalized to allow for AR(p) disturbances: $Y_t = \tilde{\alpha} + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \tilde{\delta} t + \varepsilon_t$, $\varepsilon_t \sim WN$.

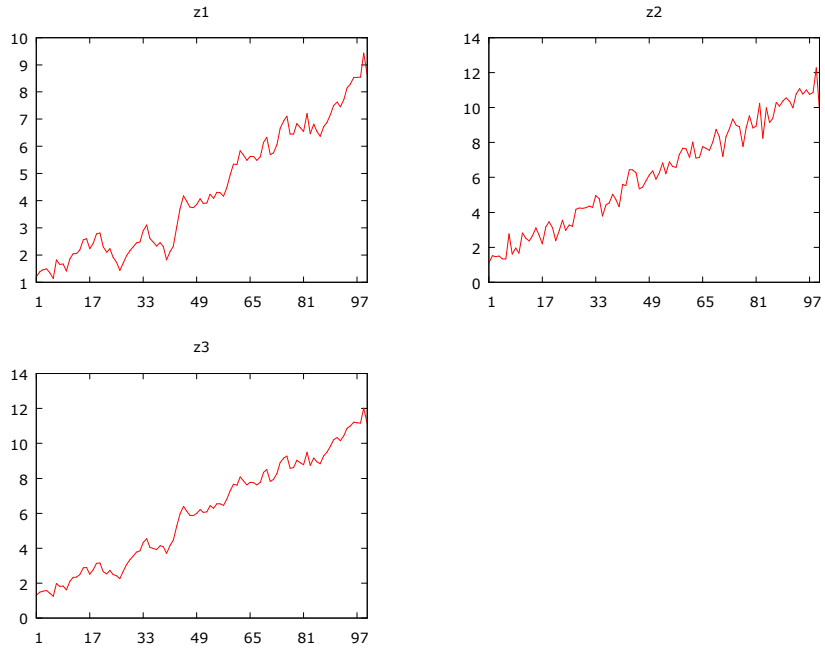


Figure 4.5. Simulated paths of the three models

All three models have very similar paths (see Fig. 4.5) but the properties of respective models are essentially different. The differences basically manifest themselves in forecasting: the forecasts of TS models tend to the trend line while DS model does not possess this **mean reverting** property. Note that in Model 2 shocks are WN, therefore already the one-step forecast $Y_{101,100}$ coincides with the trend. For Model 3 it takes some time for $Y_{100+h,100}$ to approach this trend. In contrast, the forecast of DS Model 1 starts at Y_{100} and goes **parallel** to the line δt (see Fig. 4.6). Also note that the accuracy of the forecast in Case 1 is ever decreasing (the confidence interval is broadening as $\sqrt{h}$) while in Case 2 it is always the same and in Case 3 it rapidly approaches its limit.

Thus, you should remember that **looking at time series plots alone is not enough to tell whether a series has a unit root** – we need some statistical procedures (i.e., tests) to decide upon it.

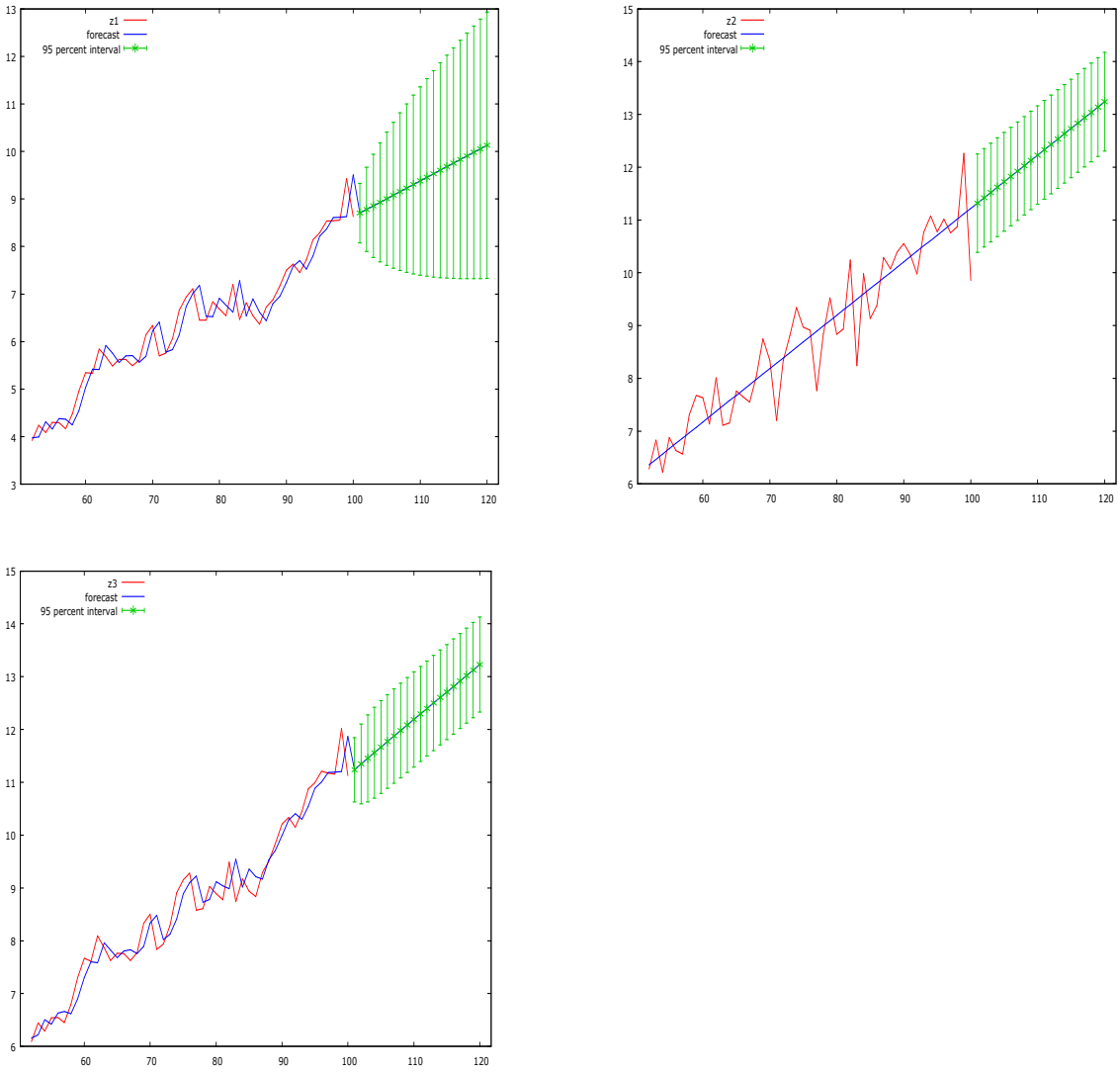


Figure 4.6. Three models and their forecasts with respective confidence bands

The LS estimate of δ in the **DS** process $Y_t = Y_0 + \delta t + \varepsilon_1 + ... + \varepsilon_t$, $\varepsilon \sim WN$, is inconsistent because the error process is not a WN. To estimate δ consistently, differentiate Y_t.
The LS estimate of δ in the stationary **TS** process
$Y_t = \alpha' + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \delta' t + \varepsilon_t, \varepsilon \sim WN$, is consistent.

The forecasts of TS models tend to the trend line while a DS model does not possess this **mean reverting** property (at any point the process begins anew.)

The discussion in the previous paragraph motivates jargon that we will use and introduce in the context of the following summary:

1. The **nonstationary** time series on which we focus are those containing a **unit root**. To put it differently, these series contain a **stochastic trend**. But if we difference these time series, the resulting time series will be stationary. For this reason, they are also called **difference stationary**.

2. The **stationary** time series on which we focus have $-2 < \rho < 0$ in the AR(p) model. However, these series can exhibit trend behavior through the incorporation of a **deterministic trend**. In this case, they are referred to as **trend stationary**.

If we add a deterministic trend to the AR(p) model, we obtain a very general model that is commonly used in univariate time series analysis:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_{p-1} \Delta Y_{t-p+1} + \delta t + \varepsilon_t \qquad (4.3)$$

We refer to the above as the **AR(p) with deterministic trend model** and use it later. We use namely this form of an AR process because it is more convenient for testing the hypothesis for unit root, i.e., $H_0 : \rho = 0$.

4.2 example. Stock prices on the NYSE (continued)

The following Table 4.1 contains output from an OLS regression of ΔY_t on $Y_{t-1}, \Delta Y_{t-1}$, $\Delta Y_{t-2}, \Delta Y_{t-3}$, and a deterministic time trend, created by using the data on stock prices from nyse.txt. In other words, it provides regression output for the AR(4) with deterministic trend model. We suspect that stock prices may contain a unit root, a supposition supported somewhat by the table. In particular, a unit root is present if ρ (the coefficient on Y_{t-1}) is zero. As we can see, the estimate of ρ is indeed very small (i.e. $\hat{\rho} =$ -0.016).

Table 4.1

```
? ols d_l_sp 0 l_sp(-1) d_l_sp(-1) d_l_sp(-2) d_l_sp(-3) time

Model 2: OLS, using observations 1952:05-1996:01 (T = 525)
Dependent variable: d_l_sp

              coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const        0.0820407     0.0395523      2.074    0.0385  **
  l_sp_1      -0.0161543     0.00831845    -1.942    0.0527  *
  d_l_sp_1     0.0513588     0.0439385      1.169    0.2430
  d_l_sp_2    -0.0273597     0.0439384     -0.6227   0.5338
  d_l_sp_3     0.0151113     0.0439494      0.3438   0.7311
  time         9.98321e-05   5.04478e-05    1.979    0.0484  **

Log-likelihood       939.5437   Akaike criterion    -1867.087
Schwarz criterion   -1841.507   Hannan-Quinn        -1857.071
rho                  0.001951   Durbin-Watson        1.991534
```

The main question we have to answer now is whether the estimate $\hat{\rho}$ is close enough to 0 to support the hypothesis $H_0 : \rho = 0$? In other words, we need a test.

4.2. Testing in the AR(p) with deterministic trend model

As we know, it is possible to use OLS to estimate the coefficients of the AR(p) model (4.3). However, in the case where we face the possibility of a unit root there is one important complication that occurs in the AR(p) model that was not present in earlier chapters. To understand it, let us divide the coefficients in the model into two groups: (1) $\alpha, \gamma_1, ..., \gamma_{p-1}$, and δ, and (2) ρ. In other words, we consider hypothesis tests involving ρ separately from those involving the other coefficients.

- **Testing involving $\alpha, \gamma_1, ..., \gamma_{p-1}$, and δ**

Many sophisticated statistical criteria and testing methods exist to determine the appropriate lag length in an AR(p) model. Nonetheless, simply looking at the t-ratio or p-values in regression outputs can be quite informative. For instance, an examination of Table 5.1 reveals that the p-values associated with the coefficients on the lagged ΔY terms are insignificant, and that they may be deleted from the regression (i.e. the p-values are greater than 0.05). Alternatively, a more common route is to proceed sequentially, that is, to choose a maximum lag length, $p_{\max}$, and then sequentially drop lag lengths if the relevant coefficients are insignificant.

More specifically, begin with an AR($p_{\max}$). If the $p_{\max}$ th lag is insignificant, we reduce the model to an AR($p_{\max}$ - 1). If the ($p_{\max}$ - 1)th lag is insignificant in the AR($p_{\max}$ - 1) then drop it and use an AR($p_{\max}$ - 2), etc. Generally, start with a fairly large choice for $p_{\max}$.

In the AR(p) with deterministic trend model we also have to worry about testing whether δ = 0. This can be accomplished in the standard way by checking whether its p-value is less than the level of significance (e.g., 0.05). This test can be done at any stage, but it is common to carry it out after following the sequential procedure for choosing p.

4.2 example. Stock prices on the NYSE (continued)

If we carry out the preceding strategy on the NYSE stock price data, beginning with $p_{\max} = 4$, the model reduces to $\Delta Y_t = \alpha + \rho Y_{t-1} + \delta t + \varepsilon_t$:

```
ols d_l_sp 0 l_sp(-1) time

Model 3: OLS, using observations 1952:02-1996:01 (T = 528)
Dependent variable: d_l_sp

             coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       0.0790879     0.0388022      2.038    0.0420  **
  l_sp_1     -0.0154821     0.00815052    -1.900    0.0580  *
  time        9.60002e-05   4.95819e-05    1.936    0.0534  *
Log-likelihood       944.6746   Akaike criterion    -1883.349
Schwarz criterion   -1870.542   Hannan-Quinn        -1878.335
rho                  0.047069   Durbin-Watson        1.905567     ◀◀
```

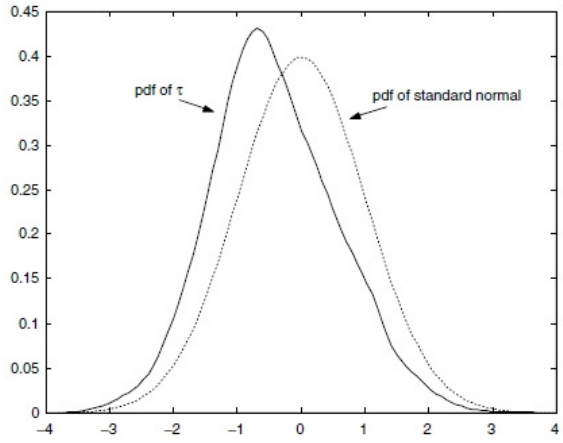
These results lead us to the next, most important, testing question: does Y contain a unit root? Remember that, if ρ = 0, then Y contains a unit root. In this case, the series must be differenced in the regression model (i.e., it is difference stationary). You may think that you can sim-

ply test $\rho = 0$ in the same manner as you tested the significance of the other coefficients. For instance, you might think that by comparing the *p*-value to a significance level (e.g., 0.05), you can test for whether $\rho = 0$. **SUCH A STRATEGY IS INCORRECT!** In hypothesis testing, ρ is different from other coefficients and thus we must treat it differently.

- **Testing involving ρ**

Most regression packages like GRETL implicitly assume that all of the variables in the model are stationary when they calculate *p*-values. If the explanatory variable Y_{t-1} is nonstationary, its *p*-value will be incorrect. A correct way of testing for a unit root, i.e., the null hypothesis $H_0 : \rho = 0$, has been developed by two statisticians named Dickey and Fuller and is known as the **Dickey–Fuller test**. They recommend using the same $t-$ratio for testing $\rho = 0$, but correcting the $p-$value.

We can motivate the Dickey–Fuller test in terms of the following: testing could be done by comparing a test statistic (here, the $t-$ratio) to a critical value to determine whether the former was either "small" (in which case the null hypothesis was accepted) or "large", i.e., "far from zero" (in which case the hypothesis was rejected). In the standard stationary case, the critical values are taken from statistical tables of the Student distribution. Dickey and Fuller demonstrated that in the unit root case this is incorrect. To explain, assume that $Y_t = \varphi Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, and test the null $H_0 : \varphi = 1$ (unit root) against the alternative stationarity hypothesis $H_1 : (-1 <) \varphi < 1$. In doing so, we used to apply the $t-$ratio statistics $\tau = (\hat{\varphi} - 1) / \widehat{s.e.(\hat{\varphi})}$; recall that if H_0 is true, τ has the Student distribution[3] (thus, if $\tau < -1.645$, we reject H_0). However, in the unit root case, the probability density function of τ is shifted to the left (see the figure on the right) and the 0.05 quantile is more negative than -1.645.



A rough rule of a thumb can be used that will not lead you too far wrong if your number of observations is moderately large (e.g., $T > 50$). This approximate rule is given in the following strategy for testing for a unit root:

1. Use the strategy outlined above to estimate the AR(*p*) with deterministic trend model. Record the $t-$statistics corresponding to ρ (i.e., the $t-$ratio for the coefficient on Y_{t-1}).
2. If the final version of your model **includes a deterministic trend**, the Dickey–Fuller critical value is approximately -3.45. If the $t-$ratio on ρ is more negative than -3.45, reject the unit root hypothesis and conclude that the series is stationary. Otherwise, conclude that the series has a unit root.

[3] Which, for big *T*, is close to standard normal.

3. If the final version of your model **does not include a deterministic trend**, the Dickey–Fuller critical value is approximately -2.89. If the $t-$statistics on ρ is more negative than -2.89, reject the unit root hypothesis and conclude that the series is stationary. Otherwise, conclude that the series has a unit root. ◀◀

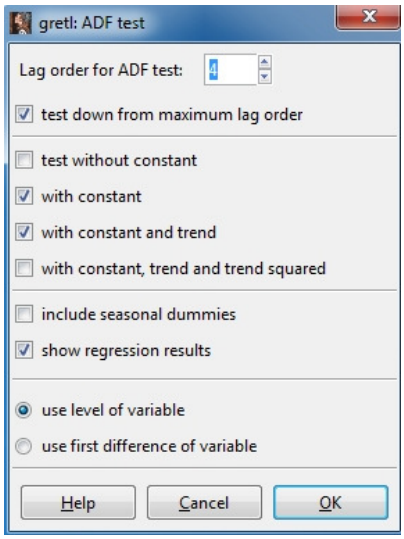
In the previous 4.2 example, the final version of the AR(p) model includes a deterministic trend. The $t-$statistics on ρ is -1.900, which is **not** more negative than -3.42 (see Table 4.2). Hence we can **accept** the hypothesis that NYSE stock prices contain a unit root and make a random walk with drift, i.e., are described by a model $\Delta \texttt{l\_sp}_t = \alpha + \varepsilon_t$ (c.f. CompLabs, p. 4-8, **3**. „`trend`“).

A more accurate estimate of $p-$value is given in Table 4.2.

Table 4.2 1% and 5% critical values for Dickey–Fuller tests

	Without trend		With trend	
Sample size	1%	5%	1%	5%
$T = 25$	−3.75	−3.00	−4.38	−3.60
$T = 50$	−3.58	−2.93	−4.15	−3.50
$T = 100$	−3.51	−2.89	−4.04	−3.45
$T = 250$	−3.46	−2.88	−3.99	−3.43
$T = 500$	−3.44	−2.87	−3.98	−3.42
$T = \infty$	−3.43	−2.86	−3.96	−3.41

Still better (once you have decided upon the form of a regression equation) is to apply to GRETL function `adf` (augmented Dickey-Fuller test).



4.2 example. Stock prices on the NYSE (continued)

```
adf 0 l_sp --ct

Dickey-Fuller test for l_sp
sample size 528
unit-root null hypothesis: a = 1

   test with constant
   model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + e
   1st-order autocorrelation coeff. for e: 0.047
   estimated value of (a - 1): -0.01548
   test statistic: tau_c(1) = -1.8995
   p-value 0.6533
```

Since the $p-$value 0.6533 isn‘t less than 0.05, we do not reject the hypothesis $\rho = 0$, i.e., once again we confirm the unit root hypothesis.

The same result can be obtained through the menu bar: in GRETL window, select `l_sp` and go to Variable| Unit root tests | Augmented Dickey-Fuller test and fill in the box as shown on the right (see p. 4-11) (thus, GRETL starts with $p = 4$ and selects $p = 1$). ◀◀

Once we have proved that the best model to describe `l_sp` is a random walk $\Delta Y_t = \alpha + \varepsilon_t$ or $\hat{Y}_t = \hat{Y}_{t-1} + \hat{\alpha} = \hat{Y}_{t-1} + 0.00686$, we create the model through Model| Ordinary Least Squares:

```
Model xx: OLS, using observations 1952:02-1996:01 (T = 528)
Dependent variable: d_l_sp

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const      0,00686458    0,00176760    3,884    0,0001  ***

Log-likelihood       942,7936   Akaike criterion    -1883,587
Schwarz criterion   -1879,318   Hannan-Quinn        -1881,916
rho                  0,039124   Durbin-Watson        1,921546
```

We can use the model to forecast `l_sp` till, say, 2000:1: go to Data| Add observations...| Number of observations to add...|48, go back to Model xx window and select Analysis| Forecasts... `l_sp`| automatic forecast (dynamic ...). After clicking OK, you will see both the graph (see Fig. 4.7) and numeric values[4].

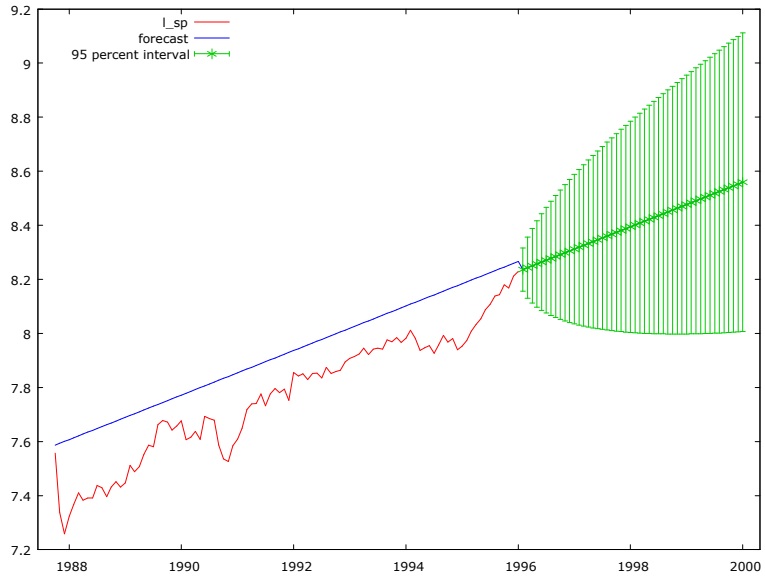


Figure 4.7. `l_sp` and its 48 months forecast

```
Obs          lsp          prediction   std. error   95% interval
.....................................................................
1996:01      8,229668     8,218999
1996:02                   8,236533     0,040578     8,157002 - 8,316064
1996:03                   8,243397     0,057386     8,130923 - 8,355871
1996:04                   8,250262     0,070283     8,112510 - 8,388014
.....................................................................
1999:11                   8,545439     0,275212     8,006032 - 9,084845
1999:12                   8,552303     0,278188     8,007065 - 9,097542
2000:01                   8,559168     0,281132     8,008160 - 9,110176
```

[4] The forecast line starts from the last point and is horizontal if there is no drift; the line is parallel to $\hat{\alpha}t$ if $\alpha \neq 0$.

4.2 exercise. Define the process $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, $\varepsilon_t \sim WN$ – is it TS or DS? How will you obtain a „good“ estimate of β_1? Define the process $Y_t = \beta_1 + Y_{t-1} + \varepsilon_t = Y_0 + \beta_1 t + \varepsilon_1 + ... + \varepsilon_t$, $\varepsilon_t \sim WN$ – is it TS or DS? Describe the procedure to obtain a „good“ estimate of β_1. What does it mean „good“?

4.3. Unit Roots - Summary

Testing for an (autoregressive) unit root is a time consuming procedure. To begin with, we rewrite our basic model $Y_t = \alpha + \delta t + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, as $\Delta Y_t = \alpha + \delta t + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$ (the latter form is more convenient since the unit root hypothesis can be simply formulated as $H_0 : \rho = 0$; the former model can also have a multicollinearity problem).

When testing for unit root, it is important to understand the meaning of the hypotheses – for example, the function `ur.df` in R package urca suggests three options: `type="none"`, `"drift"` and `"trend"`.

1. "`none`" – our time series Y_t resembles a stationary AR(1) process with zero mean (however, we suspect that it could be a random walk):

$$H_1 : Y_t = \varphi Y_{t-1} + w_t \quad \sim \quad Y_t - Y_{t-1} = (\varphi - 1) Y_{t-1} + w_t, |\varphi| < 1$$
$$H_0 : Y_t = Y_{t-1} + w_t \quad \sim \quad Y_t - Y_{t-1} = \mathbf{0} \cdot Y_{t-1} + w_t$$

In this case, the null $\varphi = 1$ means that our process is a random walk without drift. If the hypothesis is rejected, we conclude that the process is stationary AR(1) with zero mean.

2. "`drift`" – our time series resembles a stationary AR(1) process with nonzero mean (however, we suspect that it could be a random walk with drift):

$$H_1 : Y_t = \alpha + \varphi Y_{t-1} + w_t \sim Y_t - Y_{t-1} = \alpha + (\varphi - 1) Y_{t-1} + w_t, |\varphi| < 1$$
$$H_0 : Y_t = \alpha + Y_{t-1} + w_t \sim Y_t - Y_{t-1} = \alpha + \mathbf{0} \cdot Y_{t-1} + w_t$$

In this case, the null $\varphi = 1$ means that our process is a random walk with drift. If the hypothesis is rejected, we conclude that the process is stationary AR(1) with nonzero mean.

3. "`trend`" - our time series resembles a stationary AR(1) process around a linear trend (however, we suspect that it could be a random walk with drift):

$$H_1 : Y_t - a - bt = \varphi(Y_{t-1} - a - bt) + w_t \sim Y_t = [a(1-\varphi) + b\varphi] + b(1-\varphi) \cdot t + \varphi Y_{t-1} + w_t, |\varphi| < 1$$
$$H_0 : Y_t = b + Y_{t-1} + w_t \sim Y_t - Y_{t-1} = b + \mathbf{0} \cdot Y_{t-1} + w_t$$

In this case, the null $\varphi = 1$ means that our process is a random walk with drift. If the hypothesis is rejected, we conclude that our process is a stationary AR(1) around a line $a + bt$. ◀

We have to plot Y_t first and then decide upon the type of the model. In theory, our **sequential procedure** (see Sec. 4.2) gives us the right model, but be careful! Recall that when dealing with `l_sp`, our final model was (here we perform the calculation with R)

```
> library(dynlm)
> summary(dynlm(d(l_sp)~L(l_sp)+time(l_sp)))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.169524   1.124569  -1.929   0.0542 .
L(l_sp)     -0.015482   0.008150  -1.900   0.0580 .
time(l_sp)   0.001152   0.000595   1.936   0.0534 .
```

The $t-$statistics on ρ is -1.900, which is not more negative than -3.42 (see Table 4.2 and also below). Hence we can accept the hypothesis that NYSE stock price `l_sp` contains a unit root and makes a random walk with drift (i.e., $\varphi = 1$ and thus `l_sp` is described (see the "`trend`" case above) by a model $\Delta \texttt{l\_sp}_t = b + w_t$).

The same conclusion can be obtained in a simpler manner with

```
> library(urca)
> summary(ur.df(l_sp,type="trend",lags=0))

###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.918e-02  3.885e-02   2.038   0.0420 *
z.lag.1     -1.548e-02  8.151e-03  -1.900   0.0580 .
tt           9.600e-05  4.958e-05   1.936   0.0534 .

Value of test-statistic is: -1.8995 6.2954 1.877

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.96 -3.41 -3.12
phi2  6.09  4.68  4.03
phi3  8.27  6.25  5.34
```

The second approach in testing the unit root hypothesis is not to „look for the highest significant lag" but search the model with **minimum AIC** (or BIC) (this often allows to automate the model selection procedure: below, the function `ur.df` begins with `lags=4` and goes down till the minimum AIC model with `lags=1` is detected):

```
> summary(ur.df(l_sp,type="trend",lags=4,selectlags="AIC"))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.559e-02  3.915e-02   2.186   0.0292 *
z.lag.1     -1.680e-02  8.209e-03  -2.046   0.0412 *
tt           1.023e-04  4.983e-05   2.053   0.0406 *
z.diff.lag   5.261e-02  4.380e-02   1.201   0.2302
```

```
Value of test-statistic is: -2.0463 6.0166 2.1302

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.96 -3.41 -3.12
phi2  6.09  4.68  4.03
phi3  8.27  6.25  5.34
```

Now the $t-$statistics with ρ is -2.046, but it changes neither our final conclusion nor the model.

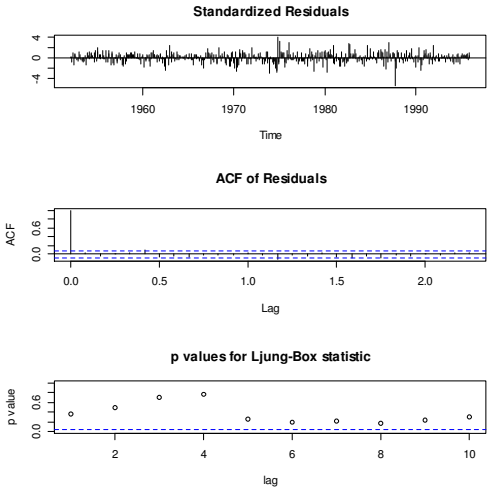
Still simpler technique to obtain the same corollary is to use

```
> library(forecast)
> aa.mod = auto.arima(l_sp)
> aa.mod

ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.0069
s.e.  0.0018

sigma^2 estimated as 0.001647:  log likelihood=942.79
AIC=-1881.59   AICc=-1881.56   BIC=-1873.05
```

The function `auto.arima` runs through many models (including those for ΔY_t) and chooses the one with minimum AIC, thus the „best" model here is ARIMA(0,1,0) which means unit root.

And **lastly**, we have to check whether the residuals of our final model make WN:

```
> tsdiag(aa.mod) # yes, residuals make WN
> plot(forecast(aa.mod),include=48)
```



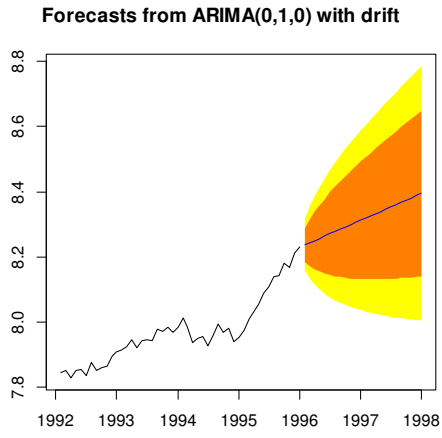


Figure 4.8. Residuals of the model `aa.mod` make white noise (left); the 24 months forecast is on the right

4.3 example. Below, we have plotted two simulated DS series (random walks without drift) and their common mean line, i.e., the x axis (black) as well as two estimates of the trend obtained with the `decompose` function (red). The function splits a time series into seasonal, „trend" and irregular components using moving averages. Note that it makes no sense to apply the function to the DS series: neither random walks nor their „trends" are close to the mean value $(EY_t \equiv)\, 0$. Random walk has no deterministic trend, it has only a stochastic „trend" which is different for every trajectory.

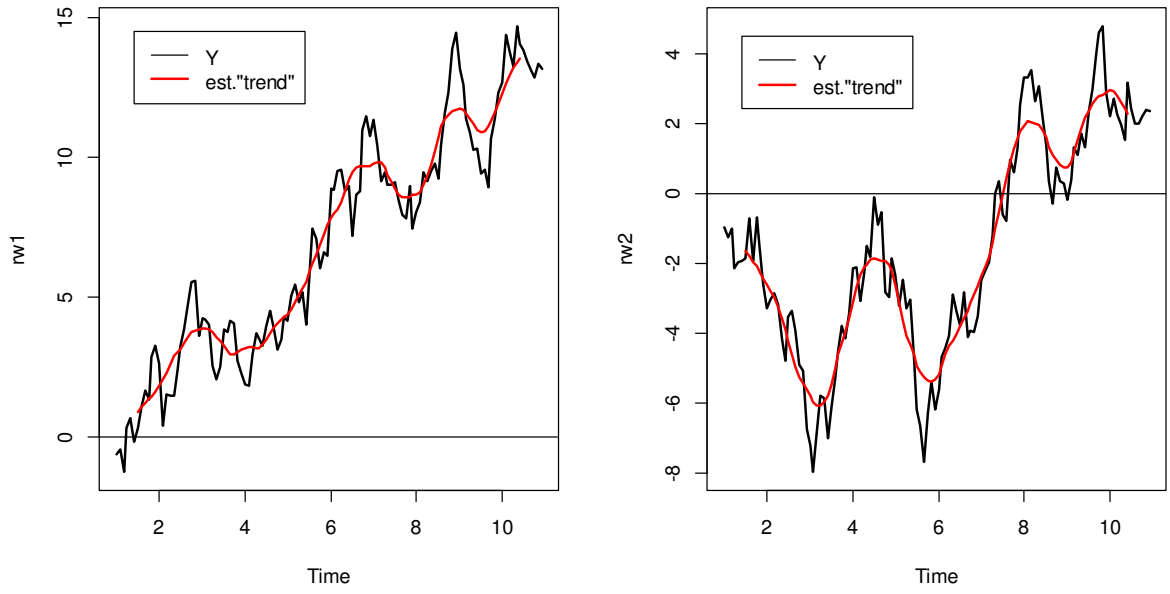


Figure 4.9. Two simulated random walks (black) and their smoothed stochastic trends (red)

4.4. Appendix

Here we shall present a more consistent definition of integrated processes.

Definition 4.1. We say that a process Y_t is $I(0)$ (integrated of zero*th* order) if it is stationary and invertible.

- White noise is $I(0)$.
- Stationary AR(p) process is $I(0)$.
- The MA(1) process $Y_t = \varepsilon_t - \varepsilon_{t-1}$ (i.e., differences of WN) is stationary but not invertible (why?), thus it is not $I(0)$.

Definition 4.2. The process Y_t is $I(1)$ (integrated of the first order) if it is not stationary but $\Delta Y_t = Y_t - Y_{t-1}$ is $I(0)$; the process Y_t is $I(2)$ (integrated of the second order) if it is not stationary, ΔY_t is also not stationary, but $\Delta(\Delta Y_t) = \Delta^2 Y_t$ is $I(0)$ etc[5].

- The special AR(1) process, namely, the random walk $Y_t = a + 1 \cdot Y_{t-1} + \varepsilon_t, \varepsilon_t \sim WN$, is an $I(1)$ process.
- The special ARMA(1,1) process $Y_t = a + 1 \cdot Y_{t-1} + \varepsilon_t - 1 \cdot \varepsilon_{t-1}, \varepsilon_t \sim WN$, is not an $I(1)$ process.
- The trend stationary process $Y_t = \alpha + \beta t + \varepsilon_t, \varepsilon_t \sim WN$, is not a stationary process (therefore it is not $I(0)$) but it is also not an $I(1)$ process. Indeed, $\Delta Y_t = \beta + \varepsilon_t - \varepsilon_{t-1}$ is not $I(0)$. Thus, if you want to get a „good" stationary process out of a TS process with a polynomial trend, do not differentiate it (you will get a „bad" process); instead, subtract the trend.

4.3 exercise. Is white noise process $Y_t = \varepsilon_t, \varepsilon_t \sim WN$ a) stationary? b) $I(0)$? c) $I(1)$?

- Let the AR(p) process $A_p(L)Y_t = \varepsilon_t, \varepsilon_t \sim WN$, is such that the inverse characteristic polynomial $A_p(x) = 1 + \varphi_1 x + ... + \varphi_p x^p$ has exactly one root equal to +1 while all the other roots are greater than 1 in modulus. Then we say that the process Y_t has (the first order) unit root. Since, in this case, $A_p(L) = A_{p-1}(L) \cdot (1 - L)$, the process Y_t is $I(1)$, i.e., ΔY_t is $I(0)$ (more specifically, ΔY_t is a stationary AR(p-1) process). Thus, if you differentiate an $I(1)$ process, you will get a stationary $I(0)$ process; if you differentiate AR(p) process with a unit root, you will get a stationary AR(p-1) process.

To sum up – the unit root process and integrated process are close but not identical concepts: the AR(p) process Y_t with one unit root is $I(1)$ and $\Delta Y_t = (1 / A_{p-1}(L))\varepsilon_t$ is $I(0)$; since an $I(1)$ process Z_t is not necessarily an AR process, $\Delta Z_t = (B(L) / A_{p-1}(L))\varepsilon_t \sim I(0)$; here $B(L)$ corresponds to an invertible process.

[5] In economics, integrated processes of order higher than 2 are rare.

5. Regression with time lags: Autoregressive Distributed Lag Models

The goal of the researcher working with time series data does not differ too much from that of the researcher working with cross-sectional data: both aim to develop a regression relating a dependent variable to some explanatory variables. However, the analyst using time series data will face two problems that the analyst using crosssectional data will not encounter: (1) one time series variable can influence another with a time lag; and (2) if the variables are nonstationary, a problem known as spurious regression may arise. However, one should aways keep in mind this general rule: **if you have nonstationary time series variables then you should not include them in a regression model.** The appropriate route is to transform the variables before running a regression in order to make them stationary. There is one exception to this general rule, which we shall discuss later, and which occurs where the variables in a regression model are nonstationary and **cointegrated**. In this chapter, we will assume all variables in the regression are stationary.

The value that the stock market places on a firm might depend not only on current income, but on historical income as well. After all, current income could be affected by short-term factors and may not be a totally reliable guide to long-run performance. Similar considerations hold for our executive compensation example where compensation might be determined not only on current profits, but also on past profits. In short, there are good reasons to include not only current values of explanatory variables, but also past values.

To put this concept in the language of regression, we say that the value of the dependent variable at a given point in time should depend not only on the value of the explanatory variable at that time period, but also on values of the explanatory variable in the past. A simple model to incorporate such **dynamic** effects has the form

$$Y_t = \alpha + \beta_0 X_t + ... + \beta_q X_{t-q} + \varepsilon_t .$$

Since the effect of the explanatory variable on the dependent variable does not happen all at once, but rather is distributed over several time periods, this model is sometimes referred to as a **distributed** (or weighted) **lag model**. Coefficients can be interpreted as always, i.e., as measures of the influence of the explanatory variable on the dependent variable. In this case, we have to be careful with timing. For instance, we interpret results as "β_2 measures the effect of the explanatory variable two periods ago on the dependent variable, *ceteris paribus*".

5.1. Selection of Lag Order

When working with distributed lag models, we rarely know *a priori* exactly how many lags we should include. That is, unlike most of the regression models considered in Ch. 0, we don't know which explanatory variables in a distributed lag model belong in the regression before we actually sit down at the computer and start working with the data. Appropriately, the issue of lag length selection becomes a databased one where we use statistical means to decide how many lags to include. There are many different approaches to lag length selection in the econometrics literature. Here we outline a common one that does not require any new statistical

techniques. This method uses *t*-tests for whether $\beta_q = 0$ to decide lag length. A common strategy is to

- Begin with a fairly large lag length, $q_{\max}$, and test whether the coefficient on the maximum lag is equal to zero (i.e., test whether $\beta_{q_{\max}} = 0$)
- If it is, drop the highest lag and re-estimate the model with maximum lag equal to $q_{\max} - 1$
- If you find $\beta_{q_{\max}-1} = 0$ in this new regression, then lower the lag order by one and re-estimate the model.
- Keep on dropping the lag order by one and re-estimating the model until you reject the hypothesis that the coefficient on the longest lag is equal to zero.

5.1 example. The effect of bad news on market capitalization

The share price of a company can be sensitive to bad news. Suppose that Company B is in an industry which is particularly sensitive to the price of oil. If the price of oil goes up, then the profits of Company B will tend to go down and some investors, anticipating this, will sell their shares in Company B driving its price (and market capitalization) down. However, this effect might not happen immediately. For instance, if Company B holds large inventories produced with cheap oil, it can sell these and maintain its profits for a while. But when new production is required, the higher oil price will lower profits. Furthermore, the effect of the oil price jump might not last forever, since Company B also has some flexibility in its production process and can gradually adjust to higher oil prices. Hence, news about the oil price should affect the market capitalization of Company B, but the effect might not happen immediately and might not last too long.

The file BADNEWS.XLS contains data collected on a monthly basis over five years (i.e., 60 months) on the following variables:

Y	market capitalization of Company B ($000)
X	the price of oil (dollars per barrel) above the benchmark price

Since this is time series data and it is likely that previous months' news about the oil price will affect current market capitalization, it is necessary to include lags of *X* in the regression. Below are present OLS estimates of the coefficients in a distributed lag model in which market capitalization is allowed to depend on present news about the oil price and news up to $q_{\max} = 4$ months ago. That is,

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_4 X_{t-4} + \varepsilon_t :$$

```
ols y 0 x x(-1) x(-2) x(-3) x(-4)

Model 1: OLS, using observations 1980:05-1984:12 (T = 56)
Dependent variable: y

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
```

```
  const      91173,3       1949,85      46,76     6,17e-043 ***
  x           -131,994       47,4361    -2,783    0,0076    ***
  x_1         -449,860       47,5566    -9,459    1,01e-012 ***
  x_2         -422,518       46,7778    -9,032    4,40e-012 ***
  x_3         -187,104       47,6409    -3,927    0,0003    ***
  x_4          -27,7710      47,6619    -0,5827   0,5627

Mean dependent var   74067,00   S.D. dependent var   10468,82
Sum squared resid    1,45e+09   S.E. of regression   5380,606
R-squared            0,759855   Adjusted R-squared   0,735840
F(5, 50)             31,64143   P-value(F)           2,22e-14
Log-likelihood      -557,3585   Akaike criterion     1126,717
Schwarz criterion    1138,869   Hannan-Quinn         1131,428
rho                 -0,161850   Durbin-Watson        2,240889
```

What can we conclude about the effect of news about the oil price on Company B's market capitalization? Increasing the oil price by one dollar per barrel in a given month is associated with:

1. An immediate reduction in market capitalization of $131,994, *ceteris paribus*.
2. A reduction in market capitalization of $449,860 one month later, *ceteris paribus*

and so on. To provide some intuition about what the *ceteris paribus* condition implies in this context note that, for example, we can also express the second of these statements as: "Increasing the oil price by one dollar in a given month will tend to reduce market capitalization in the following month by $449,860, **assuming that no other change in the oil price occurs**".

Since the *p*-value corresponding to the explanatory variable X_{t-4} is greater than 0.05 we cannot reject the hypothesis that $\beta_4 = 0$ at the 5% level of significance. Accordingly, we drop this variable from the model and re-estimate with lag length set equal to 3, yielding the results in the following table:

```
? ols y 0 x x(-1) x(-2) x(-3)

Model 2: OLS, using observations 1980:04-1984:12 (T = 57)
Dependent variable: y

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       90402,2      1643,18      55,02     9,19e-048 ***
  x            -125,900      46,2405    -2,723    0,0088    ***
  x_1          -443,492      45,8816    -9,666    3,32e-013 ***
  x_2          -417,609      45,7332    -9,131    2,18e-012 ***
  x_3          -179,904      46,2520    -3,890    0,0003    ***

Mean dependent var   74153,74   S.D. dependent var   10395,57
Sum squared resid    1,47e+09   S.E. of regression   5313,050
R-squared            0,757447   Adjusted R-squared   0,738789
F(4, 52)             40,59659   P-value(F)           2,09e-15
Log-likelihood      -567,2045   Akaike criterion     1144,409
Schwarz criterion    1154,624   Hannan-Quinn         1148,379
rho                 -0,142889   Durbin-Watson        2,234934
```

The *p*-value for testing $\beta_3 = 0$ is 0.0003, which is much less than 0.05. We therefore conclude that the variable X_{t-3} does indeed belong in the distributed lag model. Hence *q* = 3 is the lag length we select for this model. In a formal report, we would present this table of results. Since these results are similar to those discussed above, we will not repeat the interpretation of them.

5.2. Dynamic Models with Stationary Variables

In regression analysis, researchers are typically interested in measuring the effect of an explanatory variable or variables on a dependent variable. However, this goal is complicated when the researcher uses time series data since an explanatory variable may influence a dependent variable with a time lag. This often necessitates the inclusion of lags of the explanatory variable in the regression. Furthermore, the dependent variable may be correlated with lags of itself, suggesting that lags of the dependent variable should also be included in the regression.

These considerations motivate the commonly used **autoregressive distributed lag** (ADL) model:

$$Y_t = \alpha + \delta t + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \beta_0 X_t + ... + \beta_q X_{t-q} + \varepsilon_t \qquad (5.1)$$

In this model, the dependent variable *Y* depends on *p* lags of itself, the current value of an explanatory variable *X* as well as *q* lags of *X*. The model also allows for a deterministic trend *t*. Since the model contains *p* lags of *Y* and *q* lags of *X* we denote it by ADL(*p*, *q*). In this chapter, we focus on the case where there is only one explanatory variable *X*. Note, however, that we could equally allow for many expl anatory variables in the analysis.

Let us consider two <u>stationary</u> variables Y_t and X_t, and assume that it holds that

$$Y_t = \alpha + \varphi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t,\ 0 < \varphi < 1. \qquad (5.2)$$

As an illustration, we can think of Y_t as 'company sales' and X_t as 'advertising', both in month t. If we assume that ε_t is a white noise process, independent of X_t, X_{t-1} and Y_t, Y_{t-1}, the above relation can be estimated by the use of ordinary least squares.

The interesting element in (5.2) is that it describes the dynamic effects of a change in X_t upon current and future values of Y_t. Taking partial derivatives, we can derive that the immediate response is given by $\partial Y_t / \partial X_t = \beta_0$. Sometimes this is referred to as the **impact** (or **short-run**) **multiplier**. An increase in *X* with one unit has an immediate impact on *Y* of β_0 units. The effect after one period is

$$\partial Y_{t+1} / \partial X_t = \varphi\, \partial Y_t / \partial X_t + \beta_1 = \varphi\beta_0 + \beta_1$$

(this can also be derived in a more explicit way:

$$Y_{t+1} = \alpha + \varphi Y_t + \beta_0 X_{t+1} + \beta_1 X_t + \varepsilon_{t+1} =$$
$$\alpha + \varphi(\alpha + \varphi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t) + \beta_0 X_{t+1} + \beta_1 X_t + \varepsilon_{t+1} =$$
$$... + (\varphi\beta_0 + \beta_1) X_t + ...).$$

Similarly, after two periods

$$\partial Y_{t+2} / \partial X_t = \varphi \, \partial Y_{t+1} / \partial X_t = \varphi(\varphi\beta_0 + \beta_1)$$

and so on. This shows that after the first period, the effect is decreasing if $|\varphi| < 1$. Imposing this so-called stability condition allows us to determine the long-run effect of a unit change in X_t. It is given by the **long-run multiplier** (or equilibrium multiplier)

$$\beta_0 + (\varphi\beta_0 + \beta_1) + \varphi(\varphi\beta_0 + \beta_1) + ... = \beta_0 + (1 + \varphi + \varphi^2 + ...)(\varphi\beta_0 + \beta_1) = \frac{\beta_0 + \beta_1}{1 - \varphi}$$

This says that if advertising X_t increases with one unit for one moment, the expected cumulative increase in sales is given by $(\beta_0 + \beta_1) / (1 - \varphi)$.

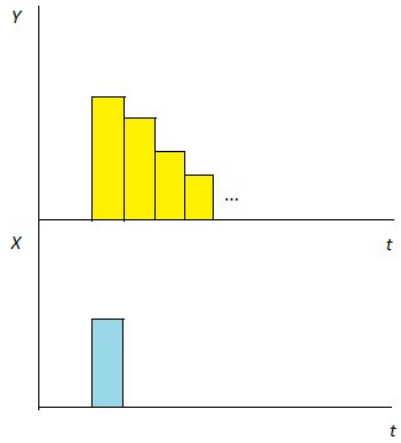


Figure 4.1. Yellow bars are the resprctive impact multipliers and their sum is the long-run multiplier.

On the other hand, if the increase in X_t is permanent, the long-run multiplier also has the interpretation of the expected long-run permanent increase in Y_t. From (5.2) the long-run equilibrium relation between Y and X can be seen to be (imposing $Y_{t-1} = Y_t = ... = Y$, $X_{t-1} = X_t = ... = X$, $\varepsilon_t = \varepsilon_{t+1} = ... = 0$)

$$Y = \alpha + \varphi Y + \beta_0 X + \beta_1 X$$

or

$$Y = \frac{\alpha}{1 - \varphi} + \frac{\beta_0 + \beta_1}{1 - \varphi} X \qquad (5.3)$$

which presents an alternative derivation of the long-run multiplier[1]. We shall write (5.3) concisely as $Y = \tilde{\alpha} + \tilde{\beta} X$, with obvious definitions of $\tilde{\alpha}$ and $\tilde{\beta}$ (see also section 6.6). Thus, if X changes to a new constant X', Y will finally change to $Y' = \tilde{\alpha} + \tilde{\beta} X'$ (but it will take some time!).

There is an alternative way to formulate the autoregressive distributed lag model from (5.2). Subtracting Y_{t-1} from both sides of (5.2) and some rewriting gives

$$\Delta Y_t = \alpha - (1-\varphi) Y_{t-1} + \beta_0 \Delta X_t + (\beta_0 + \beta_1) X_{t-1} + \varepsilon_t$$

or

$$\Delta Y_t = \beta_0 \Delta X_t - (1-\varphi)[Y_{t-1} - \tilde{\alpha} - \tilde{\beta} X_{t-1}] + \varepsilon_t \qquad (5.4)$$

This formulation[2] is an example of an **error-correction model** (ECM). It says that the change in Y_t is due to the current change in X_t plus an error-correction term: if Y_{t-1} is above the equilibrium value corresponding to X_{t-1}, that is, if the 'disequilibrium error' in square brackets is positive, then a „go to equilibrium" mechanism generates an additional negative adjustment in Y_t. The speed of adjustment is determined by $1-\varphi$, which is the adjustment parameter; note that stability assumption ensures that $0 < 1-\varphi < 1$ therefore only a part of any disequilibrium is made up for in the current period).

Notice that without prior knowledge of the long-run parameters, we cannot estimate the above ECM in the form (5.4). This is because without knowing $\tilde{\alpha}$ and $\tilde{\beta}$, we cannot construct the disequilibrium error $Y_{t-1} - \tilde{\alpha} - \tilde{\beta} X_{t-1}$. In the absence of such knowledge, to directly estimate the ECM, we must first multiply out the term in parentheses to obtain

$$\Delta Y_t = (1-\varphi)\tilde{\alpha} + \beta_0 \Delta X_t - (1-\varphi) Y_{t-1} + (1-\varphi)\tilde{\beta} X_{t-1} + \varepsilon_t . \qquad (5.5)$$

ΔY_t can now be OLS-regressed on ΔX_t, Y_{t-1}, and X_{t-1}, estimates of all short- and long-run parameters then being obtained.

We can further generalize (5.2) and (5.4). For example, if

$$Y_t = \alpha + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t ,$$

then the ECM is

$$\Delta Y_t = -\varphi_2 \Delta Y_{t-1} + \beta_0 \Delta X_t - \beta_2 \Delta X_{t-1} - (1-\varphi_1-\varphi_2)[Y_{t-1} - \breve{\alpha} - \breve{\beta} X_{t-1}] + \varepsilon_t$$

[1] If Y and X are in logarithms, $(\beta_0 + \beta_1)/(1-\varphi)$ is the long-run elasticity of Y with respect to X (what is the meaning of this multiplier if Y and X are in levels?)

[2] If Y and X are in logarithms, ΔY and ΔX can be regarded as proportional changes which are likely to form stationary series even if Y and X do not. In (5.4), β_0 is the short-run parameter whereas $\tilde{\beta}$ the long-run parameter, thus ECM contains both short- and long run multipliers.

(the original model must be rewritten in differences plus a disequilibrium error). To estimate this model, it is again necessary to express it in (5.5) form.

It is possible for more than two variables to enter into an equilibrium relationship, for example,

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \gamma_0 Z_t + \gamma_1 Z_{t-1} + \varphi Y_{t-1} + \varepsilon_t;$$

this equation transforms to

$$\Delta Y_t = \beta_0 \Delta X_t + \gamma_0 \Delta Z_t - (1-\varphi)[Y_{t-1} - \hat{\alpha} - \hat{\beta}_1 X_{t-1} - \hat{\gamma} Z_{t-1}] + \varepsilon_t$$

etc. All the ECMs may be consistently estimated via OLS provided all the predictors are stationary.

As long as it can be assumed that the error term ε_t is a white noise process, or – more generally – is stationary and independent of $X_t, X_{t-1}, \ldots$ and $Y_{t-1}, Y_{t-2}, \ldots,$ the ADL models can be estimated consistently by ordinary least squares. Problems may arise, however, if, along with Y_t and X_t, the implied ε_t is also nonstationary. This is discussed in Ch. 6.

6. Regression with Time Series Variables

Estimation and interpretation of the ADL(p, q) (see Ch. 5) model depend on whether the series X and Y are stationary or not. Note though that we assume throughout that X and Y have the same stationarity properties; that is, that they either must **both** be stationary or **both** have a unit root. Intuitively, regression analysis involves using X to explain Y. If the properties of X differ from those of Y, it becomes difficult for X to explain Y. For instance, it is hard for a stationary series to explain the stochastic trend variation in a unit root series. In practice this means that, before running any time series regression, you should examine the univariate properties of the variables you plan to use. In particular, you should carry out unit root tests along the lines described in Ch. 4 for every variable in your analysis.

6.1. Time Series Regression when X and Y are Stationary

This section is an extension of Section 5.2.

In the case of the AR(p) model (see (4.2)), it proved convenient for both the OLS estimation and interpretation to rewrite the model with ΔY as the dependent variable. Similar considerations hold for the ADL(p, q) model

$$Y_t = \alpha + \delta t + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \beta_0 X_t + ... + \beta_q X_{t-q} + \varepsilon_t \qquad (6.1)$$

which can be rewritten as:

$$\Delta Y_t = \alpha + \delta t + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_{p-1} \Delta Y_{t-p+1} + \theta X_t + \omega_1 \Delta X_t + ... + \omega_q \Delta X_{t-q+1} + \varepsilon_t. \qquad (6.2)$$

It should be emphasized that this model is the same as that in the original form of the ADL(p, q); it has merely undergone a few algebraic manipulations. This model may look complicated, but it is still nevertheless just a regression model. That is, no new mathematical techniques are required for this model, which is, after all, still based on the simple equation of a straight line.

We have already discussed how to interpret regression coefficients, placing special emphasis on *ceteris paribus* conditions. Recall that we made statements of the form: "The coefficient measures the influence of X_k on Y, *ceteris paribus*". In the ADL(p, q) model, such an interpretation can still be made, but it is not that commonly done. In economics, a common way to interpret the coefficients is through the concept of a multiplier. Thus, β_0 in (6.1) is called the **short run** or **impact multiplier** (it describes how the size 1 one-moment increase of X_t affects Y_t). However, it is common to focus on the **long run** or **total multiplier**, which is what we will do here. To motivate this measure, suppose that X and Y are in an equilibrium or steady state, i.e., are not changing over time. All of a sudden, X changed permanently to a new level one unit higher than the previous value, affecting Y, which starts to change, eventually settling down in the long run to a new equilibrium value. The difference between the old and new equilibrium values for Y can be interpreted as the long run effect of X on Y and is the long run multiplier. This multiplier is often of great interest for policymakers who want to know

the eventual effects of their policy changes in various areas. The long run multiplier measures the effect of this sort of change. In some cases, you might be interested in the effect of a temporary change in X (i.e., X starts at some original level, then increases by one unit for one period before going back to the original level the next). The long run multiplier does not measure the effect of this type of change. We can use the traditional "marginal effect" interpretation of regression coefficients for such temporary changes (recall that this is termed the short run multiplier). It can be shown that the **long run multiplier for the ADL(p, q) model is** $-\theta / \rho$. In other words, only the coefficients on X_t and Y_{t-1} in the rewritten ADL model matter for long run behavior.

To be more specific in distinguishing different multipliers, we shall analyse the ADL(1,1) model in more detail. The model $Y_t = \alpha + \varphi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$ can be rewritten in terms of X (to have convergent series, we assume that $|\varphi| < 1$) as follows:

$$(1 - \varphi L) Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$$

or

$$Y_t = (1 - \varphi L)^{-1}(\alpha + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t) = (1 + \varphi L + (\varphi L)^2 + ...)(\alpha + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t) =$$

$$\frac{\alpha}{1-\varphi} + \beta_0 X_t + (\beta_0 \varphi + \beta_1) X_{t-1} + \varphi(\beta_0 \varphi + \beta_1) X_{t-1} + \varphi^2 (\beta_0 \varphi + \beta_1) X_{t-2} + ... + v_t =$$

$$\tilde{\delta} + \delta_0 X_t + ... + \delta_s X_{t-s} + ... + v_t$$

The coefficient $(\delta_0 =)\ \beta_0$ at X_t or $\partial Y_t / \partial X_t$ is called the short-run multiplier, the coefficient $(\delta_s =)\ \varphi^s (\beta_0 \varphi + \beta_1)$ at X_{t-s} or $\partial Y_{t+s} / \partial X_t$ the s period delay multiplier, the number $\sum_{i=1}^{s} \delta_i$ the s period interim multiplier and, finally, $\sum_{i=0}^{\infty} \delta_i = \frac{\beta_0 + \beta_1}{1 - \varphi}$ the total or long-run multiplier. In other words, the short-run multiplier is the immediate Y_t 's response to a momentary change in X: $X_1 = ... = X_{t-1} = X$, $X_t = X + 1$ whereas the long-run multiplier is the eventual Y_t 's, $t \to \infty$, response to a permanent change in X: $X_1 = ... = X_{t-1} = X$, $X_t = X_{t+1} = ... = X + 1$.

Per pratybas Hill, Griffiths 381 psl

6.1 example. The effect of financial liberalization on economic growth.

Researchers in the field of international finance and development are interested in whether financial factors can play an important role in encouraging growth in a developing country. The purpose of this example is to investigate this issue empirically using time series data from a single country. Data set LIBERAL.XLS contains data from Country A for 98 quarters on GDP growth and a variable reflecting financial liberalization: the expansion of the stock market. In particular, the dependent and explanatory variables are:

- `Y = pchGDP` the percentage change in GDP.
- `X = pchSMC` the percentage change in total stock market capitalization.

The mean of these two variables is 0.30% and 0.01% per quarter, indicating that stock markets in Country A have not expanded by much **on average**. Note, however, that this average hides wide variation. In some quarters market capitalization increased considerably, while in other quarters it decreased. Assuming that both variables are stationary, we can estimate an ADL(2, 2) model using OLS. Remember that, if the variables in a model are stationary, then the standard regression quantities (e.g. OLS estimates, *p*-values, confidence intervals) can be calculated in an ordinary way. Table 6.1 contains the results of this procedure.

Table 6.1 Regression model

$$\Delta pchGDP_t = \alpha + \delta t + \rho\, pchGDP_{t-1} + \gamma_1 \Delta pchGDP_{t-1} + \theta\, pchSMC_t + \omega_1 \Delta pchSMC_t + \omega_2 \Delta pchSMC_{t-1} + \varepsilon_t$$

```
? ols dpchGDP 0 time pchGDP(-1) dpchGDP(-1) pchSMC dpchSMC dpchSMC(-1)

Model 2: OLS, using observations 3-98 (T = 96)
Dependent variable: dpchGDP
              coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------------
  const       -0,0297651     0,0426415     -0,6980   0,4870
  ttt          0,000729571   0,000741733    0,9836   0,3280
  pchGDP_1    -0,119616      0,0126390     -9,464    4,11e-015 ***
  dpchGDP_1    0,794232      0,0309911     25,63     7,41e-043 ***
  pchSMC       0,125388      0,0481253      2,605    0,0108    **
  dpchSMC      0,837527      0,0438253     19,11     2,96e-033 ***
  dpchSMC_1    0,00224124    0,0217915      0,1028   0,9183

Log-likelihood       22,54807   Akaike criterion    -31,09614
Schwarz criterion   -13,14570   Hannan-Quinn        -23,84028
```

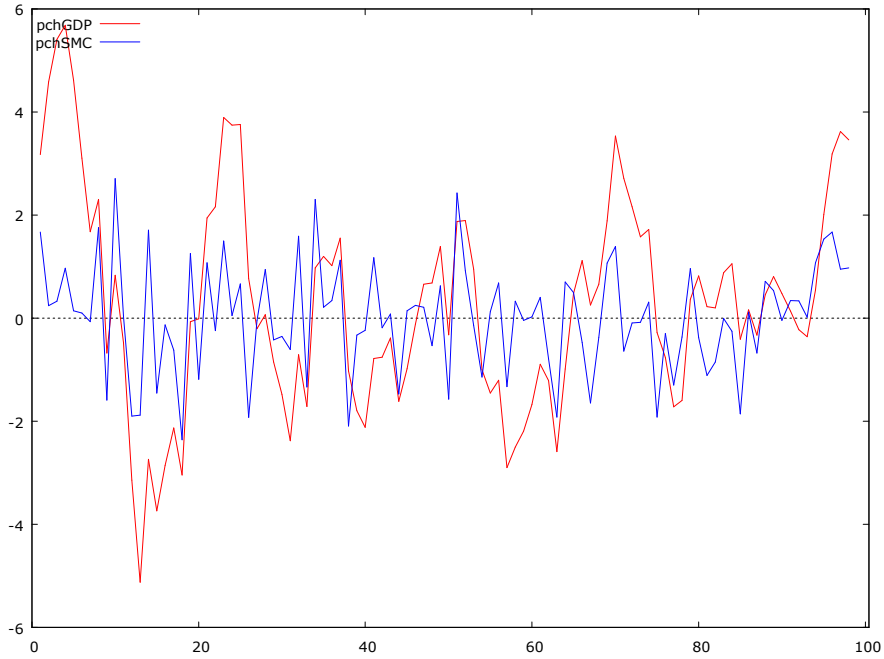


Figure 6.1. Graphs of two stationary time series: `pchGDP` (red) and `pchSMC` (blue)

Using the formula for the long run multiplier, we can see that its OLS estimate is -(.125/-.120) = 1.042. There are different ways of expressing this information verbally (remember that the dependent and explanatory variables are percentage changes):

1. On average, market capitalization in Country A has been increasing by 0.01% per quarter and GDP by 0.30% per quarter. If Country A's total stock market capitalization increases by 1.01% in each month (i.e., increase by one unit from 0.01 each month to 1.01 each month), then in the long run its GDP should start increasing by 1.342% per quarter (i.e., the initial 0.30 plus the long run multiplier of 1.042). Can you imagine how the graphs in Fig. 6.1 change?
2. The long run multiplier effect of financial liberalization on GDP growth is 1.042%.
3. If *X* permanently increases by 1%, the equilibrium value of *Y* will increase by 1.042%.

The statistical information, though, indicates that this might not be a good model, since some of the explanatory variables are not significant (e.g. the *p*-values for the coefficients on ΔX_{t-1} and the time trend both imply insignificance at the 5% level). This raises the issue of lag length selection in the ADL(*p*, *q*) model. We will not discuss this topic here, other than to note that the strategy for selecting *q* in the regression model with lagged explanatory variables and the strategy for selecting *p* in the AR(*p*) model can be combined. There is no general convention about whether you should first select *p*, then *q*, then decide whether the deterministic trend should be included, or make another ordering (e.g., select *q*, then *p* then trend or select *q* then trend then *p*, etc.). As long as you are careful, you will not be led too far wrong in selecting a good model. The author's final ADL model is

$$\begin{aligned}\Delta pchGDP_t = 0.007 + (\rho =)(-0.120)\cdot pchGDP_{t-1} + 0.800\cdot \Delta pchGDP_{t-1} + \\ (\theta = 0.124)\cdot pchSMC_t + 0.839\cdot \Delta pchSMC_t + \varepsilon_t,\end{aligned}$$

the long-run multiplier $-\theta/\rho = 1.033$, and the equilibrium model $pchGDP = 0.007 + 1.033\ pchSMC$ (thus, if *pchSMC* increases by 1 to a new permanent level, *pchGDP* will eventually increase by 1.033).

6.2. Time series regression when *Y* and *X* have unit roots: spurious regression

For the remainder of this chapter, we will assume that *Y* and *X* have unit roots. In practice, of course, you would have to test whether this was the case using the last chapter sequential procedure and/or Dickey–Fuller (or any other unit root) test. We begin by focussing on the case of regression models without lags, then proceed to similar models to the ADL(*p*, *q*) model. Suppose we are interested in estimating the following regression:

$$Y_t = \alpha + \delta t + \beta X_t + \varepsilon_t . \qquad (6.3)$$

If *Y* and *X* contain unit roots, then OLS estimation of this regression can yield results which are completely wrong. For instance, even if the true value of β is 0, OLS can yield an estimate $\hat{\beta}$ which is very different from zero. Statistical tests (using the *t*-ratio or *p*-value) may indicate that β is not zero. Furthermore, if $\beta = 0$, then the R^2 should be zero. In fact, the R^2 will often be quite large. To put it another way: **if *Y* and *X* have unit roots then all the usual OLS regression results might be misleading and incorrect**. This is the so-called **spurious regression problem**.We do not have the statistical tools to prove that this problem occurs, but it is important to stress the practical implication. With the one exception of cointegration that

we note below, **you should never run a regression of *Y* on *X* if the variables have unit roots**.

In Fig. 6.2 we illustrate the above statements. x and y are two independent random walks with $\delta = 0$ (as it follows from independence, the true β equals 0). However, the OLS estimate $\hat{\beta}$ differs from 0 and, as its $p-$value is almost always less than 0.05 (the last graph, blue line), we "conclude" that β is almost always „significant". Do not trust OLS in such a case!

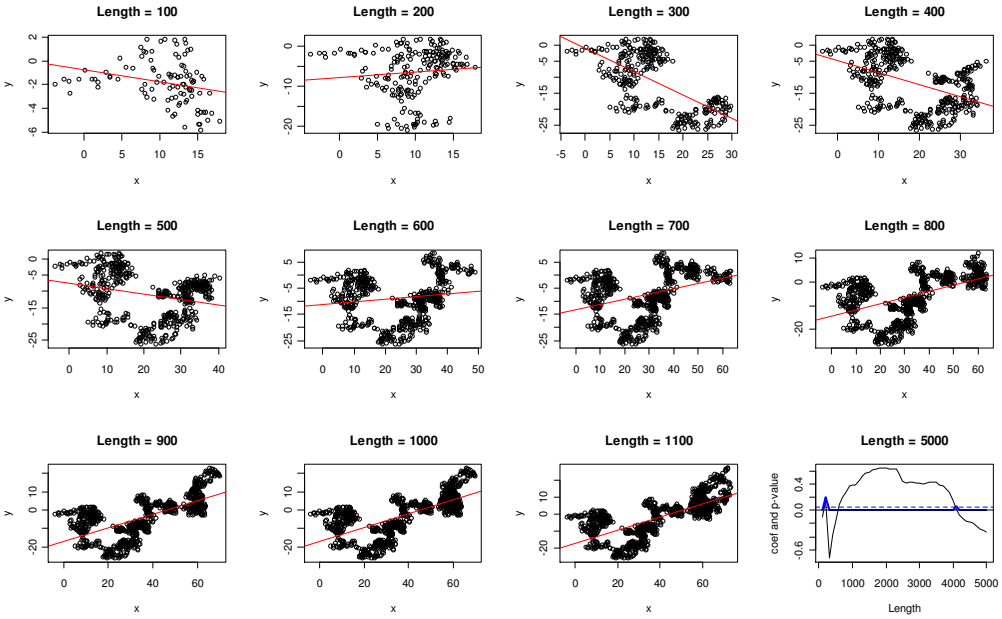


Figure 6.2. *x* and *y* are two independent random walks.

We will see soon that the reason of these wrong conclusions is the fact that $\varepsilon_t \sim I(1)$. Here are some properties of popular statistics in that „bad" case where Y_t, X_t, and $\varepsilon_t \sim I(1)$:

- DW statistics tends to 0 when $T \to \infty$
- R^2 tends to a random limit
- $\hat{\beta}^{OLS}$ is not consistent
- t-statistics on $\hat{\beta}^{OLS}$ diverge – need to be divided by $\sqrt{T}$ to obtain a random variable with well defined distribution

In short, do not regress Y_t on X_t in such a case. However, the question remains - what to do if Y_t and X_t have unit roots and are not cointegrated? (on the "good" case of cointegration, see Sct. 6.3 below). The general advice is to rethink your model, for example, instead of working with $Y_t = \alpha + \beta X_t + \varepsilon_t$, analyze the model $Y_t = \alpha + \varphi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$ (in this case the OLS estimators are consistent [redacted]). Another possibility is to create an ADL model for stationary differences ΔY_t and ΔX_t, for example:

$$\Delta Y_t = \alpha + \varphi \Delta Y_{t-1} + \beta_0 \Delta X_t + \beta_1 \Delta X_{t-1} + \varepsilon_t .$$

If, for example, Y_t is log of GDP and X_t is log of the price level, then ΔY_t and ΔX_t are the growth rate and inflation, respectively, and the above equation has a nice interpretation.

6.3. Time Series Regression when Y and X Have Unit Roots: Cointegration

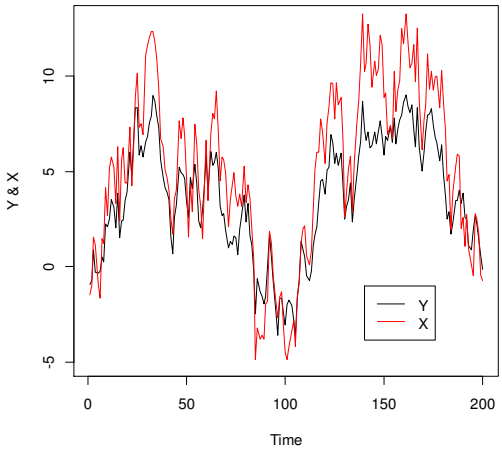
The one time where you do not have to worry about the spurious regression problem occurs when *Y* and *X* are cointegrated. This case not only surmounts the spurious regression problem, but also provides some nice financial intuition. Cointegration has received a great deal of attention recently in the financial literature and many theoretical finance models imply cointegration should occur, so it is worthwhile to discuss the topic in detail here.

Some intuition for cointegration can be obtained by considering the errors in the above regression model: $\varepsilon_t = Y_t - \alpha - \beta X_t$. Written in this way, it is clear that the errors are just a linear combination of *Y* and *X*. However, *X* and *Y* both exhibit nonstationary unit root behavior such that you would expect the error to also exhibit nonstationary behavior. (After all, if you add two things with a certain property together the result generally tends to have that property.) The error does indeed usually have a unit root. Statistically, it is this unit root in the error term that causes the spurious regression problem. However, it is possible that the unit roots in *Y* and *X* "cancel each other out" and that the resulting error is stationary (or, in other words, *Y* and *X* will trend together.). In this special case, called cointegration, the spurious regression problem vanishes and it is valid to run a regression of *Y* on *X*. To summarize, here is a definition: **if *Y* and *X* have unit roots, but some linear combination of them, $\gamma_1 Y_t + \gamma_2 X_t$, is stationary, then we say that *Y* and *X* are *cointegrated***. Note that the coefficients γ_1 and γ_2 are not uniquely defined: if $\gamma_1 Y_t + \gamma_2 X_t$ is stationary, then, for example, $1 \cdot Y_t + (\gamma_2 / \gamma_1) \cdot X_t$ is stationary, too[1] (we say that the latter expression is *normalized* with respect to Y).

How to make sure that such constants exist? Stationarity means little variability around the mean therefore, to best approximate Y_t in terms of X_t, we can use OLS.

6.2 example. Let

$$\begin{cases} Y_t = 1.3 X_t + u_t, u_t = 0.6 u_{t-1} + \varepsilon_{1t} \\ X_t = X_{t-1} + \varepsilon_{2t} \end{cases}, \ \varepsilon_{it} \sim iid\ N(0,1).$$



Here, both X_t and Y_t are quite irregular random walks, but the linear combination $1 \cdot Y_t - 1.3 \cdot X_t$ is a stationary AR(1) process u_t; thus Y_t and X_t cointegrate (they have a common stochastic trend X_t). In practice, we have to test the residuals of the regression $Y_t = \alpha + \beta X_t + u_t$ for stationarity (see Engle-Granger test in Sct. 6.4). ◀◀

[1] Why?

In the real world, it is unlikely that a financial system will ever be in precise equilibrium since shocks and unexpected changes to it will always occur. However, departures u_t from equilibrium should not be too large and there should always be a tendency to return to equilibrium after a shock occurs. Hence, if a financial model which implies an equilibrium relationship exists, then we should observe Y and X as being cointegrated.
Common sense tells you that, if two assets are close substitutes for one another, then their prices should not drift too far apart. After all, if one asset becomes much more expensive than a similar asset, then investors will sell the first asset in order to buy the cheaper alternative. But if many investors are selling the expensive asset, then its price should drop. And if many investors are buying the cheap asset its price would rise. Thus, the prices of the expensive and cheap assets would move closer to one another. Many financial theories formalize this intuition to imply different cointegrating relationships.

Cointegration often arises in models of the **term structure of interest rates** and the **yield curve**. The basic idea is that bonds can have different maturities or repayment periods. So you can have a bond which promises to pay a fixed interest rate for one year, or two years, or 10 years, etc. The interest rates paid on bonds of different maturities can be different since investors have different time preferences and long maturities are less flexible since they lock the investor in for a longer time period. That is, an investor could either buy a five-year bond, or a sequence of one-year bonds each year for five years. The latter strategy would be more flexible since the investor could always change her mind after each year. Hence, long-term interest rates often tend to be higher than short-term interest rates to compensate the buyer for a loss of flexibility. The exact shape of the relationship between interest rates at different maturities is called the term structure of interest rates or the yield curve (a yield is the return to holding the bond for the entire time until it matures). This provides much useful information about investor's beliefs about the future and is, thus incorporated in many financial theories some of which imply cointegrating relationships. For instance, in [CLM, Ch.10] is outlined an argument where yield spreads (i.e., the difference between the yield of a bond with an N period maturity and the yield of a bond with a 1 period maturity) are stationary time series variables and show how this implies yields of different maturities should be cointegrated.

In futures markets, theories involving investors having rational expectations tend to imply cointegrating relationships. For instance, in foreign exchange markets you can buy any major currency (e.g., the $ or the £) in the conventional manner (i.e., for immediate delivery at a specified rate). This is referred to as the **spot exchange rate** or **spot rate**. However, it also possible to agree an exchange rate now, but carry out the actual trade at some future date (e.g., a deal might have the form "I will guarantee that one year from now, I will give you $2.00 for your £1"). Such an exchange rate, agreed now but with the actual trade to be carried out later, is called the **forward exchange rate** or **forward rate**. Similar contracts (and much more complicated ones) can be written in stock markets and, indeed, such **financial derivatives** play a huge role in modern financial markets. Many financial theories, involving market efficiency and rational expectations of investors, imply that forward rates should be good predictors of future spot rates. Empirically, it seems that prices of assets (and an exchange rate is a price of an asset) often have unit roots in them (with returns being stationary). If we combine the financial theory with this empirical regularity, it turns out that they imply that spot and forward rates should be cointegrated. In foreign exchange markets, there are many theories which imply such cointegrating relationships. We will not explain them here, but just drop a few of names such as purchasing power parity, uncovered interest parity and covered interest parity. As we have touched on previously, there are also many financial theories which come out of basic present value relationships which imply cointegration. For instance, one such the-

ory implies that stock prices and dividends should be cointegrated. Another financial theory implies that consumption c, assets a and income y should be cointegrated. Such so-called ***cay*** relationships have received a great deal of attention in the recent empirical finance literature. Furthermore, theory suggests that the cointegrating error from the *cay* relationship plays a very important role: it should have predictive power for future stock returns.
In short, financial theory suggests cointegrating relationships between many different financial time series should exist. Hence, it is important to test whether cointegration is present (i.e., to see whether financial theory holds in practice) and, if it is present, to estimate models involving cointegrated variables (e.g., to estimate the cointegrating error from the *cay* relationship). Accordingly, we now address these issues, beginning with an empirical example.

6.3 example. Cointegration between the spot and forward rates

We have discussed previously how financial theory suggests spot and forward rates should be cointegrated. As an example, forexN.xls contains time series data for 181 months on the spot and one-month forward exchange rates of a certain foreign currency (both variables are measured in foreign currency units per dollar). Figure 6.1 plots these two series and provides strong visual evidence that the spot and forward rates are indeed cointegrated. That is, even though they are not identical to one another, the general trend behavior in the two variables looks quite similar.

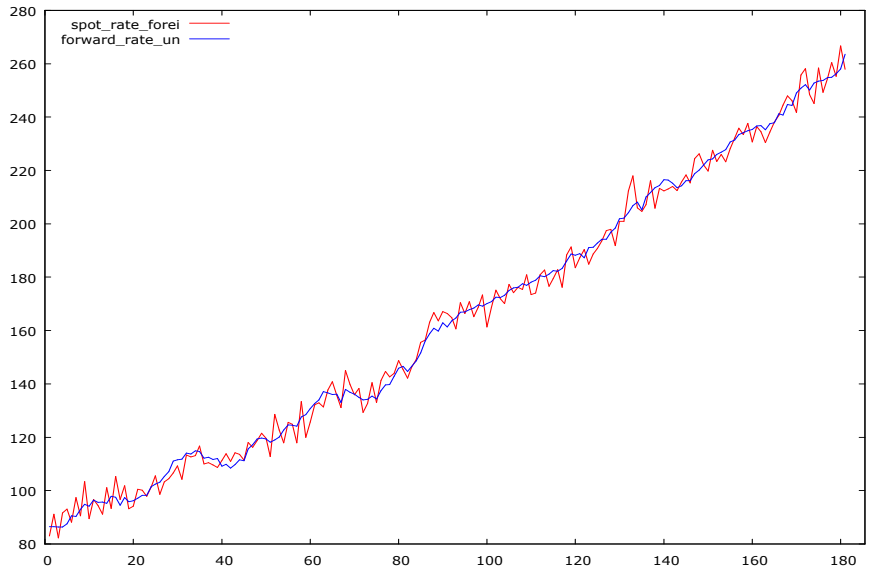


Figure 6.3. `spot_rate` and `forward_rate` plots

6.4. Estimation and Testing with Cointegrated Variables

As mentioned above, if Y and X are cointegrated, then the spurious regression problem does not apply; consequently, we can run an OLS regression of Y on X and obtain valid results. Furthermore, the coefficient from this regression is the long run multiplier. Thus, insofar as interest centers on the long run multiplier, then estimation with cointegrated variables is very easy.

The test for cointegration described here is referred to as the Engle–Granger test, after the two econometricians who developed it. It is based on the regression of *Y* on *X*. Remember that, if cointegration occurs, then the errors from this regression will be stationary. Conversely, if cointegration does not occur, then the errors will have a unit root. Given the close relationship between the errors and the residuals, it is reasonable to examine the properties of the residuals in order to investigate the presence of cointegration. In Ch. 5 we discussed testing for a unit root in a time series variable. Here, we test for a unit root in the ***residuals*** using the same techniques. In particular, the test for cointegration involves the following steps:

1. Run the regression of *Y* on *X* and save the residuals.
2. Carry out a unit root test on the residuals (without including a constant and a deterministic trend).
3. If the unit root hypothesis is rejected, then conclude that *Y* and *X* are cointegrated. However, if the unit root is accepted then conclude cointegration does not occur.

Thus if Y_t and X_t are cointegrated in $Y_t = \alpha + \beta X_t + \varepsilon_t$, the error term is $I(0)$. If not, ε_t will be $I(1)$. Hence one can test for the presence of a cointegrating relationship by testing for a unit root in the OLS residuals e_t. It seems that this can be done by using the Dickey–Fuller tests of the previous chapter. For example, one can run the regression

$$\Delta e_t = \rho e_{t-1} + \sum_{j=1}^{p} \beta_j \Delta e_{t-j} + u_t,\ u_t \sim WN \qquad (6.4)$$

and test whether $\rho = 0$ (a unit root). There is, however, an additional complication in testing for unit roots in OLS residuals rather than in observed time series. Because the OLS estimators „choose“ the residuals in the cointegrating regression to have as small variance as possible, even if the variables are not cointegrated, the OLS estimator will make the residuals „look“ as stationary as possible. Thus, using standard DF or ADF tests, we may reject the null hypothesis of nonstationarity too often. As a result, the appropriate critical values are more negative than those for the standard Dickey–Fuller tests in Table 4.2 and are presented in Table 6.2. If e_t is not appropriately described by a first order autoregressive process, one should add lagged values of Δe_t to (6.4), leading to the augmented Dickey–Fuller (ADF) tests, with the same asymptotic critical values. This test can be extended to test for cointegration between three or more variables. If more than a single X_t variable is included in the cointegrating regression, the critical values shift further to the left. This is reflected in the additional rows in Table 6.2.

Table 6.2 Asymptotic 5% critical values residual unit root tests for cointegration

Number of X‘s in the rhs of (6.3)	No deterministic terms in (6.3)	Only a constant in (6.3)	Constant and trend in (6.3)
1	-2.76	-3.37	-3.80
2	-3.27	-3.77	-4.16
3	-3.74	-4.11	-4.49

4	-4.12	-4.45	-4.74
5	-4.40	-4.71	-5.03

Thus if the ADF test statistic $\hat{\rho} / se(\hat{\rho})$ is closer to zero than respective critical value, we do not reject the unit root hypothesis. Note also that in the Dickey–Fuller test, we test the hypothesis that $\rho = 0$ (i.e., the null hypothesis is the unit root). In the cointegration test, we use the Dickey–Fuller methodology but cointegration is found if we reject the unit root hypothesis for the residuals. In other words, the null hypothesis in the Engle–Granger test is "no cointegration" and we conclude "cointegration is present" only if we reject this hypothesis.

6.2 example. **Cointegration between the spot and forward rates (continued)**

First we manually implement the following strategy:

(**1**) carry out Dickey-Fuller tests on the null hypothesis that each of the variables listed has a unit root
(**2**) estimate the cointegrating regression
(**3**) run a DF test on the residuals from the cointegrating regression.

(**1**) Let us suppose that spot and forward rates both have unit roots (probably in a regression with trend *t*; how to test the claim?).

(**2**) If we run a regression of `sr` = the spot rate on `fr` = the forward rate using the data in forexN.xls, we obtain the following fitted regression model:

```
ols sr 0 fr

Model 1: OLS, using observations 1-181
Dependent variable: sr

             coefficient   std. error   t-ratio     p-value
  ------------------------------------------------------------
  const       0,774335     0,974207       0,7948   0,4278
  fr          0,995574     0,00567534   175,4      3,09e-202 ***

rho                -0,085248   Durbin-Watson        2,153426
```

(**3**) The above presented strategy suggests that we should next carry out a unit root test on the residuals, u_t from this regression. The first step in doing this is to correctly select the lag length using the sequential strategy outlined in Ch. 4. Suppose we have done so and conclude that an AR(1) specification for the residuals is appropriate (how to test this claim?). The Dickey–Fuller strategy suggests we should regress Δu_t on u_{t-1}:

```
# save the residuals
    genr ui = $uhat
```

```
    ols diff(ui) 0 ui(-1) # cointegrating regression

Model 4: OLS, using observations 2-181 (T = 180)
Dependent variable: d_ui

             coefficient   std. error    t-ratio     p-value
  ------------------------------------------------------------
  const       0,0241279    0,291848      0,08267   0,9342
  ui_1       -1,08529      0,0748477    -14,50     5,80e-032 ***
```

Since *t*-ratio -14.50 is definitely less than (recall that T=181) -3.37 (see Table 6.2), we reject the unit root hypothesis and conclude that the residuals do not have a unit root. In other words, we conclude that the spot and forward rates are indeed cointegrated. ◀◀

In GRETL, the Engle-Granger cointegration test can be performed in one shot with the function `coint`: the flag `--ct` (see below) means that both constant and linear trend must be included in all regression models; flag `--test-down` means that the given value 4 is taken as the maximum and the actual lag order used in each case is obtained by testing down):

```
? coint 4 sr fr --ct --test-down

Step 1: testing for a unit root in sr

Augmented Dickey-Fuller test for sr
including 2 lags of (1-L)sr
sample size 176
unit-root null hypothesis: a = 1
   with constant and trend
   model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: -0,023
   lagged differences: F(2, 171) = 28,563 [0,0000]
   estimated value of (a - 1): -0,172594
   test statistic: tau_ct(1) = -2,77991
   asymptotic p-value 0,2047

Step 2: testing for a unit root in fr

Augmented Dickey-Fuller test for fr
including 2 lags of (1-L)fr
sample size 176
unit-root null hypothesis: a = 1

   with constant and trend
   model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: -0,012
   lagged differences: F(2, 171) = 3,861 [0,0229]
   estimated value of (a - 1): -0,0394676
   test statistic: tau_ct(1) = -1,77939
   asymptotic p-value 0,715

Step 3: cointegrating regression

Cointegrating regression -
OLS, using observations 1-181



Dependent variable: sr
```

```
            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       2,52803      3,91577      0,6456   0,5194
  fr          0,971740     0,0518494   18,74     5,57e-044 ***
  time        0,0236185    0,0510713    0,4625   0,6443

Mean dependent var   163,8089   S.D. dependent var   51,52887
Sum squared resid    2760,724   S.E. of regression   3,938234
R-squared            0,994224   Adjusted R-squared   0,994159
Log-likelihood      -503,4179   Akaike criterion     1012,836
Schwarz criterion    1022,431   Hannan-Quinn         1016,726
rho                 -0,086001   Durbin-Watson        2,156768

Step 4: testing for a unit root in uhat

Augmented Dickey-Fuller test for uhat
including 2 lags of (1-L)uhat
sample size 176
unit-root null hypothesis: a = 1

   model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: 0,004
   lagged differences: F(2, 173) = 2,324 [0,1009]
   estimated value of (a - 1): -0,869617
   test statistic: tau_ct(2) = -6,51698
   asymptotic p-value 1,638e-007

There is evidence for a cointegrating relationship if:
(a) The unit-root hypothesis is not rejected for the individual variables.
(b) The unit-root hypothesis is rejected for the residuals (uhat) from the
    cointegrating regression.
```

Since the *p*-value is less than 0.05, we reject unit root in residuals and once again conclude cointegration between `sr` and `fr`. ◀◀

We have found cointegration and therefore we do not need to worry about the spurious regressions problem. Hence, we can proceed to an interpretation of our coefficients without worrying that the OLS estimates are meaningless. The coefficient on the forward rate is 0.996 which is very close to the value of 1 predicted by financial theory (i.e., financial theory says that spot and forward rates should on average be the same as the latter should be an optimal predictor for the former). Alternatively, we can interpret this coefficient estimate as saying that the long run multiplier is 0.996 (see Model 1 on p. 6-9).

6.5. Time Series Regression when *Y* and *X* are Cointegrated: the Error Correction Model

You may encounter instances where unit root tests indicate that your time series have unit roots, but the Engle–Granger test indicates that the series are not cointegrated (they have no long-run relationship). That is, the series may not be trending together and may not have an equilibrium relationship. In these cases, you should not run a regression of Y on X due to the spurious regression problem. The presence of such characteristics suggests that you should rethink your basic model and include other explanatory variables. Instead of working with Y and X themselves, for example, you could difference them. (Remember that if Y and X have one unit root, then ΔY and ΔX should be stationary.) In this case, you could work with the

changes in your time series and estimate the ADL model using the techniques described at the beginning of this chapter. In other words, you may wish to estimate the original ADL model, but with changes in the variables:

$$\Delta Y_t = \varphi + \gamma_1 \Delta Y_{t-1} + ... + \gamma_p \Delta Y_{t-p} + \omega_0 \Delta X_t + ... + \omega_q \Delta X_{t-q} + \varepsilon_t .$$

Although such models can be estimated, it is important to understand that they describe only short-run interactions (since there is no equilibrium, there is no sence to speak about long run multiplier). Furthermore, if the variables are cointegrated, these models are not satisfactory, because they ignore the long-run relationship between the variables Y_t and X_t.

The last remark is best understood through example. A principal feature of cointegrated variables is that their time paths are influenced by the extent of any deviation from long-run equilibrium. After all, if the system is to return to the long-run equilibrium, the movements of at least some of the variables must respond to the magnitude of the disequilibrium. For example, theories of the term structure of interest rates imply a long-run relationship between long- and short-term rates, that is, between $R_{L,t}$ and $R_{S,t}$. In other words, these two variables are cointegrated: there exists such a constant β that the disequilibrium error $R_{L,t} - \beta R_{S,t}$ is stationary. If the gap $R_{L,t-1} - \beta R_{S,t-1}$ between the long- and short-term rates is positive, the short-term rate must ultimately rise relative to the long-term rate.

The dynamic model implied by this discussion is one of **error correction**. In an error-correction model, the short-term dynamics of the variables in the system are influenced by the deviation from equilibrium. If we assume that both interest rates are $I(1)$, a simple vector error-correction model (VECM[2]; see more on this in Ch. 7, Section 7.4) that could be applied to the term structure of interest rates is

$$\begin{cases} \Delta R_{L,t} = -\alpha_L (R_{L,t-1} - \beta R_{S,t-1}) + \varepsilon_{L,t} & (\alpha_L > 0, \varepsilon_L \sim WN) \\ \Delta R_{S,t} = \alpha_S (R_{L,t-1} - \beta R_{S,t-1}) + \varepsilon_{S,t} & (\alpha_S > 0, \varepsilon_S \sim WN) \end{cases} \qquad (6.5a)$$

where R_L and R_S are the long- and short-term interest rates, respectively. As specified, the short- and long-term interest rates change in response to stochastic shocks (represented by $\varepsilon_{S,t}$ and $\varepsilon_{L,t}$) and to the previous period's deviation from long-run equilibrium. Everything else equal, if this deviation happened to be positive (so that $R_{L,t-1} - \beta R_{S,t-1} > 0$), the short-term interest rate would rise and the long-term rate would fall (long-run equilibrium is attained when $R_{L,t} = \beta R_{S,t}$).

Note that (6.5a) can be written in a matrix form as a vector error correction model (VECM):

$$\begin{pmatrix} \Delta R_L \\ \Delta R_S \end{pmatrix}_t = \begin{pmatrix} -\alpha_L \\ \alpha_S \end{pmatrix} \begin{pmatrix} 1 & -\beta \end{pmatrix} \begin{pmatrix} R_L \\ R_S \end{pmatrix}_{t-1} + \begin{pmatrix} \varepsilon_L \\ \varepsilon_S \end{pmatrix}_t \qquad (6.5b)$$

[2] The systems (6.5) describes the behavior of two variables, R_S and R_L, therefore the model is called a vector ECM.

(we shall generalize it in Section 7.4; it is important that this model can be used to analyze and forecast the vector $\begin{pmatrix} R_L \\ R_S \end{pmatrix}_t$, $t > T$).

These results are unaltered if we add to the model more stationary terms and obtain a general VECM in two variables:

$$\begin{cases} \Delta R_{S,t} = \alpha_{10} + \alpha_S (R_{L,t-1} - \beta R_{S,t-1}) + \sum a_{11}(i)\Delta R_{S,t-i} + \sum a_{12}(i)\Delta R_{L,t-i} + \varepsilon_{S,t} \\ \Delta R_{L,t} = \alpha_{20} - \alpha_L (R_{L,t-1} - \beta R_{S,t-1}) + \sum a_{21}(i)\Delta R_{S,t-i} + \sum a_{22}(i)\Delta R_{L,t-i} + \varepsilon_{L,t} \end{cases} \quad (6.5c)$$

So far the presented material was explained on an intuitive level. However, there exists the **Granger Representation Theorem** that says that if Y and X are cointegrated, then the relationship between them can always be expressed as an ECM.

A more common expression for ECM in <u>two</u> variables (Y and X) is a single equation form

$$\Delta Y_t = \varphi + \lambda e_{t-1} + \omega_0 \Delta X_t + \varepsilon_t$$

where e_{t-1} is the error obtained from the regression model with Y and X (i.e., $e_{t-1} = Y_{t-1} - \alpha - \beta X_{t-1}$)[3]. Note that if we knew e_{t-1}, then the ECM is just a regression model. We can also generalize the above expression to

$$\Delta Y_t = \varphi + \delta t + \lambda e_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_p \Delta Y_{t-p} + \omega_0 \Delta X_t + ... + \omega_q \Delta X_{t-q} + \varepsilon_t .$$

Some sophisticated statistical techniques have been developed to estimate the ECM, but the simplest thing to do is merely to replace the unknown errors by the residuals from the regression of Y on X (i.e., replace e_{t-1} by $\hat{e}_{t-1}$). That is, a simple technique based on two OLS regressions proceeds as follows:

Step 1. Run a regression of Y on X and save the residuals.
Step 2. Run a regression of ΔY_t on all regressors from the above equation including the residuals from Step 1 lagged one period.

6.4 example. Cointegration between the spot and forward rates (continued from 6.2 example).

In the previous part of this example, we found that the variables, $Y =$ `sr` = the spot rate and $X =$ `fr` = the forward rate, were cointegrated. This suggests that we can estimate an error correction model. To do so, we begin by running a regression of Y on X and saving the residuals (which was done in the previous part of the example). The residuals u_t can then be included in the following regression (in lagged form):

[3] If the model contains two or more X's, there can be more than one cointegrating relationship and more than one error correction term (see Section 7.4).

$$\Delta Y_t = \varphi + \lambda u_{t-1} + \omega_0 \Delta X_t + \varepsilon_t$$

Below you can see results from OLS estimation of this model. The statistical information can be interpreted in the standard way. We can say that (with the exception of the intercept) all the coefficients are strongly statistically significant (since their *p*-values are much less than 0.05).

```
? ols sr 0 fr

Model 1: OLS, using observations 1-181
Dependent variable: sr

             coefficient   std. error   t-ratio     p-value
  ---------------------------------------------------------
  const       0,774335     0,974207      0,7948   0,4278
  fr          0,995574     0,00567534   175,4     3,09e-202 ***

?  series srfrRES=$uhat
Generated series srfrRES (ID 7)
? ols diff(sr) 0 srfrRES(-1) diff(fr)

Model 2: OLS, using observations 2-181 (T = 180)
Dependent variable: d_sr

              coefficient   std. error    t-ratio     p-value
  ------------------------------------------------------------
  const       -0,0231999    0,342999      -0,06764   0,9461
  srfrRES_1   -1,08504      0,0750502    -14,46      8,69e-032  ***
  d_fr         1,04369      0,181935       5,737     4,11e-08   ***
```

We noted before that $\hat{\beta} = 0.996$ and this is the estimate of the long run multiplier. The point estimates in the table of λ and ω_0 summarize the short run properties. To aid in interpretation note that all variables in the model are percentages. The coefficient on u_{t-1} of -1.085 measures how much Y responds to equilibrium errors. Since this coefficient is negative, positive errors tend to cause ΔY to be negative and hence Y to fall. In particular, an equilibrium error of one unit tends to cause the spot rate to fall by 1.085 units in the next period, *ceteris paribus*. This is a very quick adjustment to an equilibrium error! The coefficient on $\Delta X_t = 1.044$. Imagine, in other words, what would happen if the forward rate were to remain unchanged for some time ($\Delta X = 0$), but then suddenly were to change by one unit. The ECM implies that Y would instantly change by 1.044 units. In other words, the spot rate responds very quickly to changes in the forward rate.

6.6. The Long-run and Short-run Multipliers*)

Now we shall get back to the material which was earlier exposed in Ch. 5 and the opening pages of this chapter. First, assume that both Y_t and X_t are stationary.

1. Let us consider two processes Y_t and X_t, bound together by the static equation $Y_t = \alpha + \beta X_t + \varepsilon_t$, $t = ..., T-1, T, T+1, ...$ If the values of Y_t and X_t do not change:

$... = X_{T-1} = X_T = X$, $... = Y_{T-1} = Y_T = Y(= \alpha + \beta X)$, and $\varepsilon_t \equiv 0$, we say that X_t and Y_t are in equilibrium. Now, assume that in time $T+1$ the variable X_t changes its value to $X_{T+1} = X_{T+2} = ... = X+1$ – the (X_t, Y_t) system will move to a new equilibrium but, generally speaking, it could take some time. A (usually not very realistic) static system is remarkable in the sense that it reaches the new equilibrium in no time:

$$Y_{T+1} = \alpha + \beta X_{T+1} = \alpha + \beta \cdot (X+1) \quad \Rightarrow \quad Y_{T+1} - Y = \beta$$
$$Y_{T+h} = \alpha + \beta X_{T+h} = \alpha + \beta \cdot (X+1) \quad \Rightarrow \quad Y_{T+h} - Y = \beta,\ h \geq 2,$$

Here β is both the short-run (when $h=1$) and long-run (when $h \to \infty$) multiplier.

2. Transition to the new equilibrium is slower in dynamic case. Assume that DGP is described by $Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$. Now

$$Y_{T+1} = \alpha + \beta_0 X_{T+1} + \beta_1 X_T = \alpha + \beta_0 \cdot (X+1) + \beta_1 X \quad \Rightarrow \quad Y_{T+1} - Y = \beta_0$$
$$Y_{T+h} = \alpha + \beta_0 X_{T+h} + \beta_1 X_{T+h-1} = \alpha + (\beta_0 + \beta_1) \cdot (X+1) \quad \Rightarrow \quad Y_{T+h} - Y = \beta_0 + \beta_1,\ h \geq 2$$

i.e., the short-run (or impact) multiplier is β_0 and the long-run multiplier is $\beta_0 + \beta_1$. Note that $\beta_0 + \beta_1$ is also the slope of the equilibrium equation which can be obtained from the original dynamic equation as $t \to \infty$, $X_{t-1} \to X$ and $X_t \to X$: $(\lim_{t\to\infty} Y_t =)\ Y = \alpha + (\beta_0 + \beta_1) X$.

7.1 exercise. Find the short- and long-run multipliers if the DGP is the ADL(0,2) process: $Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t$. Can you rewrite the original equation as $Y_t = \alpha + \beta_0 \Delta X_t + (\beta_0 + \beta_1)\Delta X_{t-1} + (\beta_0 + \beta_1 + \beta_2) \cdot X_{t-2} + \varepsilon_t$? Which coefficient is the short-run and which one the long-run multiplier?

3. Assume that the DGP is described as ADL(1,1) process: $Y_t = \alpha + \varphi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$. In equilibrium, at moment T we have $Y = \alpha + \varphi Y + \beta_0 X + \beta_1 X$ or $Y_T = (\alpha + (\beta_0 + \beta_1) \cdot X) / (1 - \varphi)$. Thus

$$Y_{T+1} = \alpha + \varphi Y_T + \beta_0 (X+1) + \beta_1 X \quad \Rightarrow Y_{T+1} - Y_T = ... = \beta_0,$$

i.e., the short-run multiplier equals β_0. To find the limit $Y_{T+h} - Y_T$ as $h \to \infty$, repeat the reasoning of p. 4-4[4] which implies that the long-run multilier equals $(\beta_0 + \beta_1) / (1 - \varphi)$.

7.2 exercise. Can you rewrite the above equation as $\Delta Y_t = \alpha + (\varphi - 1) Y_{t-1} + \beta_0 \Delta X_t + (\beta_0 + \beta_1) X_{t-1} + \varepsilon_t$? Verify that the short-run multiplier is β_0 and the long-run one equals $(\beta_0 + \beta_1) / (1 - \varphi)$.

4. The ADL(1,1) process can be rewritten as the error correction model (ECM):

[4] Or apply the equilibrium equation.

$$\Delta Y_t = \alpha + \beta_0 \Delta X_t + (\varphi - 1)\left(Y_{t-1} - \frac{\beta_0 + \beta_1}{1-\varphi} X_{t-1} \right) + \varepsilon_t =$$

$$= \beta_0 \Delta X_t + (\varphi - 1)\left(Y_{t-1} - \frac{\alpha}{1-\varphi} - \frac{\beta_0 + \beta_1}{1-\varphi} X_{t-1} \right) + \varepsilon_t$$

where the short-run and long-run features of the dynamic relationship are modelled separately. Here β_0 describes the short-run relationship, $(\beta_0 + \beta_1)/(1-\varphi)$ the long-run multiplier, and $\varphi - 1$ speed of adjustment to equilibrium described by the error correction term.

5. Let us assume that the above equations are not in levels but in logarithms. Then respective multipliers are called short-run elasticity and long-run elasticity. ◀◀

Now, assume that both Y_t and X_t are $I(1)$ processes.

6. Regression analysis applied to non-stationary time series can produce misleading results (spurious regression), therefore (if the series are not cointegrated, i.e., they are not trending together) we should not look for a long-run multiplier in equilibrium equation $Y_t = \alpha + \beta X_t + \varepsilon_t$. Instead, to find the short-run multiplier, assume that the DGP for stationary differences can be expressed as ADL(1,1): $\Delta Y_t = \varphi \Delta Y_{t-1} + \beta_0 \Delta X_t + \beta_1 \Delta X_{t-1} + \varepsilon_t$. For example, say $y_t = \log GDP_t$ and $x_t = \log P_t$ (here P_t is stock exchange index) are both integrated but not cointegrated. In this case, regression $y_t = \alpha + \beta x_t + \varepsilon_t$ is senseless, but ADL(p,q) equation for stationary GDP growth rate Δy_t and stock returns Δx_t can be reasonable (the coefficient β_0 is the short run multiplier).

7. If Y_t and X_t are cointegrated, i.e., the equilibrium relationship exists, the ECM equation $\Delta Y_t = \varphi + \lambda(Y_{t-1} - \alpha - \beta X_{t-1}) + \omega_0 \Delta X_t + \varepsilon_t$ contains both short run multiplier ω_0 and long run multiplier β.

6.7. Summary

The static (or long-run) model $Y_t = \beta_0 + \beta_1 X_t^{(1)} + \beta_2 X_t^{(2)}$ depicts an equilibrium relationship such that for given values of the right-hand-side variables and their long-run impact on Y (i.e., the β_i), there is no reason for the response to be at any other value than Y_t. However, it is also of interest to consider the short-run evolution of the variables under consideration, especially since equilibrium (i.e., the steady-state) may rarely be observed. The major reason why relationships are not always in equilibrium centres on the inability of economic agents to adjust to new information instantaneously. There are often substantial costs of adjustment (both pecuniary and non-pecuniary) which result in the current value of the dependent variable Y being determined not only by the current value of some explanatory variables X_t but also by their past values $X_{t-1}, ..., X_{t-q}$. In addition, as Y evolves through time in reac-

tion to current and previous values of X, past (i.e., lagged) values of itself (i.e., $Y_{t-1},...,Y_{t-p}$) will also enter the short-run (dynamic) model. This inclusion of lagged values of the dependent variable as regressors is a means of simplifying the form of the dynamic model (which would otherwise tend to have a large number of highly correlated lagged values of X and lead to the problem of multicollinearity – high R^2 but imprecise parameter estimates and low t- values, even if though the model may be correctly specified). A very simple dynamic model of short-run adjustments is

$$y_t = \alpha_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \alpha_1 y_{t-1} + \varepsilon_t, \varepsilon_t \sim WN \qquad (6.6)$$

(here variables in lower case are in logarithms). Clearly, the parameter γ_0 denotes the short-run reaction of y_t to a change in x_t, and not the long-run effect that would occur if the model were in equilibrium (the latter is defined as $y_t = \beta_0 + \beta_1 x_t$). So in the long-run, the elasticity between Y and X is $\beta_1 = (\gamma_0 + \gamma_1)/(1-\alpha_1)$, assuming that $|\alpha_1| < 1$ (which is a necessary if the short-run model is to converge to a long-run solution).

The dynamic model represented by (6.6) is easily generalised to allow for more complicated, and often more realistic adjustment processes by increasing the lag-lengths. However, there are several potential problems with this form of the dynamic model. The **first** has already been mentioned and concerns multicollinearity. Also, some (if not all) of the variables in a dynamic model are likely to be non-stationary, since they enter in levels (this leads to the potential problem of spurious regression). A solution might be to respecify the dynamic model in (first) differences. However, this then removes any information about the long-run from the model (are we above or below the equilibrium?) and consequently is unlikely to be useful for forecasting purposes. A more suitable approach is to adopt the error-correction (ECM) formulation of the dynamic model:

$$\Delta y_t = \gamma_0 \Delta x_t - (1-\alpha_1)[y_{t-1} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{t-1}] + \varepsilon_t \qquad (6.7)$$

Equations (6.6) and (6.7) are equivalent but the ECM has some advantages.

A **second** feature of the ECM is that all the terms in the model are stationary so standard regression techniques are valid, assuming cointegration and that we have estimates of β_0 and β_1.

The simple ECM depicted in (6.7) can be generalised in several directions. First, increasing the lag-length p and/or q in (6.6) results in additional lagged first differences entering (6.7):

$$A(L)\Delta y_t = B(L)\Delta x_t - (1-\pi)[y_{t-1} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{t-1}] + \varepsilon_t,$$

$\pi = \alpha_1 + ... + \alpha_p$. Second, if more than one X enters the model, we have to consider the possibility for more than one cointegrating relationship to exist. This leads to the ECM in multivariate form, VECM, which will be explored in Ch. 7.

Revision questions

6.1 Let $Y_t = 3 + 2X_{t-1} - 1.3X_{t-2} + \varepsilon_t$, $\varepsilon_t \sim WN$. Find the short- and long-run multipliers.

6.2 wytė

7. Multivariate Models

In the present chapter, we discuss methods which involve more than one equation. To motivate why multiple equation methods are important, we begin by discussing **Granger causality** before discussing the most popular class of multiple-equation models: so-called **Vector Autoregressive** (VAR) models. VARs can be used to investigate Granger causality, but are also useful for many other things in finance. Time series models for integrated series are usually based on applying VAR to first differences. However, differencing eliminates valuable information about the relationship among integrated series – this is where **Vector Error Correction** model (VECM) is applicable. Using financial examples, we will show their importance.

7.1. Granger Causality

Recall our warnings made earlier about interpreting correlation and regression results as reflecting causality. For instance, alcohol drinking and lung cancer rates were correlated with one another, even though alcohol drinking does not cause lung cancer. Here correlation did not imply causality. In fact, it was cigarette smoking that caused lung cancer, but a correlation between cigarette smoking and alcohol drinking produced an apparent relationship between alcohol and lung cancer.

In our discussion of regression, we were on a little firmer ground, since we attempted to use common sense in labeling one variable the dependent variable and the others the explanatory variables. In many cases, because the latter "explained" the former it was reasonable to talk about X "causing" Y. For instance, the price of the house can be said to be "caused" by the characteristics of the house (e.g., number of bedrooms, number of bathrooms, etc.). However, one can ran a regression of Y = stock prices in Country B on X = stock prices in Country A. It is possible that stock price movements in Country A cause stock markets to change in Country B (i.e., X causes Y). For instance, if Country A is a big country with an important role in the world economy (e.g., the USA), then a stock market crash in Country A could also cause panic in Country B. However, if Country A and B were neighboring countries (e.g., Thailand and Malaysia) then an event which caused panic in either country could affect both countries. In other words, the causality could run in either direction – or both! Hence, when using the word "cause" with regression or correlation results a great deal of caution has to be taken and common sense has to be used.

However, with time series data we can make slightly stronger statements about causality simply by exploiting the fact that time does not run backward! That is, if event A happens before event B, then it is possible that A is causing B. However, it is not possible that B is causing A. In other words, events in the past can cause events to happen today. Future events cannot.

These intuitive ideas can be investigated through regression models incorporating the notion of **Granger** or regressive **causality**. The basic idea is that a variable X **Granger causes** Y if past values of X can help explain Y. Of course, if Granger causality holds this does not guarantee that X causes Y[1]. This is why we say "Granger causality" rather than just "causality". Nevertheless, if past values of X have explanatory power for current values of Y, it at least su-

[1] If you know the weather forecast (X), it will help you to predict rain (Y); however, the forecast is not the cause of the rain.

ggests that *X* might be causing *Y*. Granger causality is only relevant with time series variables. To illustrate the basic concepts we will consider Granger causality between two variables (*X* and *Y*) which are both stationary. A nonstationary case, where *X* and *Y* have unit roots but are cointegrated, will be mentioned below.

Since we have assumed that X and Y are stationary, the discussion of Ch. 4 suggests an ADL model is appropriate. Suppose that the following simple ADL (only lags on the right hand side!) model holds:

$$Y_t = \alpha + \varphi_1 Y_{t-1} + \beta_1 X_{t-1} + \varepsilon_t .$$

This model implies that last period's value of X has explanatory power for the current value of Y. The coefficient β_1 is a measure of the influence of X_{t-1} on Y_t. If $\beta_1 = 0$, then past values of X have no effect on *Y* and there is no way that *X* could Granger cause *Y*. In other words, if $\beta_1 = 0$ then X does not Granger cause *Y*. An alternative way of expressing this concept is to say that "if $\beta_1 = 0$ then past values of X have no explanatory power for Y beyond that provided by past values of Y ". Since we know how to estimate the ADL and carry out hypothesis tests, it is simple to test Granger causality or, in other words, to test $H_0 : \beta_1 = 0$: if $\hat{\beta}_1$ is statistically significant (e.g., its *p*-value < 0.05), then we conclude that X Granger causes Y. Note that the null hypothesis being tested here is $H_0 : \beta_1 = 0$ which is a hypothesis that Grangercausality **does not occur**. So we should formally refer to the test of $\beta_1 = 0$ as a test of **Granger non-causality**, but we will adopt the more common informal terminology and just refer to this procedure as a **Granger causality test**.

Of course, the above ADL model is quite restrictive in that it incorporates only one lag of X and Y. In general, we could assume that the (X,Y) interaction is described by an ADL(*p*, *q*) model of the form (only lags on the right hand side!):

$$Y_t = \alpha + \delta t + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \beta_1 X_{t-1} + ... + \beta_q X_{t-q} + \varepsilon_t ; \qquad (7.1)$$

we say that X does not Granger cause Y if all $\beta_i = 0$. In practice, we use the methods described in Ch. 6 to select the lag length and then test the joint significance of $\hat{\beta}_i$ (we conclude that X Granger causes Y if any or all of $\hat{\beta}_1, ..., \hat{\beta}_q$ are statistically significant). In other words, if X at any time in the past has explanatory power for the current value of Y, then we say that *X* Granger causes *Y*. Since we are assuming X and Y do not contain unit roots, OLS regression analysis can be used to estimate this model. To test $H_0 : \beta_1 = 0, ..., \beta_q = 0$ we have to compare two models: the **unrestricted** model (7.1) and **restricted** model

$$Y_t = \alpha + \delta t + \varphi_1 Y_{t-1} + ... + \varphi_p Y_{t-p} + \varepsilon_t .$$

We do not reject H_0 if models are „more or less the same", i.e., if $SSR_{UR} \approx SSR_R$. Most popular here is the F test: if test statistics $F = \dfrac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(T-q-(p+2))}$ is greater than the 0.95 quantile of the *F* distribution with $(q, T-q-(p+2))$, we say that *X* Granger causes *Y*.

7.1 example. Do stock price movements in country A Granger cause stock price movements in country B?

133 monthly data on logged stock prices for Countries A and B are located in data set STOCKPAB.XLS, where

lspA	log stock price in country A
lspB	log stock price in country B
pchA	% change in A's price
pchB	% change in B's price

It is easy to test that stock prices in both countries appear to have unit roots, but are not cointegrated (test it yourself). However, the differences of these series are stationary and can be nicely interpreted as stock market returns (exclusive of dividends). We will use these differenced variables to investigate whether stock returns in country A Granger cause those in country B.

```
# generate linear trend
genr time
# create unrestricted model
ols pchA 0 time pchA(-1) pchA(-2) pchA(-3) pchA(-4) \
  pchB(-1) pchB(-2) pchB(-3) pchB(-4)
# create restricted model without pchB; note the Wald F value at the end
omit pchB(-1) pchB(-2) pchB(-3) pchB(-4)

**********

? ols pchA 0 time pchA(-1) pchA(-2) pchA(-3) pchA(-4) pchB(-1) pchB(-2) \
  pchB(-3) pchB(-4)

Model 1: OLS, using observations 6-133 (T = 128)
Dependent variable: pchA

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       -0,609430     0,834940    -0,7299   0,4669
  time         0,0462238    0,0131545    3,514    0,0006   ***
  pchA_1       0,0525275    0,168148     0,3124   0,7553
  pchA_2      -0,0396629    0,168608    -0,2352   0,8144
  pchA_3      -0,0584789    0,167816    -0,3485   0,7281
  pchA_4       0,0359476    0,167210     0,2150   0,8302
  pchB_1       0,853797     0,199501     4,280    3,83e-05 ***
  pchB_2      -0,216937     0,218358    -0,9935   0,3225
  pchB_3       0,233913     0,219217     1,067    0,2881
  pchB_4      -0,271694     0,205388    -1,323    0,1884

R-squared             0,604812   Adjusted R-squared   0,574671
Log-likelihood     -362,2816     Akaike criterion     744,5631
Schwarz criterion   773,0834     Hannan-Quinn         756,1511
rho                -0,015744     Durbin-Watson        2,030963

? omit pchB(-1) pchB(-2) pchB(-3) pchB(-4)

Model 2: OLS, using observations 6-133 (T = 128)
Dependent variable: pchA
             coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const       -0,863807     0,880810    -0,9807   0,3287
```

```
  time         0,0431803    0,0138237     3,124    0,0022    ***
  pchA_1       0,645935     0,0897174     7,200    5,40e-011 ***
  pchA_2      -0,0494193    0,107041     -0,4617   0,6451
  pchA_3       0,0285230    0,106882      0,2669   0,7900
  pchA_4      -0,122913     0,0896255    -1,371    0,1728

Mean dependent var   4,308594   S.D. dependent var   6,550260
Sum squared resid    2548,150   S.E. of regression   4,570172
R-squared            0,532368   Adjusted R-squared   0,513203
F(5, 122)            27,77779   P-value(F)           1,07e-18
Log-likelihood      -373,0541   Akaike criterion     758,1081
Schwarz criterion    775,2203   Hannan-Quinn         765,0609
rho                 -0,002375   Durbin-Watson        2,002890

Comparison of Model 1 and Model 2:

  Null hypothesis: the regression parameters are zero for the variables
    pchB_1, pchB_2, pchB_3, pchB_4

  Test statistic: F(4, 118) = 5,40781, with p-value = 0,000493296
  Of the 3 model selection statistics, 0 have improved.
```

The last lines of the output show that the *p*-value of the *F*-test is much less than 5%, i.e., exclusion of `pchB` makes the sum of residuals SSR_R much bigger than SSR_{UR} and, therefore, we reject the hypothesis $H_0 : \beta_1 = 0,..., \beta_4 = 0$. In other words, the lags of `pchB` improve forecasting, they Granger cause stock returns in country A. ◀◀

In many cases, it is not obvious which way causality should run. For instance, should stock markets in country A affect markets in country B or should the reverse hold? In such cases, when causality may be in either direction, it is important that you check for it. If Y and X are the two variables under study, in addition to running a regression of Y on lags of itself and lags of X (as above), you should also run a regression of X on lags of itself and lags of Y. In other words, you should work with two separate equations: one with Y being the dependent variable and one with X being the dependent variable. This is a simple example of a regression model with more than one equation.

Note that it is possible to find that Y Granger causes X and that X Granger causes Y. In the case of complicated models, such bi-directional causality is quite common and even reasonable. Think, for instance, of the relationship between interest rates and exchange rates. It is not unreasonable that interest rate policy may affect future exchange rates. However, it is also equally reasonable to think that exchange rates may also affect future interest rate policy (e.g., if the exchange rate is perceived to be too high now the central bank may be led to decrease interest rates in the future).

7.1 example (continued)

```
# create unrestricted model
ols pchB 0 time pchA(-1) pchA(-2) pchA(-3) pchA(-4) \
pchB(-1) pchB(-2) pchB(-3) pchB(-4)
# create restricted model without pchA; note the Wald F value at the end
omit pchA(-1) pchA(-2) pchA(-3) pchA(-4)
**********************************
[...]
Comparison of Model 3 and Model 4:
[...]
```

```
Null hypothesis: the regression parameters are zero for the variables
  pchA_1, pchA_2, pchA_3, pchA_4

Test statistic: F(4, 118) = 0,0710735, with p-value = 0,990678
Of the 3 model selection statistics, 3 have improved.
```

Now we do not reject $H_0 : \beta_1 = 0, ..., \beta_4 = 0$, i.e., pchA does not Granger cause pchB. ◀◀

This brief discussion of Granger causality has focussed on two variables, X and Y. However, there is no reason why these basic techniques cannot be extended to the case of many variables. For instance, if we had three variables, X, Y and Z, and were interested in investigating whether X or Z Granger cause Y, we would simply regress Y on lags of Y, lags of X and lags of Z. If, say, the lags of Z were found to be significant and the lags of X not, then we could say that Z Granger causes Y, but X does not.

Testing for Granger causality among cointegrated variables is very similar to the method outlined above. Remember that, if variables are found to be cointegrated (something which should be investigated using unit root and cointegration tests), then you should work with an error correction model (ECM) involving these variables. In the case where you have two variables, this is given by:

$$\Delta Y_t = \varphi + \delta t + \lambda e_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_p \Delta Y_{t-p} + \omega_1 \Delta X_{t-1} + ... + \omega_q \Delta X_{t-q} + \varepsilon_t$$

As noted in Ch. 6, this is essentially an ADL model except for the presence of the term λe_{t-1}. Remember that $e_{t-1} = Y_{t-1} - \alpha - \beta X_{t-1}$, an estimate of which can be obtained by running a regression of Y on X and saving the residuals. Intuitively, X Granger causes Y if past values of X have explanatory power for current values of Y. Applying this intuition to the ECM, we can see that past values of X appear in the terms $\Delta X_{t-1}, ..., \Delta X_{t-q}$ and e_{t-1}. This implies that X does not Granger cause Y if $\omega_1 = ... = \omega_q = \lambda = 0$. Chapter 6 discussed how we can use two OLS regressions to estimate ECMs, and then use their *p*-values or confidence intervals to test for causality. Thus, *t*-statistics and *p*-values can be used to test for Granger causality in the same way as the stationary case. Also, the *F* - tests can be used to carry out a formal test of $H_0 : \omega_1 = ... = \omega_q = \lambda = 0$.

The bottom line – if X Granger-causes Y, this does not mean that X causes Y, it only means that X improves Y 's predictability (i.e., reduces residuals of the model).

7.2. VAR: Estimation and Forecasting

Our discussion of Granger causality naturally leads us to an interest in models with several equations and the topic of **Vector Autoregressions or VARs**. Before discussing their popularity and estimation, we will first define what a VAR is. Initially, we will assume that all variables are stationary. If the original variables have unit roots, then we assume that differences have been taken such that the model includes the changes in the original variables (which do not have unit roots). The end of this section will consider the extension of this case to that of cointegration.

When we investigated Granger causality between *X* and *Y*, we began with an ADL(*p*, *q*) model for *Y* as the dependent variable. We used it to investigate if *X* Granger caused *Y*. We then went on to consider causality in the other direction, which involved switching the roles of *X* and *Y* in the ADL; in particular, X became the dependent variable. We can write the two equations as follows:

$$Y_t = \alpha_1 + \delta_1 t + \varphi_{11} Y_{t-1} + ... + \varphi_{1p} Y_{t-p} + \beta_{11} X_{t-1} + ... + \beta_{1q} X_{t-q} + \varepsilon_{1t}$$

and

$$X_t = \alpha_2 + \delta_2 t + \varphi_{21} Y_{t-1} + ... + \varphi_{2p} Y_{t-p} + \beta_{21} X_{t-1} + ... + \beta_{2q} X_{t-q} + \varepsilon_{2t}.$$

The first of these equations tests whether *X* Granger causes Y; the second, whether *Y* Granger causes *X*. Note that now the coefficients have subscripts indicating which equation they are in. For instance, α_1 is the intercept in the first equation, and α_2 the intercept in the second. Furthermore, the errors now have subscripts to denote the fact that they will be different in the two equations.

These two equations comprise a VAR. A VAR is the extension of the autoregressive (AR) model to the case in which there is more than one variable under study. Remember that the AR model introduced in Ch. 3 involved one dependent variable Y_t which depended only on lags of itself (and possibly a deterministic trend). A VAR has more than one dependent variable (e.g., *Y* and *X*) and, thus, has more than one equation (e.g., one where Y_t is the dependent variable and one where X_t is). Each equation uses as its explanatory variables lags of *all the variables under study* (and possibly a deterministic trend).

The term „VAR“ becomes more transparent if we use a matrix notation. A <u>first</u> order VAR in <u>two</u> variables would be given by

$$\begin{aligned} Y_t &= \alpha_1 + \varphi_{11} Y_{t-1} + \varphi_{12} X_{t-1} + \varepsilon_{1t} \\ X_t &= \alpha_2 + \varphi_{21} Y_{t-1} + \varphi_{22} X_{t-1} + \varepsilon_{2t} \end{aligned} \qquad (7.2)$$

where ε_{1t} and ε_{2t} are two white noise processes (independent of the history of *Y* and *X*) that may be correlated. If, for example, $\varphi_{12} \neq 0$, this means that the history of X helps explaining *Y*, that is, X is a Granger cause of Y. The system (7.2) can be written as

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \qquad (7.3)$$

or, with relevant definitions, as

$$\vec{Y}_t = \vec{\alpha} + \Theta_1 \vec{Y}_{t-1} + \vec{\varepsilon}_t \qquad (7.4)$$

where $\vec{Y}_t = (Y_t, X_t)'$ is a column vector, Θ_1 is a 2×2 matrix etc. This extends the first order autoregressive model AR(1) from Ch. 2 to the more dimensional case. In general, a VAR(*p)* model for a M - dimensional vector $\vec{Y}_t$ is given by

$$\vec{Y}_t = \vec{\alpha} + \vec{\delta} t + \Theta_1 \vec{Y}_{t-1} + ... + \Theta_p \vec{Y}_{t-p} + \vec{\varepsilon}_t \qquad (7.5)$$

where each Θ_j is a $d \times d$ matrix and $\vec{\varepsilon}_t$ is a d - dimensional vector of white noise terms with covariance matrix Σ. For example,

$$\begin{pmatrix} Y_t \\ X_t \\ Z_t \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varphi_{11}^{(1)} & \varphi_{12}^{(1)} & \varphi_{13}^{(1)} \\ \varphi_{21}^{(1)} & \varphi_{22}^{(1)} & \varphi_{23}^{(1)} \\ \varphi_{31}^{(1)} & \varphi_{32}^{(1)} & \varphi_{33}^{(1)} \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ X_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} \varphi_{11}^{(2)} & \varphi_{12}^{(2)} & \varphi_{13}^{(2)} \\ \varphi_{21}^{(2)} & \varphi_{22}^{(2)} & \varphi_{23}^{(2)} \\ \varphi_{31}^{(2)} & \varphi_{32}^{(2)} & \varphi_{33}^{(2)} \end{pmatrix} \begin{pmatrix} Y_{t-2} \\ X_{t-2} \\ Z_{t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix}$$

is a VAR(2) model in three variables. The VAR(p) model implies univariate ARMA models for each of its components (the AR order is (at most) $d\,p$ and MA order (at most) $(d-1)p$ [HBFKD, p. 660]). However, the advantages of considering the components simultaneously include that the model may be more parsimonious and includes fewer lags, and that more accurate forecasting is possible, because the information set is extended to also include the history of the other variables. Determining the lag length p in an empirical application is not always easy and univariate autocorrelation or partial autocorrelation functions will not help. A reasonable strategy is to estimate a VAR model for different values of p and then select on the basis of the Akaike or Schwarz information criteria.

Similarly to one-dimensional case, a VAR(p) is stationary if all the roots of the equation $\det(I_k - \Theta_1 z - \Theta_2 z^2 - ... - \Theta_p z^p) = 0$ are outside a unit complex circle. The VAR is said to have a single unit root if the above equation has exactly one root $z = +1$, i.e., $\det(I_k - \Theta_1 - \Theta_2 - ... - \Theta_p) = 0$. This will hold if at least one of the variables in the VAR contains a unit root.

Why we would want to work with such models? One reason has to be Granger causality testing. That is, VARs provide a framework for testing for Granger causality between each set of variables. Should interest rates cause exchange rates to change or vice versa? Both? Should GDP growth cause interest rates to change? The opposite? Both? However, there are also many other reasons. For instance, a point which we will discuss below is that VARs are often used for forecasting. However, financial researchers also use VARs in many other contexts. Models involving so-called present value relationships often work with VARs using the (log) dividend-price ratio and dividend growth. VARs have been used to investigate issues relating to the term structure of interest rates (using interest rates of various maturities, interest rate spreads, etc.), intertemporal asset allocation (using returns on various risky assets), the rational valuation formula (using the dividend-price ratio and returns), the interaction of bond and equity markets (using stock and bond return data), etc. In other words, VARs have been used in a wide variety of financial problems.

7.2 example. A VAR(1) with RMPY variables

Economists often use such important macroeconomic variables as: `R` = the interest rate, `M` = the money supply, `P` = the price level and `Y` = real GDP. Due to the symbols used, models using these variables are sometimes informally referred to as RMPY models (pronounced

"rumpy"). The file rmpy.xls contains quarterly data on the variables for the US from 1947:1 through 1992:4. To be precise:

- `r` three-month Treasury bill rate
- `m` money supply (M2) measured in billions of dollars
- `p` price level measured by the GDP deflator (a price index with 1987 = 1.00)
- `y` GDP measured in billions of 1987 dollars

Before carrying out an analysis using time series data, you must conduct unit root tests. Remember that, if unit roots are present but cointegration does not occur, then the spurious regression problem exists. In this case, you should work with differenced data. Alternatively, if unit roots exist and cointegration does occur, then you will have important economic information that the series are trending together and use ECM.

In the present case, tests indicate (check) that we cannot reject the hypothesis that unit roots exist in all variables and that cointegration does not occur. In order to avoid the spurious regression problem, we work with differenced data. In particular, we take logs of each series, then take differences of these logged series, then multiply them by 100. This implies that we are working with percentage changes in each variable (e.g., a value of 1 implies a 1% change). Thus,

- `dr` percentage change in the interest rate.
- `dm` percentage change in the money supply.
- `dp` percentage change in the price level (i.e., inflation).
- `dy` percentage change in GDP (i.e., GDP growth).

We choose somewhat arbitrarily a VAR(1) model with a linear trend.

```
var 1 dR dM dP dY ; time
```

In Table 7.1 we present a short version of the output of the `var` function.

Table 7.1. The RMPY VAR(1) using d*R*, d*M*, d*P* and d*Y* as dependent variables

	Dependent Variable ΔR		Dependent Variable ΔM		Dependent Variable ΔP		Dependent Variable ΔY	
	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.
Intercept	−3.631	0.162	0.335	0.001	0.161	0.138	0.495	0.005
ΔR_{t-1}	0.222	0.003	−0.013	2.0E − 5	0.010	0.002	3.8E − 4	0.940
ΔM_{t-1}	3.391	0.007	0.749	1.0E − 33	0.121	0.021	0.283	9.3E − 4
ΔP_{t-1}	1.779	0.228	0.061	0.303	0.519	1.0E − 14	−0.117	0.242
ΔY_{t-1}	3.224	0.004	−0.032	0.480	−0.039	0.407	0.309	7.0E − 5
Time	−0.056	0.011	3.4E − 4	0.695	0.002	0.048	−0.003	0.035

If we examine the significant coefficients (i.e., those with p-value less than 0.05), some interesting patterns emerge. First, in every equation, the lag of the dependent variable is significant. For example, in the equation with ΔR_t as the dependent variable, ΔR_{t-1} provides significant explanatory power. Secondly, the results for the four equations demonstrate some inte-

resting patterns of Granger causality. In the equation with ΔR as the dependent variable, we can see that both GDP growth and money growth Granger cause interest rate changes. In other words, past values of GDP and money growth have explanatory power for current interest rate changes. In the case of the $\Delta R / \Delta M$ (interest rate/money supply) relationship, the equation with ΔM as the dependent variable shows that the causality flows in both directions since interest rate changes also Granger cause money growth. However, interest rate changes do not Granger cause GDP growth. The Granger causality results in respect to inflation are particularly interesting since it can be seen that inflation does not Granger cause any other variable, but that both ΔR and ΔM Granger cause inflation. A macroeconomist could use these results to address theoretical questions of interest (e.g., Is inflation purely a monetary phenomenon? Are monetarist views of the economy supported? Are Keynesian views of the economy supported? Is the real economy affected by inflation?, etc.), but it is beyond the scope of this course to discuss them in detail.

The results in the previous example are based on a VAR(1). That is, we set $p = 1$ and used one lag of each variable to explain the dependent variable. In general, of course, we might face the cases where the interaction between the components of Y may well be described by VAR(p), $p > 1$. The literature on choosing the right lag in VARs is voluminous, but usually the selection is based on the generalized versions of AIC, BIC and like criteria. The following GRETL command allows to compare three models with lags 1, 2, and 3:

```
var 3 dY dP dR dM; time -- lagselect

lags        loglik    p(LR)        AIC           BIC           HQC

   1   -1225,28426             13,880936    14,306664*   14,053550
   2   -1194,73952  0,00000    13,719328    14,428874    14,007018*
   3   -1176,87524  0,00316    13,698614*   14,691978    14,101380
```

The most restrictive BIC recommends the 1st order and most liberal AIC the 3rd order. Note that VAR(3) must estimate (1+4+4+4+1)*4=56 coefficients which means a loss of many degrees of freedom and not very accurate estimates (below, we shall choose $p = 2$). ◀◀

Once the order p has beeb established, we have to estimate the coefficients in (7.5). It appears that to get BLUE&C estimators, we can apply OLS to every equation individually (this is what, in GRETL, the `var` procedure does and, in R, `VAR` of the the package vars or `lineVar` of the package tsDyn, do).

Below we shall examine **forecasting** a VAR(2) model, but at first we present a brief introduction to some of the practical issues and intuitive ideas relating to forecasting. All our discussion will relate to forecasting with VARs but it is worth noting that the ideas also relate to forecasting with univariate time series models. After all, an AR model is just a VAR with only one equation.

Forecasting is usually done using time series variables. The idea is that you use your observed data to predict what you expect to happen in the future. In more technical terms, you use data for periods $t = 1, \ldots, T$ to forecast periods $T + 1$, $T + 2$, etc. To provide some intuition for how forecasting is done, consider a VAR(1) involving two variables, Y and X:

$$Y_t = \alpha_1 + \delta_1 t + \varphi_{11} Y_{t-1} + \varphi_{12} X_{t-1} + \varepsilon_{1t}$$

and

$$X_t = \alpha_2 + \delta_2 t + \varphi_{21} Y_{t-1} + \varphi_{22} X_{t-1} + \varepsilon_{2t}\,.$$

You cannot observe Y_{T+1} but you want to make a guess of what it is likely to be. Using the first equation of the VAR and setting $t = T + 1$, we obtain an expression for Y_{T+1}:

$$Y_{T+1} = \alpha_1 + \delta_1 (T+1) + \varphi_{11} Y_T + \varphi_{12} X_T + \varepsilon_{1,T+1}$$

This equation cannot be directly used to obtain Y_{T+1} since we don't know $\varepsilon_{1,T+1}$. In words, we don't know what unpredictable shock or surprise will hit the economy next period. Furthermore, we do not know what the coefficients are. However, if we ignore the error term (which cannot be forecast since it is unpredictable) and replace the coefficients by their estimates we obtain a forecast which we denote as $\hat{Y}_{T+1}$:

$$\hat{Y}_{T+1} = \hat{\alpha}_1 + \hat{\delta}_1 (T+1) + \hat{\varphi}_{11} Y_T + \hat{\varphi}_{12} X_T$$

We can use the same strategy for two periods, provided that we make one extension. In the one period case, we used X_T and Y_T to create $\hat{Y}_{T+1}$ and $\hat{X}_{T+1}$. In the two period case, $\hat{Y}_{T+2}$ and $\hat{X}_{T+2}$ depend on Y_{T+1} and X_{T+1}. But since our data only runs until period *T*, we do not know what Y_{T+1} and X_{T+1} are. Consequently, we replace Y_{T+1} and X_{T+1} by $\hat{Y}_{T+1}$ and $\hat{X}_{T+1}$ (this called a dynamic forecast). That is, use the relevant equation from the VAR, ignore the error, replace the coefficients by their estimates and replace past values of the variables that we do not observe by their forecasts. In a formula:

$$\hat{Y}_{T+2} = \hat{\alpha}_1 + \hat{\delta}_1 (T+2) + \hat{\varphi}_{11} \hat{Y}_{T+1} + \hat{\varphi}_{12} \hat{X}_{T+1}\,.$$

$\hat{X}_{T+2}$ can be calculated in a similar manner using the formula

$$\hat{X}_{T+2} = \hat{\alpha}_2 + \hat{\delta}_2 (T+2) + \hat{\varphi}_{21} \hat{Y}_{T+1} + \hat{\varphi}_{22} \hat{X}_{T+1}$$

We can use the general strategy of ignoring the error, replacing coefficients by their estimates and replacing lagged values of variables that are unobserved by forecasts, to obtain forecasts for any number of periods in the future for any VAR(*p*).

The above described strategy of **dynamic** forecast is implemented in GRETL. As always, we can choose either GRETL terminal or menu based approach. This time we shall proceed from the toolbar. Assuming you have already imported rmpy.xls, go to Model|Time series|Vector Autoregression..., select `dY`, `dP`, `dR`, `dM`, and lag order: 2, tick Include a trend, and click OK. In the window appeared, we can go to Analysis and explore different possibilities. For example, to forecast `dY` for the comming 12 quarters in 1993:1 – 1995:4, go to Forecast section and fill up relevant boxes.

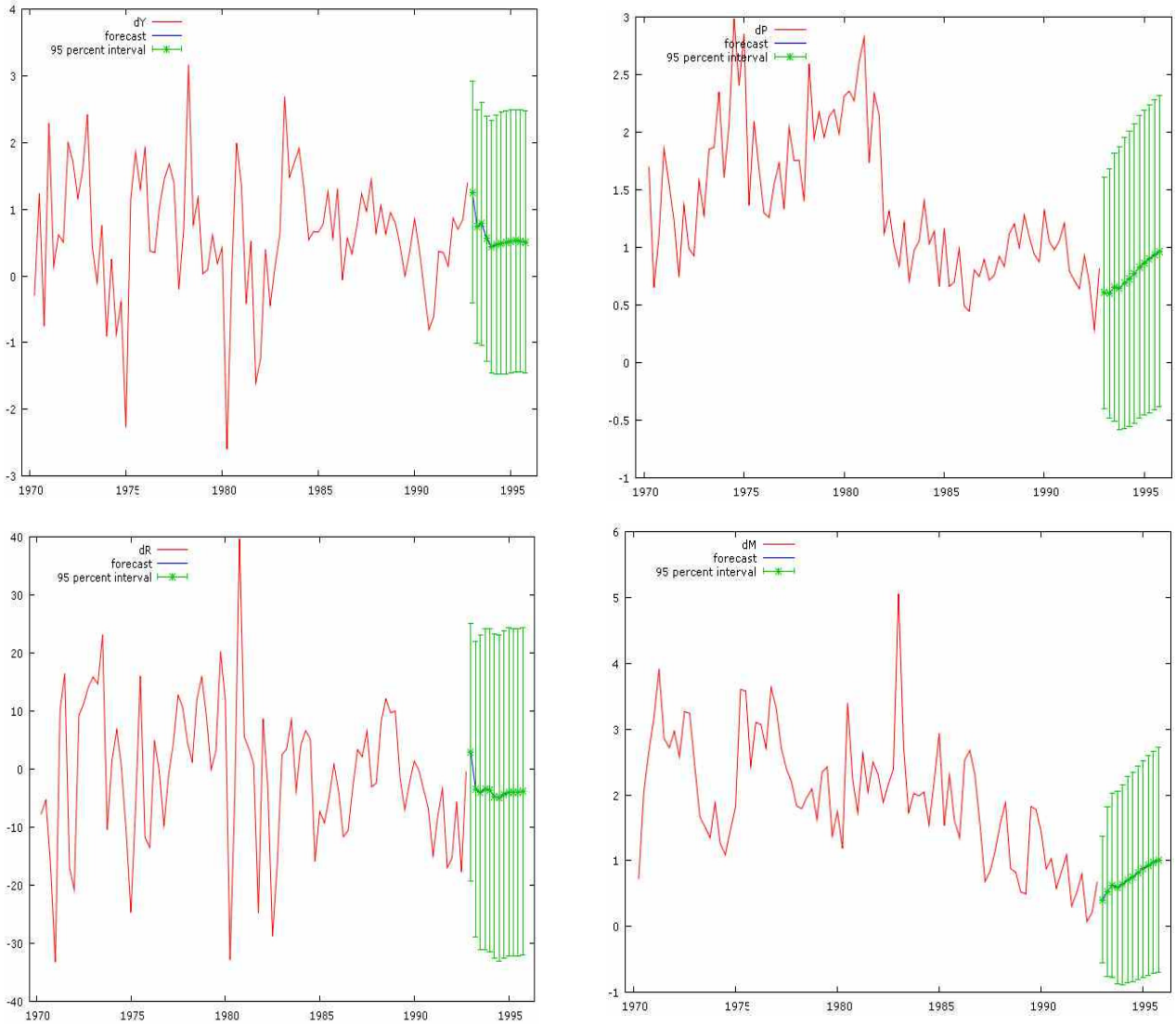


Figure 7.1. 12 quarters forecast in VAR(2) model for `dY`, `dP`, `dR`, and `dM`

In summary, building a VAR model involves three steps: (a) use some information criterion to identify the order, (b) estimate the specified model by using the least squares method and, if necessary, reestimate the model by removing statistically insignificant parameters, and (c) use the Portmanteau test statistic of the residuals to check the adequacy of a fitted model (this is a multivariate analogue of Q-stat in ARIMA model and is to test for autocorrelation and cross-correlation in residuals). If the fitted model is adequate, then it can be used to obtain forecasts.

7.3 example. The file var1a.txt contains 100 elements of (trend) stationary VAR(1) simulated with R (the code is in Computer Labs, 7.4 example) and described by the formula

$$\vec{Y}_t = \begin{pmatrix} 1.5 \\ 4 \end{pmatrix} + \begin{pmatrix} 0 \\ 0.08 \end{pmatrix} t + \begin{pmatrix} 1.2 & -0.5 \\ 0.6 & 0.3 \end{pmatrix} \vec{Y}_{t-1} + \vec{\varepsilon}_t, \quad \text{cov}(\vec{\varepsilon}) = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.25 \end{pmatrix}, \quad t = 1, ..., 100.$$

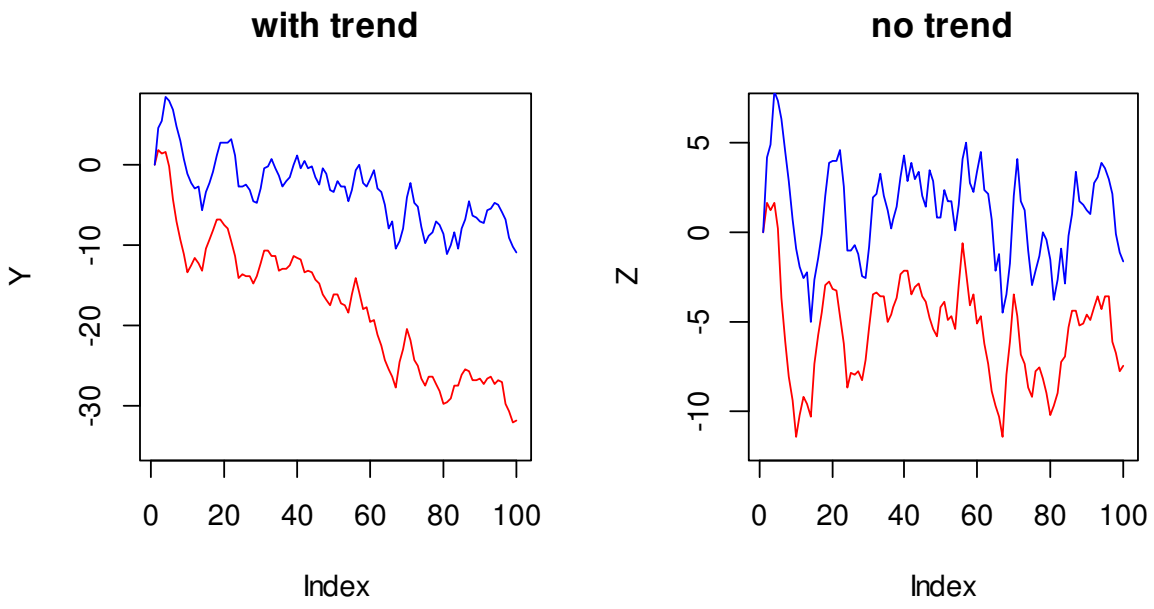


Figure 7.2. The trend coefficient $\delta_1 = 0$ but, because of the interaction between components, Y_{1t} is nevertheless trending (left); if we also take $\delta_2 = 0$, both components become stationary (why?; right)

Now we pretend that we have forgotten the formula and shall try to „restore“ the model. The first step in describing $\vec{Y}$ is to choose the „right“ order: go to Model * Time series * VAR lag selection and fill the respective boxes with 4, `y1` and `y2` and tick Include a constant and Include a trend – all the criterions recommend the 1st order (this is in accordance with our formula). Next, we shall estimate the coefficients of VAR(1): go to Model * Time series * Vector Autoregression and fill the boxes as previously.

```
Equation 1: v1

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       2.00766      0.380238     5.280     8.16e-07  ***
  v1_1        1.26420      0.0528852   23.90      5.73e-042 ***
  v2_1       -0.539874     0.0568344   -9.499     1.94e-015 ***
  time        0.0117678    0.0106652    1.103     0.2726

Equation 2: v2

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       4.46168      0.418784    10.65      6.67e-018 ***
  v1_1        0.657071     0.0582464   11.28      3.13e-019 ***
  v2_1        0.257945     0.0625959    4.121     8.07e-05  ***
  time        0.0899141    0.0117464    7.655     1.60e-011 ***
```

(note that all the estimates are close to the true values of the coefficients.) Finally, we shall predict both components for 20 periods ahead: in the vector autoregression model window, go to Analysis * Forecasts etc

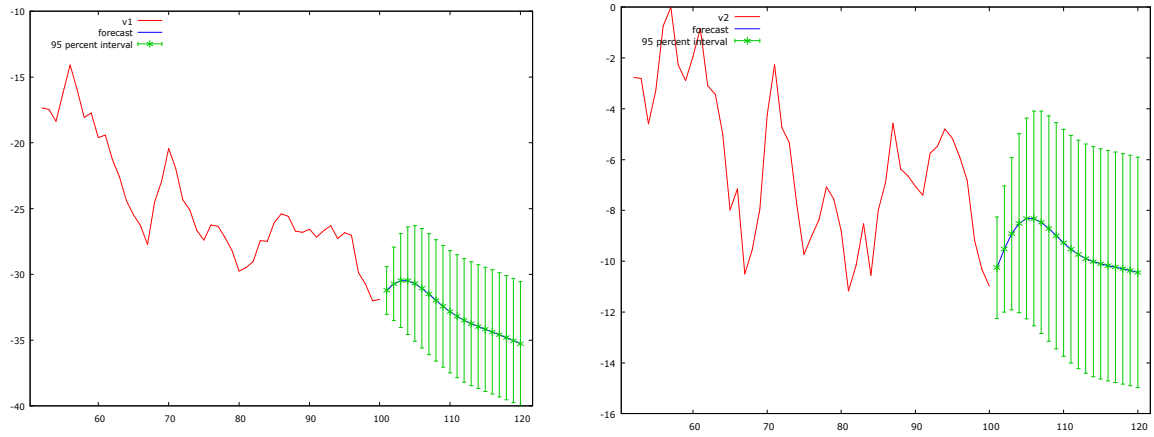


Figure 7.3. 20-periods-ahead forecasts for Y_1 and Y_2. Note different scales in the two graphs and a "strange" (nonmonotone) behavior of the forecasts.

7.1 exercise. a) The file var3a.txt contains a 150-long realization of VAR(1) described by the formula $\vec{Y}_t = \begin{pmatrix} 5/8 & 1/2 \\ 1/4 & 1/3 \end{pmatrix}\vec{Y}_{t-1} + \vec{\varepsilon}_t$, $\mathrm{cov}(\vec{\varepsilon}) = \begin{pmatrix} 64 & 0 \\ 0 & 64 \end{pmatrix}$. Plot $\vec{Y}_t$. Test both coordinates for the presence of unit root. Repeat the analysis of 8.3 example. b) The file var3b.txt contains a 150-long realization of VAR(1) described by the formula $\vec{Y}_t = \begin{pmatrix} 5/8 & 1/2 \\ 1/4 & 2/3 \end{pmatrix}\vec{Y}_{t-1} + \vec{\varepsilon}_t$, $\mathrm{cov}(\vec{\varepsilon}) = \begin{pmatrix} 64 & 0 \\ 0 & 64 \end{pmatrix}$. Plot $\vec{Y}_t$. Test both components for the presence of unit root.

7.4 example. ⇒ **In R**

The data file us-tbill.txt contains monthly, 1964:01 through 1993:12, interest rates of US treasure bills for maturities of one month `Y_1M` and five years `Y_5Y`. Both series are integrated, therefore we fit a VAR(2) model to the first differences[2] (the 2nd order was selected with `VARselect` by Schwarz's SC).

```
rate=ts(read.table(file.choose(),header=TRUE),start=1964,freq=12)
rate
      date  Y_1M  Y_1Y  Y_5Y
[1,]     1 3.419 3.789 3.984
[2,]     2 3.500 3.947 4.023
.........................
rrate=rate[,c(2,4)]
library(vars)
VARselect(rrate, lag.max = 5, type="const")
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     5      2      2      5
$criteria
                1           2           3           4           5
```

[2] Since the series are cointegrated (check), we should also include an error correction term in our model. We shall do it later (see 7.4 exercise).

```
AIC(n) -2.33355888 -2.4113507 -2.40700353 -2.40319058 -2.41265107
HQ(n)  -2.30752343 -2.3679583 -2.34625415 -2.32508424 -2.31718775
SC(n)  -2.26811464 -2.3022769 -2.25430029 -2.20685785 -2.17268884
library(urca)
N=120      # prediction horizon = 120(months)
par(mfrow=c(1,2))
d.rate=diff(rrate)                                  # create differences.
var.diff=VAR(d.rate, p = 2, type = "none")          # „none“ because rrate
print(summary(var.diff))                            # does not drift

Endogenous variables: Y_1M, Y_5Y
Deterministic variables: none

Estimation results for equation Y_1M:
=====================================
Y_1M = Y_1M.l1 + Y_5Y.l1 + Y_1M.l2 + Y_5Y.l2

         Estimate Std. Error t value Pr(>|t|)
Y_1M.l1 -0.19831    0.05795  -3.422 0.000694 ***
Y_5Y.l1  0.62395    0.10715   5.823  1.3e-08 ***
Y_1M.l2  0.01030    0.05651   0.182 0.855530
Y_5Y.l2 -0.27583    0.11218  -2.459 0.014415 *

Multiple R-Squared: 0.1163

Estimation results for equation Y_5Y:
=====================================
Y_5Y = Y_1M.l1 + Y_5Y.l1 + Y_1M.l2 + Y_5Y.l2

         Estimate Std. Error t value Pr(>|t|)
Y_1M.l1  0.01457    0.03143   0.463   0.6433
Y_5Y.l1  0.06346    0.05812   1.092   0.2756
Y_1M.l2  0.04838    0.03065   1.578   0.1154
Y_5Y.l2 -0.12976    0.06084  -2.133   0.0336 *

Multiple R-Squared: 0.0187
```

The model `var.diff` is for differences. Since $Y_{T+h} = \Delta Y_{T+h} + ... + \Delta Y_{T+1} + Y_T$, the forecast $Y_{T+h,T}$ equals the cumulative sum of the forecasts for differences:

```
var.pred=predict(var.diff, n.ahead = N, ci = 0.95)
R1.d=numeric(360+N);R2.d=numeric(360+N)
# insert historical data
R1.d[1:360]=rrate[,1]
R2.d[1:360]=rrate[,2]
# predict levels from differences
R1.d[361:(360+N)]= R1.d[360]+cumsum(var.pred$fcst[["Y_1M"]][,1])
R2.d[361:(360+N)]= R2.d[360]+cumsum(var.pred$fcst[["Y_5Y"]][,1])
Ra1.d=ts(R1.d,start=1964,freq=12)
Ra2.d=ts(R2.d,start=1964,freq=12)
plot(Ra1.d,ylab="Rates",main="VAR in differences")
lines(Ra2.d,col=2)
legend(1992,15,c("Y_1M","Y_5Y"),lty=1,col=1:2)      # see Fig. 7.4, left
```

The estimation results show that `Y_5Y` is the Granger cause of `Y_1M` but not vice versa. Also note that the dependence of the one-month rate on the five-year rate is much stronger (as can be seen from R^2 : $0.1163 > 0.0187$).

Figure 7.4 (left) shows the dynamic out-of-sample forecasts (starting in January 1994) of the two interest rates. Later, in 7.4 exercise, we shall correct the specification by adding a cointegration term to the right-hand sides of the model.

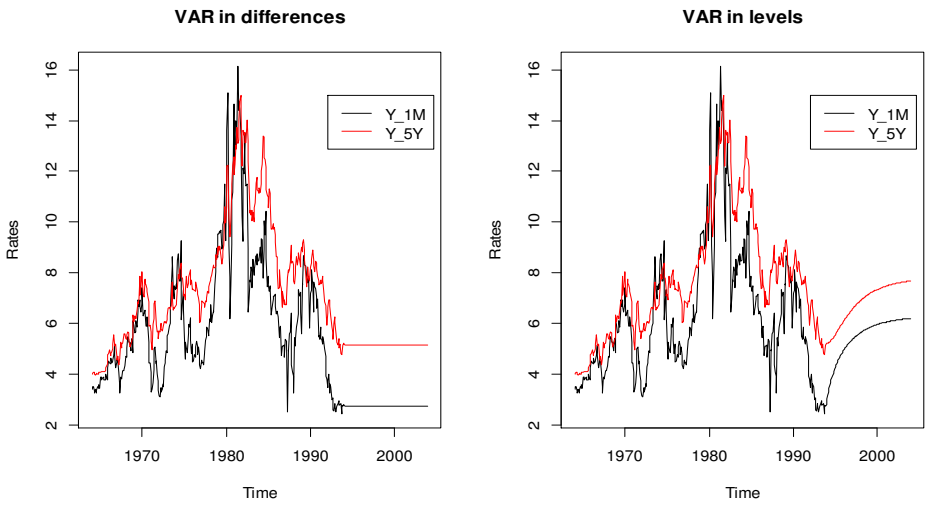


Figure 7.4. Forecasting the levels with two models - VAR in differences (left) and VAR in levels (right)

Note that VAR model should always include variables with the same order of integration – this allows us to create a model in levels[3] (instead of differences.) However, this model again has some drawbacks[4].

⇒In R

```
var.lev=VAR(rrate, p = 2, type = "const")
summary(var.lev)

Endogenous variables: Y_1M, Y_5Y
Deterministic variables: const

Estimation results for equation Y_1M:
=====================================
Y_1M = Y_1M.l1 + Y_5Y.l1 + Y_1M.l2 + Y_5Y.l2 + const

         Estimate Std. Error t value Pr(>|t|)
Y_1M.l1  0.71030    0.05664  12.540  < 2e-16 ***
Y_5Y.l1  0.70251    0.10873   6.461 3.46e-10 ***
Y_1M.l2  0.17968    0.05690   3.158  0.00173 **
Y_5Y.l2 -0.62927    0.10615  -5.928 7.31e-09 ***
const    0.12092    0.13944   0.867  0.38644
Multiple R-Squared: 0.913

Estimation results for equation Y_5Y:
=====================================
Y_5Y = Y_1M.l1 + Y_5Y.l1 + Y_1M.l2 + Y_5Y.l2 + const

         Estimate Std. Error t value Pr(>|t|)
Y_1M.l1  0.01034    0.03086   0.335   0.7379
```

[3] See [H], p.651 or the last pages of this chapter.
[4] See also [E2], p. 358.

```
Y_5Y.l1   1.03517    0.05923  17.477   <2e-16 ***
Y_1M.l2   0.02965    0.03100   0.957   0.3394
Y_5Y.l2  -0.09011    0.05783  -1.558   0.1201
const     0.17660    0.07596   2.325   0.0206 *

Multiple R-Squared: 0.9703
```

```
var.pred.lev=predict(var.lev, n.ahead = N, ci = 0.95)
R1.lev=numeric(360+N)
R2.lev=numeric(360+N)
R1.lev[1:360]=rrate[,1]
R2.lev[1:360]=rrate[,2]
R1.lev[361:(360+N)]=var.pred.lev$fcst[["Y_1M"]][,1]
R2.lev[361:(360+N)]=var.pred.lev$fcst[["Y_5Y"]][,1]
Ra1.lev=ts(R1.lev,start=1964,freq=12)
Ra2.lev=ts(R2.lev,start=1964,freq=12)
plot(Ra1.lev,ylab="Rates",main="VAR in levels")
lines(Ra2.lev,col=2)
legend(1992,15,c("Y_1M","Y_5Y"),lty=1,col=1:2)
```

Both levels (see Fig. 7.4, right) revert to (the values close to) their means[5] (check) which contradicts the unit root behavior of each series. The explanation lies in the fact that we estimated unrestricted[6] VAR model while actually the coefficients should reflect cointegration and obey some constrains (read the VECM section). ◀◀

⇒ In gretl

Lag selection (Model|Time series|VAR lag selection...)

```
lags        loglik    p(LR)       AIC          BIC          HQC

   1    -573.10448             3.447050     3.515213     3.474222
   2    -556.33423  0.00000    3.371037     3.484642*    3.416323*
   3    -553.24230  0.18583    3.376442     3.535489     3.439843
   4    -550.10035  0.17893    3.381550     3.586038     3.463064
   5    -544.67823  0.02837    3.373085     3.623015     3.472714
   6    -539.60533  0.03804    3.366698*    3.662070     3.484442
```

We choose (as in R) lag=2. Now go to Model|Time series|Vector Autoregression...:

```
  Equation 1: Y_1M

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        0.120917    0.139439     0.8672   0.3864
  Y_1M_1       0.710303    0.0566441   12.54     4.36e-030 ***
  Y_1M_2       0.179677    0.0569002    3.158    0.0017    ***
  Y_5Y_1       0.702505    0.108732     6.461    3.46e-010 ***
  Y_5Y_2      -0.629267    0.106154    -5.928    7.31e-09  ***

[...]
```

[5] This is what we expect for stationary sequences.
[6] No restrictions on the coefficients.

```
Equation 2: Y_5Y

             coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        0.176603     0.0759583     2.325    0.0206     **
  Y_1M_1       0.0103356    0.0308565     0.3350   0.7379
  Y_1M_2       0.0296520    0.0309960     0.9566   0.3394
  Y_5Y_1       1.03517      0.0592311    17.48     1.02e-049 ***
  Y_5Y_2      -0.0901103    0.0578267    -1.558    0.1201

[...]
```

These equations coincide with those obtained with R. To forecast, in the window appeared, go to Analysis | Forecasts|...|Number of observations to add 120 – you will get the same graphs as before, in Fig. 8.4:

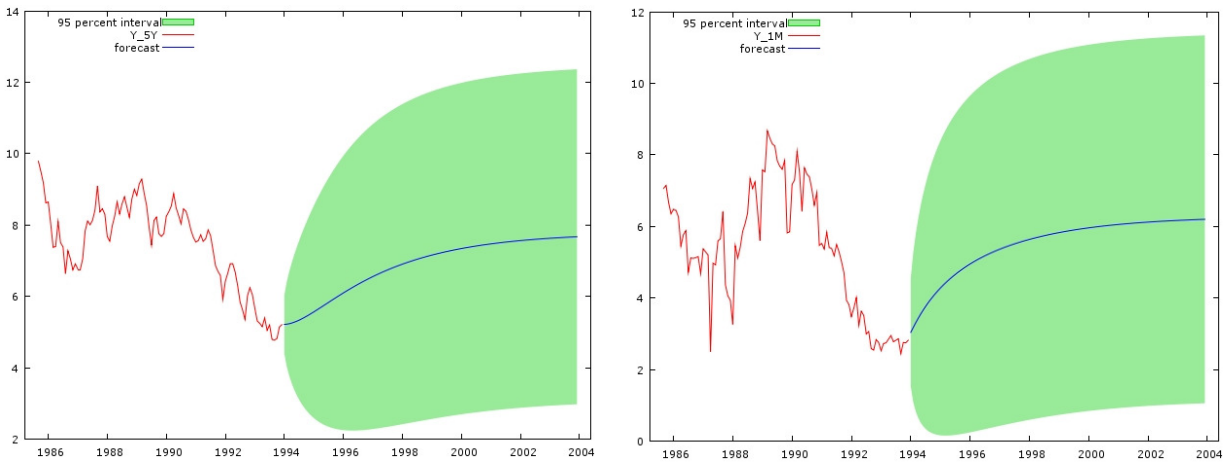


Figure 7.5. Historical data and forecast of Y_5Y (left) and the same for Y_1M (right)

7.2 exercise. The file ccva.txt contains 1893-1997 yearly data on prices and dividends on S&P 500 stocks as well as data on inflation and short-term interest rates.

```
          DYD       RTBILL       SPREAD        XBOND       XSTOCK       YIELD
1   -2.849642  0.149256883 -0.005428600  0.009500733 -0.245791525 0.049742092
2   -3.002845  0.036565583  0.010258374 -0.007351078 -0.017624430 0.034208171
.....................................................................
```

Fit a VAR model using all series in the file and interpret the results. Fit a VAR model using only data from 1893 to 1981. Obtain dynamic forecasts for all series until 1997 and interpret the results.

7.3. VAR: Impulse-Response Function

The **impulse-response function** is yet another device that helps us to learn about the dynamic properties of vector autoregressions of interest to forecasters. The question of interest is simple and direct: How does a unit innovation to a series affect it, now and in the future?

To clarify the issue, let us start with one-dimensional case. Let $Y_1 = ... = Y_{T-1} = 0$, $\varepsilon_1 = ... = \varepsilon_{T-1} = 0$ and at moment $t = T$ a unit shock comes: $\varepsilon_T = \sigma$, $\varepsilon_{T+1} = ... = 0$.

- If $Y_t = \varepsilon_t$, i.e., Y_t is WN, then $Y_T = \sigma, Y_{T+1} = \varepsilon_{T+1} = 0$ and $Y_{T+h} = \varepsilon_{T+h} \equiv 0$ - WN has no memory, no dynamics.
- If $Y_t = \varphi Y_{t-1} + \varepsilon_t, |\varphi| < 1$, i.e., Y_t is AR(1), then $Y_T = \varphi \cdot 0 + \sigma = \sigma$, $Y_{T+1} = \varphi\sigma + 0 = \varphi\sigma$, ..., $Y_{T+h} = \varphi^h \sigma \to 0$ as $h \to \infty$ - the impulse response is dying down.

Now consider again the two-variable, first-order system

$$Y_t = \varphi_{11} Y_{t-1} + \varphi_{12} X_{t-1} + \varepsilon_{1t},$$
$$X_t = \varphi_{21} Y_{t-1} + \varphi_{22} X_{t-1} + \varepsilon_{2t}.$$

A perturbation in ε_{1t} has an immediate and one-for-one effect on Y_t, but no effect on X_t. In period $t+1$, that perturbation in Y_t affects Y_{t+1} through the first equation and also affects X_{t+1} through the second equation. These effects work through to period $t + 2$, and so on. Thus a perturbation in one innovation in the VAR sets up a chain reaction over time in all variables in the VAR. Impulse response functions calculate these chain reactions.

7.5 example. Suppose we have a first-order VAR system defined by

$$\Theta_1 = \begin{pmatrix} 0.4 & 0.1 \\ 0.2 & 0.5 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 16 & 14 \\ 14 & 25 \end{pmatrix},$$

where Θ_1 is the coefficient matrix as in (7.4) and $\Omega = E\vec{\varepsilon}_t \vec{\varepsilon}_t'$ is the covariance matrix of the shock vector $\vec{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t})$. It is possible to show that respective VAR process is stationary (there is little point in studying impulse response functions for nonstationary systems.) Set $(Y_0, X_0) = (0,0)$ [7] and postulate $\varepsilon_1' = (4,0)'$. This vector sets a one-standard-deviation innovation in the first equation and a zero innovation in the second equation in period one. Assume further that both innovations are zero in periods 2, 3, and so on. The first few (Y, X) vectors are then given by

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} Y_2 \\ X_2 \end{pmatrix} = \Theta_1 \begin{pmatrix} Y_1 \\ X_1 \end{pmatrix} + \varepsilon_2' = \begin{pmatrix} 0.4 & 0.1 \\ 0.2 & 0.5 \end{pmatrix} \begin{pmatrix} 4 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1.6 \\ 0.8 \end{pmatrix}, \quad \begin{pmatrix} Y_3 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0.72 \\ 0.72 \end{pmatrix}$$

The impulse responses in the first five periods for a perturbation of one standard deviation in ε_{1t} are given in Table 7.2. Similarly some impulse responses for a perturbation of one standard deviation in ε_{2t} are presented in Table 7.3.

[7] That is, the process (Y_t, X_t) is in its equilibrium. What initial condition would you take if the free term $\vec{\alpha} \neq 0$?

Table 7.2

Impulse responses from $\varepsilon_1 = (4,0)'$

Period	Y	X
1	4	0
2	1.6	0.8
3	0.72	0.72
4	0.36	0.504
5	0.194	0.324

Table 7.3

Impulse responses from $\varepsilon_1 = (0,5)'$

Period	Y	X
1	0	5
2	0.5	2.5
3	0.45	1.35
4	0.315	0.765

An objection to the procedure just illustrated for the computation of impulse response functions is that the innovations in the VAR are, in general, not contemporaneously independent of one another. That one innovation receives a perturbation and the other does not is implausible. A widely used "solution" to the problem is to transform the $\vec{\varepsilon}$ innovations to produce a new set of uncorrelated unit variance innovations $\vec{u}$. We will illustrate the procedure for the two-variable case. Let $E(u_1, u_2) = (0,0)$ and $\text{cov}(u_1, u_2) = I$. We want to find four (in fact, three) coefficients $P = (p_{ij})$ such that the covariance matrix of $(\varepsilon_1, \varepsilon_2) = P\vec{u} = \begin{pmatrix} p_{11} & 0 \\ p_{21} & p_{22} \end{pmatrix} \vec{u}$:

$$\begin{aligned} \varepsilon_1 &= p_{11}u_1 + p_{12}u_2 = p_{11}u_1 + 0 \cdot u_2 = p_{11}u_1 \\ \varepsilon_2 &= p_{21}u_1 + p_{22}u_2 \end{aligned} \tag{7.6}$$

be the same as the sample covariance matrix of $(\varepsilon_1, \varepsilon_2)$, i.e., $\hat{\Omega}$. In order to do this, we have to solve the following three equation system:

$$\begin{aligned} D\varepsilon_1 &= p_{11}^2 = s_1^2 \\ D\varepsilon_2 &= p_{21}^2 + p_{22}^2 = s_2^2 \\ \text{cov}(\varepsilon_1, \varepsilon_2) &= Ep_{11}u_1 \cdot (p_{21}u_1 + p_{22}u_2) = p_{11}p_{21} = s_{12}; \end{aligned}$$

its solution is $p_{11} = s_1$, $p_{21} = s_{12} / s_1$, $p_{22} = \sqrt{s_2^2 - (s_{12} / s_1)^2}$. Note that in higher-dimensional VAR's, the equation that is first in the ordering has only one uncorrelated innovation, u_1. The equation that is second has only u_1 and u_2, the equation that is third has only u_1, u_2, and u_3, and so on. What is really important here are the equations relating $\vec{\varepsilon}$ and $\vec{u}$: $\vec{u}_t = P^{-1}\varepsilon_t$ and

$$\vec{\varepsilon}_t = Pu_t . \tag{7.7}$$

These relations imply the **Cholesky factorization** of the matrix $\hat{\Omega}$: $\hat{\Omega} = PP'$ (the matrix is expressed as the product of a lower triangular matrix P and its transpose P', which is upper triangular.) This imposes an ordering of the variables in the VAR and attributes all of the effect of any common component to the variable that comes first in the VAR system (in our case it is *Y*). Note that responses can change dramatically if you change the ordering of the variab-

les. In Gretl, you may reorder the variables by changing the order of entering VAR variables in gretl:VAR window.

7.5 example (continued). Continuing the previous numerical example, we will suppose that the values in the A and Ω matrices have been estimated from sample data. It is easy to check that

$$P = \begin{pmatrix} 4 & 0 \\ 3.5 & 3.5707 \end{pmatrix}.$$

Suppose that we postulate a $u_1 = (1,0)'$ vector and set all subsequent u vectors to zero. This vector gives a one standard deviation perturbation in the first component. From (7.7) this implies

$$\vec{\varepsilon}_1 = P\vec{u}_1 = \begin{pmatrix} 4 & 0 \\ 3.5 & 3.5707 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 3.5 \end{pmatrix}.$$

The second element in $\vec{\varepsilon}_1$ is now nonzero. It is, in fact, the expected value of ε_{21}, given that $\varepsilon_{11} = 4$. The values of the (Y, X) vector may then be calculated as before. The first few values are presented in Table 7.4. Compared with the earlier assumption of a one standard deviation perturbation in just ε_{11}, there is now an important impact on X in the first period, followed by noticeably greater impacts in subsequent periods. If a perturbation of one standard deviation in the second innovation is assumed, the ε_1 vector is given by

Table 7.4

Period	Y	X
1	4	3.5
2	1.95	2.55
3	1.035	1.665
4	0.580	1.039

$$\varepsilon_1 = Pu_1 = \begin{pmatrix} 4 & 0 \\ 3.5 & 3.5707 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 3.5707 \end{pmatrix}$$

and the successive (Y, X) vectors may be computed in the usual way.

Uncorrelated innovations were developed to deal with the problem of nonzero correlations between the original innovations. However, the solution of one problem creates another. The new problem is that the order in which the $\vec{u}$ variables are entered can have dramatic effects on the numerical results. The interpretation of impulse response functions is thus a somewhat hazardous operation, and there has been intense debate on their possible economic significance.

7.6 example. We estimate a bivariate VAR for U.S. seasonally adjusted housing `starts` and `completions`, two widely watched business cycle indicators for 1968.01-1996.06 (these monthly data are available as house.dat). We use the VAR to produce point extrapolation forecasts. We show housing starts and completions in Figure 7.2. Both are highly cyclical, increasing during business-cycle expansions and decreasing during contractions. Moreover, completions tend to lag behind starts, which makes sense because a house takes time to complete.

We split the data into an estimation sample, 1968:01-1991:12, and a holdout sample, 1992:01-1996:06 for forecasting. We therefore perform all model specification analysis and estimation, to which we now turn, on the 1968:01-1991:12 data.

To model starts and completions, we need to select the order, p, of our VAR(p). Exploration of VARs of order 1 through 36 revealed that the AIC achieves a distict minimum at $p = 4$, so we adopt VAR(4).

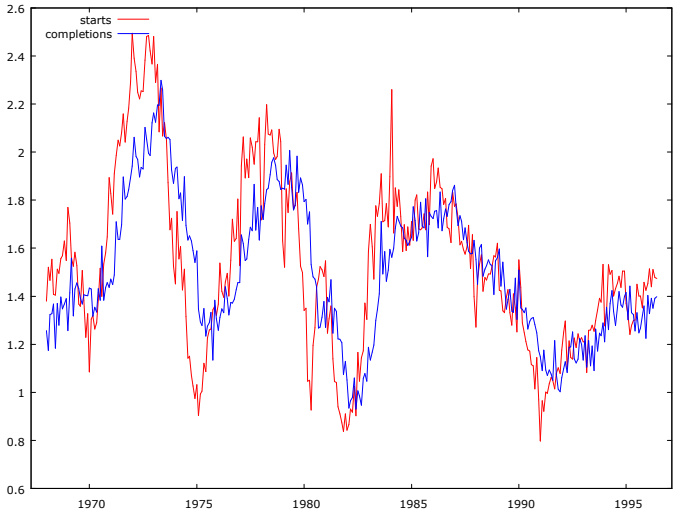


Figure 7.6. US housing starts and completions, 1968:1 – 1996:06

```
smpl 1968:01 1991:12
var 36 starts completions --lagselect
```

```
lags         loglik     p(LR)        AIC           BIC           HQC

   1     425,18118              -3,326835     -3,242801     -3,293021
   2     466,27329   0,00000    -3,621217     -3,481160*    -3,564861
   3     477,23394   0,00021    -3,676460     -3,480380     -3,597562*
   4     483,68704   0,01174    -3,695929*    -3,443827     -3,594488
   5     484,47037   0,81477    -3,670400     -3,362275     -3,546417
   6     487,36627   0,21524    -3,661637     -3,297490     -3,515112
   7     488,02559   0,85821    -3,635124     -3,214954     -3,466056
   8     490,08218   0,39091    -3,619700     -3,143507     -3,428090
   ...............................................................................
```

After we have chosen VAR(4), go to Model|Time series|Vector Autoregression… and then Analysis|Forecasts. The resulting forecasts are presented in Fig. 7.7. Starts begin their recov-

ery before 1992:01, and the VAR projects continuation of the recovery. The VAR forecast captures the general pattern quite well, but it forecasts quicker mean reversion than actually occurs, as is clear when comparing the forecast and realization in Fig. 7.7. This figure also makes clear that the recovery of housing starts from the recession of 1990 was slower than the previous recoveries in the sample, which naturally makes for difficult forecasting. The completions forecast suffers the same fate, as shown in Fig. 7.7, right. Interestingly, however, completions had not yet turned by 1991:12, but the forecast nevertheless correctly predicts the turning point.

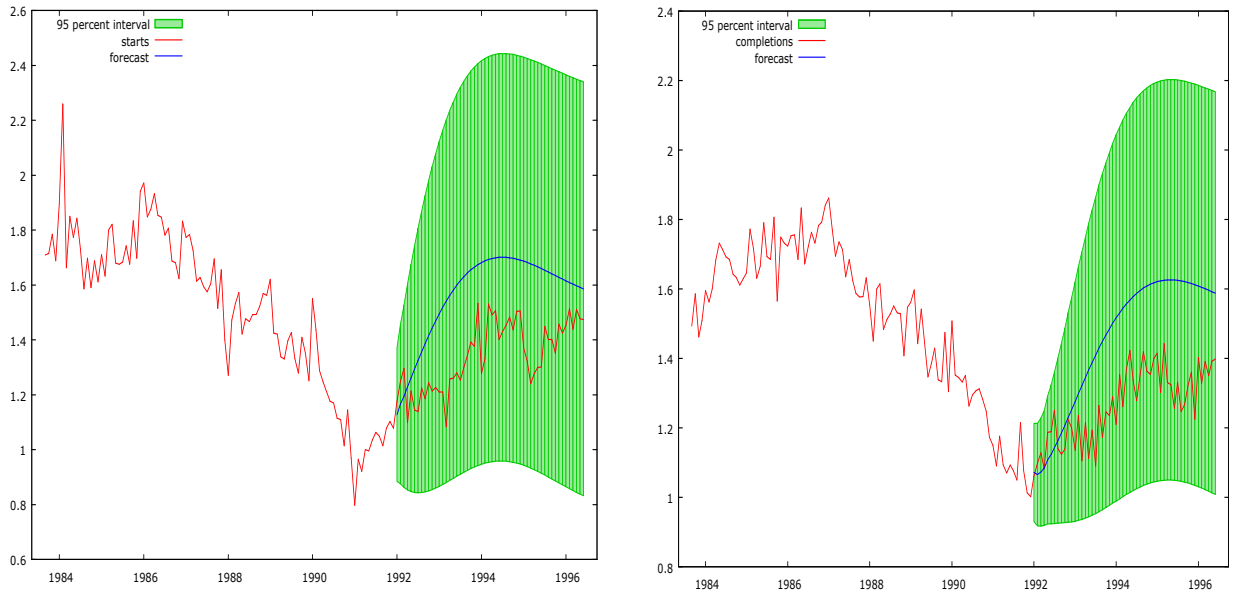


Figure 7.7. Housing starts: history 1968:01-1991:12 and forecast and realization 1992:01-1996:06 (left) and the same for completions (right)

In Fig. 7.8 we display the impulse-response functions. First we consider the own-variable impulse responses – that is, the effects of the starts innovation on subsequent starts (top, left) or a completions innovation on subsequent completions (bottom, right). The effects are similar: in each case, the impulse response is large and decays in a slow, approximately monotonic fashion. In contrast,

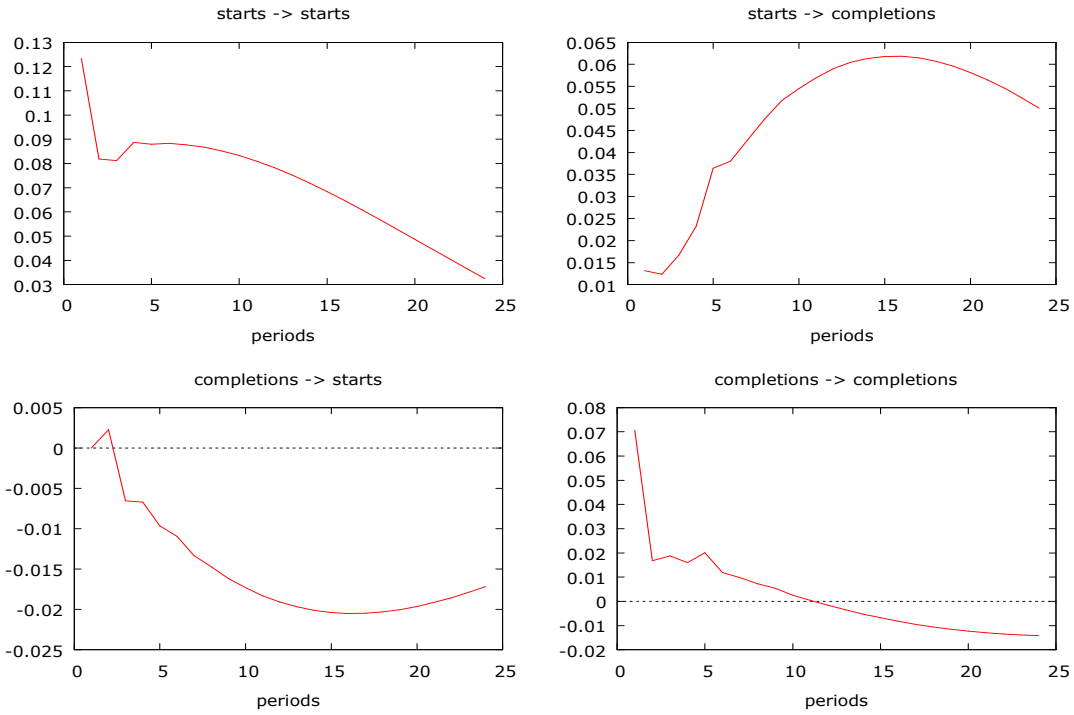


Figure 7.8. Responses to one standard deviation innovations

the cross-variable impulse responses are very different. An innovation to starts produces no movement in completions at first, but the effect gradually builds and becomes large, peaking at about 14 months (it takes time to build houses). An innovation to completions, however, produces little movement in starts at any time.

Note that in our case reversing the order of starts and completions in var box does not considerably change Fig. 7.8.

7.4. Vector Error Correction Model (VECM)

Recall that the AR(p) process

$$Y_t = \mu + \delta t + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \varepsilon_t$$

(see (5.1)) can be rearranged into

$$\Delta Y_t = \mu + \delta t + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t ;$$

the latter form is more convenient for the unit root testing[8]: $\rho = 0$ implies that Y_t has a unit root. Similarly, the d - dimensional VAR process

$$\vec{Y}_t = \vec{\mu}_t + \Theta_1 \vec{Y}_{t-1} + \ldots + \Theta_p \vec{Y}_{t-p} + \vec{\varepsilon}_t$$

[8] Recall that have used the ADF test to this end.

(here the deterministic part $\vec{\mu}_t$ is typically a polynomial in time, i.e., $\vec{\mu}_t = \vec{\mu}$ or $\vec{\mu}_t = \vec{\mu} + \vec{\delta} t$, or $\vec{\mu}_t = \vec{\mu} + \vec{\delta} t + \vec{\gamma} t^2$) can be rewritten in VEC form (c.f. (6.5c)):

$$\Delta \vec{Y}_t = \vec{\mu}_t + \Pi \vec{Y}_{t-1} + \Gamma_1 \Delta \vec{Y}_{t-1} + ... + \Gamma_{p-1} \Delta \vec{Y}_{t-p+1} + \vec{\varepsilon}_t, \qquad (7.6a)$$

where the long-run matrix $\Pi = \sum_{i=1}^{p} \Theta_i - I$ and $\Gamma_i = -\sum_{j=i+1}^{p} \Theta_j$ or, alternatively, as

$$\Delta \vec{Y}_t = \vec{\mu}_t + \tilde{\Gamma}_1 \Delta \vec{Y}_{t-1} + ... + \tilde{\Gamma}_{p-1} \Delta \vec{Y}_{t-p+1} + \tilde{\Pi} \vec{Y}_{t-p} + \vec{\varepsilon}_t .$$

Below we shall use Johansen test to test whether Π equals 0 in some sense (in $\det \Pi = 0$ or $rank\ \Pi = r < d$ sense; more precisely, the test is aimed to test the number r of cointegrating relationships). Thus the Johansen approach can be interpreted as a multivariate unit root test. Note the central fact: if Π has rank r (in other words, if $(Y_{t,1},...,Y_{t,d})$ has r linearly independent cointegrating relations), it can be written as the product $\Pi_{d\times d} = \alpha_{d\times r}\, \beta^T_{r\times d}$, $rank\, \alpha = rank\, \beta = r$. The rows of the matrix β^T form a basis for the r cointegrating vectors and the elements of α distribute the impact of the cointegrating vectors to the evolution of $\Delta \vec{Y}_t$ (they are usually interpreted as speed of adjustment to equilibrium coefficients).

In the preceding discussion of VARs, we assumed that all variables were stationary. If all of the original variables have unit roots and are **not** cointegrated, then they should be differenced and the resulting stationary variables should be used in the VAR. This covers every case except one where the variables have unit roots and are cointegrated. Recall that in this case in the discussion of Granger causality, we recommended that you work with an ECM. The same strategy can be employed here. In particular, instead of working with a vector autoregression (VAR), you should work with a vector error correction model (VECM).

To outline the strategy of dealing with multivariate time series, it is better to start with a two-dimensional case $\vec{Y}_t = (Y_t, X_t)'$ and recall the Engle-Granger (EG) procedure:

1. Test whether each series, Y_t and X_t, is integrated of the same order.
2. If both series are $I(0)$, estimate VAR model in levels (no need for VECM).
3. If both series are $I(1)$, estimate the *cointegration regression* $Y_t = \gamma_0 + \gamma_1 X_t + Z_t$, then test whether the residuals $\hat{Z}_t$ are stationary (this is called the Engle-Granger (EG) test, it is close to the ADF test).
4. If Z_t is $I(1)$, estimate a VAR model in differences ΔY_t and ΔX_t.
5. If Z_t is stationary, Y_t and X_t are *cointegrated*; in this case, estimate the VEC model

$$\begin{cases} \Delta Y_t = \alpha_2 \hat{Z}_{t-1} + \mu_2 + \gamma^1_{21} \Delta X_{t-1} + \gamma^1_{22} \Delta Y_{t-1} + ... + \gamma^p_{21} \Delta X_{t-p} + \gamma^p_{22} \Delta Y_{t-p} + \varepsilon^y_t \\ \Delta X_t = \alpha_1 \hat{Z}_{t-1} + \mu_1 + \gamma^1_{11} \Delta X_{t-1} + \gamma^1_{12} \Delta Y_{t-1} + ... + \gamma^p_{11} \Delta X_{t-p} + \gamma^p_{12} \Delta Y_{t-p} + \varepsilon^x_t \end{cases} \qquad (7.7a)$$

or, to express it in a matrix form,

$$\begin{pmatrix}\Delta Y\\ \Delta X\end{pmatrix}_t = \begin{pmatrix}-\alpha_L\\ \alpha_S\end{pmatrix}\begin{pmatrix}1 & -\gamma_1\end{pmatrix}\begin{pmatrix}Y\\ X\end{pmatrix}_{t-1} + \Gamma_1\begin{pmatrix}\Delta Y\\ \Delta X\end{pmatrix}_{t-2} + ... + \Gamma_{p-1}\begin{pmatrix}\Delta Y\\ \Delta X\end{pmatrix}_{t-p} + \begin{pmatrix}\varepsilon_Y\\ \varepsilon_X\end{pmatrix}_t \qquad (7.7b)$$

The order p of this VEC model is chosen such that VAR($p+1$) model fitted to the levels has minimum AIC or SC. If $p=0$, i.e., a level VAR(1) has minimum, this may indicate that that the original series are stationary.

6. If necessary, use the model obtained to forecast Y_t and X_t (this can be done by rewriting the VEC model as a VAR model). For example, the model $\Delta\vec{Y}_t = \alpha\beta^T\vec{Y}_{t-1} + \Gamma_1\Delta\vec{Y}_{t-1} + \vec{\varepsilon}_t$ can be expressed as VAR(2): $\vec{Y}_t = (I + \Gamma_1 + \alpha\beta^T)\vec{Y}_{t-1} - \Gamma_1\vec{Y}_{t-2} + \vec{\varepsilon}_t$. ◀◀

Thus, in the case where our data consists of **two** $I(1)$ components, we use the EG test for cointegration (see 6.4 and 6.5; the test was to check whether H_0 : *no cointegration* or H_1 : *Y and X cointegrate* is true). In a **multidimensional case** we use another cointegration test called the **Johansen test**. The first thing to note is that it is possible for more than one cointegrating relationship to exist if you are working with several time series variables (all of which you have tested and found to have unit roots). To be precise, if you are working with d variables, then it is possible to have up to $d-1$ cointegrating relationships (and, thus, up to $d-1$ cointegrating residuals included in the VECM). For instance, in Ch. 6 we mentioned a financial theory arguing that the *cay* variables (i.e., (logarithms of) consumption, assets and income) are cointegrated. As we shall see below, there probably is just one cointegrating relationship between these variables. That is, all *c*, *a* and *y* have unit roots, but $c_t - \beta_1 a_t - \beta_2 y_t$ is stationary. However, in theory it would be possible to have two cointegrating relationships (e.g., if $c_t - 1\cdot y_t$ were stationary[9]; it is evident that the VEC models depends on the number of cointegrating relationships).

To begin with, let us return to (7.6a):

$$\Delta\vec{Y}_t = \vec{\mu}_t + \alpha\beta^T\vec{Y}_{t-1} + \Gamma_1\Delta\vec{Y}_{t-1} + ... + \Gamma_{p-1}\Delta\vec{Y}_{t-p+1} + \vec{\varepsilon}_t, \qquad (7.6b)$$

- If $rank\,\beta = 0$, then only $\beta^T\vec{Y}_{t-1} = 0\cdot Y_{1,t-1} + ... + 0\cdot Y_{d,t-1} \equiv 0$ is stationary, in other words, $\vec{Y}_t$ is not cointegrated and VECM reduces to VAR(p-1) in differences.
- If $0 < rank\,\beta = r < d$, $\vec{Y}_t$ is $I(1)$ with r linearly independent cointegrating vectors and $\beta^T\vec{Y}_{t-1} \sim I(0)$
- If $rank\,\beta = d$, then Π has full rank and is invertible, therefore $\vec{Y}_{t-1}$ be will a linear combination of stationary differences, therefore stationary itself.

[9] The ratio C_t / Y_t should not deviate much from a constant, therefore it is probable that $\log(C_t / Y_t) = \log C_t - \log Y_t = c_t - y_t$ is close to (another) constant, i.e., stationary.

Thus, it is often of interest to **test**, not simply for whether cointegrating is present or not, but also **for the number of cointegrating relationships**. Recall that any hypothesis is rejected if the test statistics exceeds the critical value. However, these critical values depend on the deterministic components of VECM such as constants and linear trends. In a similar situation when testing for a unit root, i.e., the hypothesis $H_0 : \rho = 0$ (see Table 5.2), we used two different critical values for the t-statistics of $\hat{\rho}$ depending on the presence or absence of a deterministic trend. Now we have five different variants for calculating $p-$values. The above-mentioned five cases are:

1. $\vec{\mu}_t = \vec{0}$ (**no constant**) – all the series in $\vec{Y}_t$ are I(1) without drift and the cointegrating relations $\beta^T \vec{Y}_t$ $(= \beta_1 Y_t^{(1)} + ... + \beta_M Y_t^{(M)})$ have zero mean:

$$\Delta \vec{Y}_t = \alpha \beta^T \vec{Y}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta \vec{Y}_{t-i} + \vec{\varepsilon}_t \, .$$

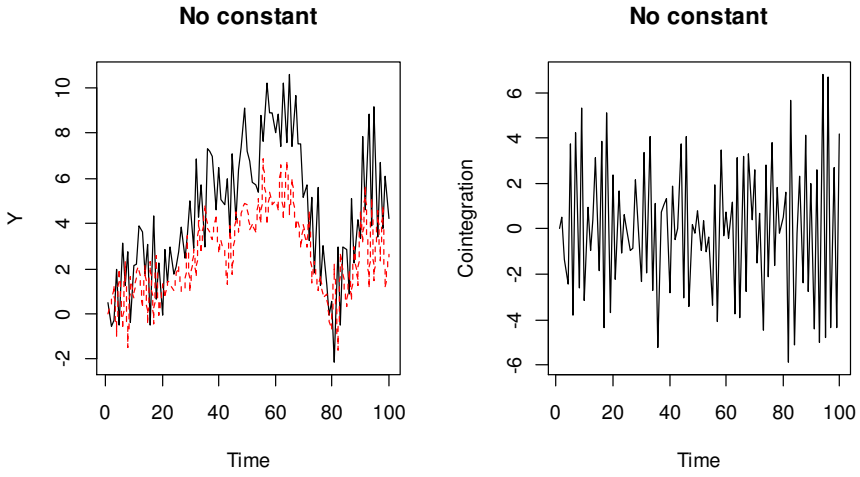


Figure 7.9. (Bivariate VECM) Case 1, simulated $\vec{Y}_t$ (left, no components have a drift); cointegrating term $Y_{1t} - \beta Y_{2t}$ (right)

2. $\vec{\mu}_t = \vec{\mu}_0 = \alpha \vec{\rho}_0$ (**restricted constant**) – the series in $\vec{Y}_t$ are I(1) without drift and the cointegrating relations $\beta^T \vec{Y}_t$ have a non-zero mean:

$$\Delta \vec{Y}_t = \alpha(\beta^T \vec{Y}_{t-1} + \vec{\rho}_0) + \sum_{i=1}^{p-1} \Gamma_i \Delta \vec{Y}_{t-i} + \vec{\varepsilon}_t \, .$$

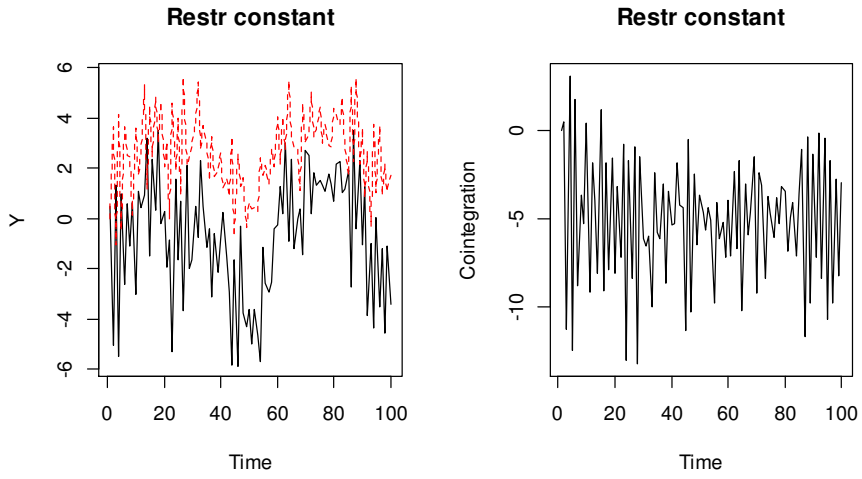


Figure 7.10. Case 2 - no components have a drift

3. $\vec{\mu}_t = \vec{\mu}_0$ (**unrestricted constant**) – the series in $\vec{Y}_t$ are I(1) with drift vector $\vec{\mu}_0$ and the cointegrating relations $\beta^T \vec{Y}_t$ may have a non-zero mean:

$$\Delta \vec{Y}_t = \vec{\mu}_0 + \alpha \beta^T \vec{Y}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta \vec{Y}_{t-i} + \vec{\varepsilon}_t$$

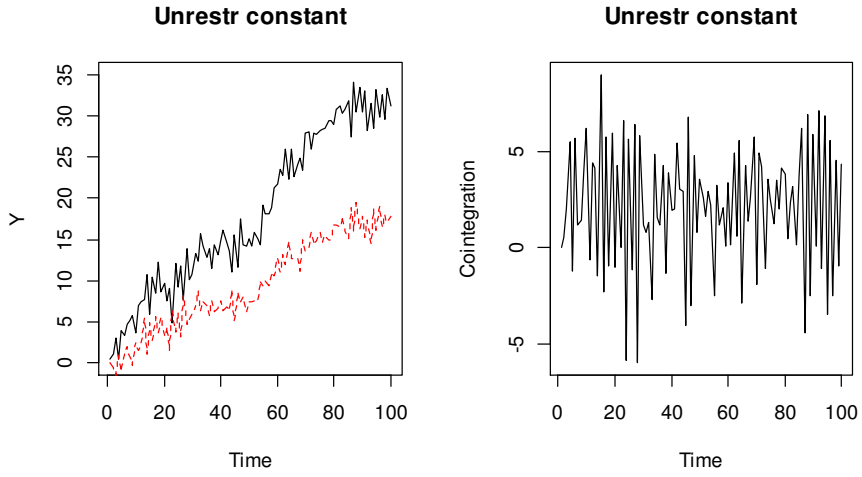


Figure 7.11. Case 3 - at least one component drifts

4. $\vec{\mu}_t = \vec{\mu}_0 + \alpha \vec{\rho}_1 t$ (**restricted trend**) – the series in $\vec{Y}_t$ are I(1) with drift vector $\vec{\mu}_0$ and the cointegrating relations $\beta^T \vec{Y}_t$ may have a linear trend $\vec{\rho}_1 t$:

$$\Delta \vec{Y}_t = \vec{\mu}_0 + \alpha(\beta^T \vec{Y}_{t-1} + \vec{\rho}_1 t) + \sum_{i=1}^{p-1} \Gamma_i \Delta \vec{Y}_{t-i} + \vec{\varepsilon}_t$$

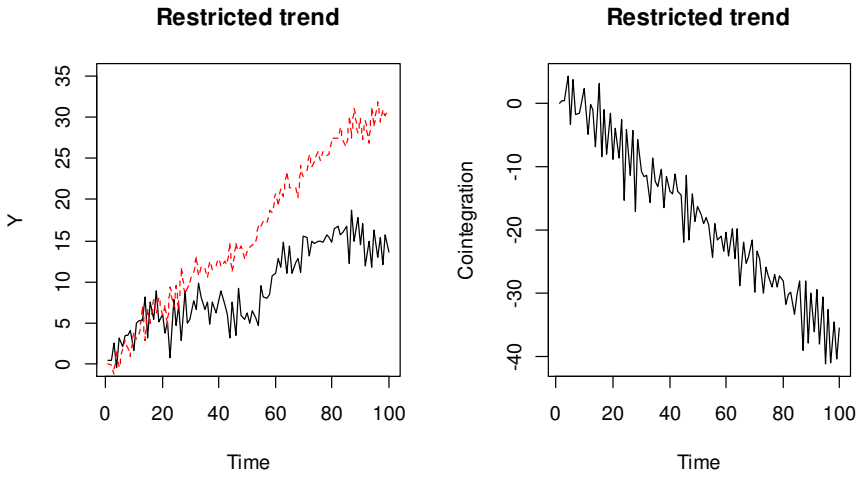


Figure 7.12. Case 4 - at least one component drifts

5. $\vec{\mu}_t = \vec{\mu}_0 + \vec{\mu}_1 t$ (**unrestricted constant and trend**) – the series in $\vec{Y}_t$ are I(1) with a linear trend in VECM (and quadratic trend in levels) and the cointegrating relations $\beta^T \vec{Y}_t$ have a linear trend.

$$\Delta \vec{Y}_t = \vec{\mu}_0 + \vec{\mu}_1 t + \alpha \beta^T \vec{Y}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta \vec{Y}_{t-i} + \vec{\varepsilon}_t$$

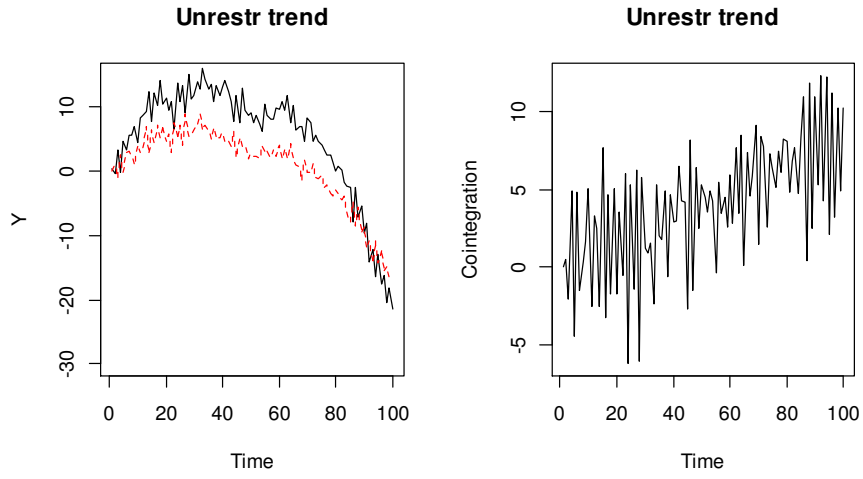


Figure 7.13. Case 5 - at least one component has a quadratic trend

Note – if no components of $\vec{Y}_t$ drift, we use Cases 1 or 2; if at least one components of $\vec{Y}_t$ drifts, we use Cases 3 or 4; and if at least one components of $\vec{Y}_t$ has a quadratic trend, we use Case 5.

Simulated data for the five cases for a bivariate cointegrated VAR(1) model (available, respectively, in vecm1.txt,..., vecm5.txt files) are illustrated in Figures 7.9, ... 7.13. Case 1 is not really relevant for empirical work. The restricted contstant Case 2 is appropriate for non-trending I(1) data like interest rates and exchange rates. The unrestriced constant Case 3 is appropriate for trending I(1) data like asset prices, macroeconomic aggregates (real GDP, consumption, employment etc). The restricted trend case 4 is also appropriate for trending I(1) as in Case 3. However, notice the deterministic trend in the cointegrating residual in Case 4 as opposed to the stationary residuals in case 3. Finally, the unrestricted trend Case 5 is appropriate for I(1) data with a quadratic trend. An example might be nominal price data during times of extreme inflation.

The above-given figures and considerations are important in choosing the right variant to define critical values of the Johansen test. The basic steps in Johansen's methodology are (we assume that all the series in $\vec{Y}_t$ are I(1)):

1. Choose the right order p for a VAR(p) model for levels
2. Choose the right case out of the five ones (use graphs of $\vec{Y}_t$)
3. Apply Johansen's test and find the number of cointegrating relations
4. Create VECM
5. Use it to forecast $\vec{Y}_t$

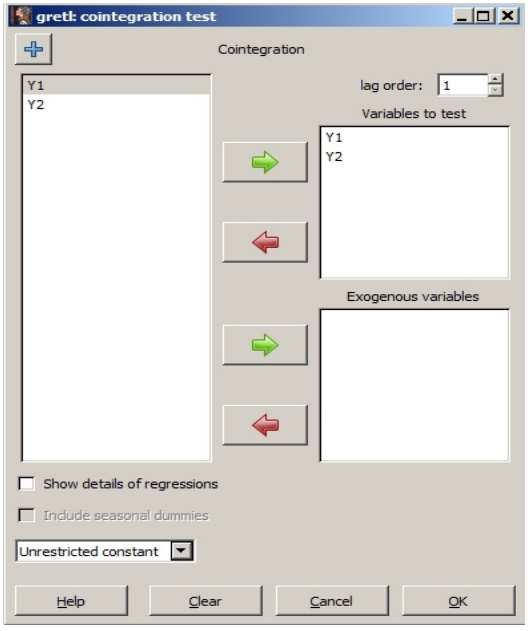
7.7 example. Import vecm3.txt as Other time series starting at 1 and, following the above recommendations, create its VECM model.

0. It is easy to show (through Variable| Unit root tests| Augmented Dickey-Fuller test) that both Y1 and Y2 have unit roots.

1. Using Model| Time series| VAR lag selection..., we obtain that (Y1,Y2) is described as VAR(1).

2. As both Y1 and Y2 are drifting, we should try both Cases 3 or Case 4 and choose the "better" one.

3. To find the number of cointegrating relationships[10] (in our two dimensional case, it can be zero or one), we shall apply Johansen's test: go to Model| Time series| Cointegration test| Johansen... and fill the box as shown on the right (thus, we begin with Case 3).



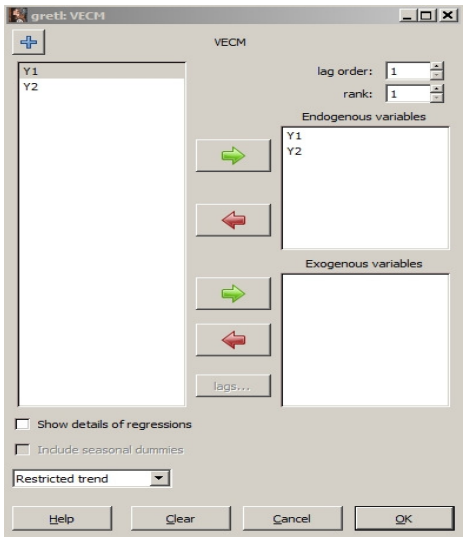
The output of the test (see below) alows us to determine the rank r of Π. The estimated eigenvalues λ_i are sorted from largest to smallest and we apply a sequential procedure: both

[10] Or the rank of the matrix Π (see (7.6)) or the number of non-zero eigenvalues of this matrix.

trace and Lmax tests rejects the null hypothesis $H_0: r = 0$, but accepts the next null $H_0: r = 1$ (we derive the same conclusion, i.e., $r = 1$, if the option Unrestricted constant is replaced by Restricted trend).

```
Case 3: Unrestricted constant

Rank Eigenvalue Trace test  p-value  Lmax test  p-value
   0    0.89453     223.86 [0.0000]     222.69 [0.0000]
   1   0.011804     1.1755 [0.2783]     1.1755 [0.2783]
```

4. Which one of the two competing models to choose? Go to Model| Time series| VECM... and fill the box as shown on the right. The output reads AIC = 6.0256, which is slightly less than the corresponding AIC in the Unrestricted constant case. Thus, this Restricted trend case is our final VEC model and its output can be written more explicitely as given below (note that we misclassify the Case 3 process as Case 4).




```
VECM system, lag order 1
Cointegration rank = 1
Case 4: Restricted trend, unrestricted constant

beta (cointegrating vectors, standard errors in parentheses)

Y1          1.0000
           (0.00000)
Y2          -1.8278
           (0.078852)
trend       0.026320
           (0.015617)

alpha (adjustment vectors)

Y1         -0.76565
Y2          0.53356

AIC = 6.0256
BIC = 6.1829
HQC = 6.0893

Equation 1: d_Y1

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        2.02582     0.128566     15.76    2.27e-028 ***
  EC1         -0.765654    0.0320605   -23.88    3.59e-042 ***
```

```
Equation 2: d_Y2

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -1.01562     0.132060     -7.691    1.28e-011 ***
  EC1          0.533560    0.0329319    16.20     3.23e-029 ***
```

Below is the same output given in a system of equations form:

$$\begin{cases} \Delta Y1_t &= 2.03 - 0.77 \cdot (Y1_{t-1} - 1.83 Y2_{t-1} + 0.03(t-1)) \\ \Delta Y2_t &= -1.02 + 0.53 \cdot (Y1_{t-1} - 1.83 Y2_{t-1} + 0.03(t-1)) \end{cases}$$

(the true, i.e., simulated) model was

$$\begin{cases} \Delta Y1_t &= 1.9 - 0.8 \cdot (Y1_{t-1} - 1.7 Y2_{t-1}) + \varepsilon_{1t} \\ \Delta Y2_t &= -1 \quad + 0.6 \cdot (Y1_{t-1} - 1.7 Y2_{t-1}) + \varepsilon_{2t} \end{cases}$$

thus the estimate is close to it).

The coefficient $(\alpha_1 =)$ 0.53 is called an adjustment coefficient and indicates that $Y2$ will return to equilibrium in $(1/0.53 \approx)$ 2 steps, *ceteris paribus*.

5. Now we can use the model obtained to forecast $\vec{Y}_t$: in VECM window go to Analysis| Forecast and choose 10-years-ahead forecast for, first, YY and then XX.

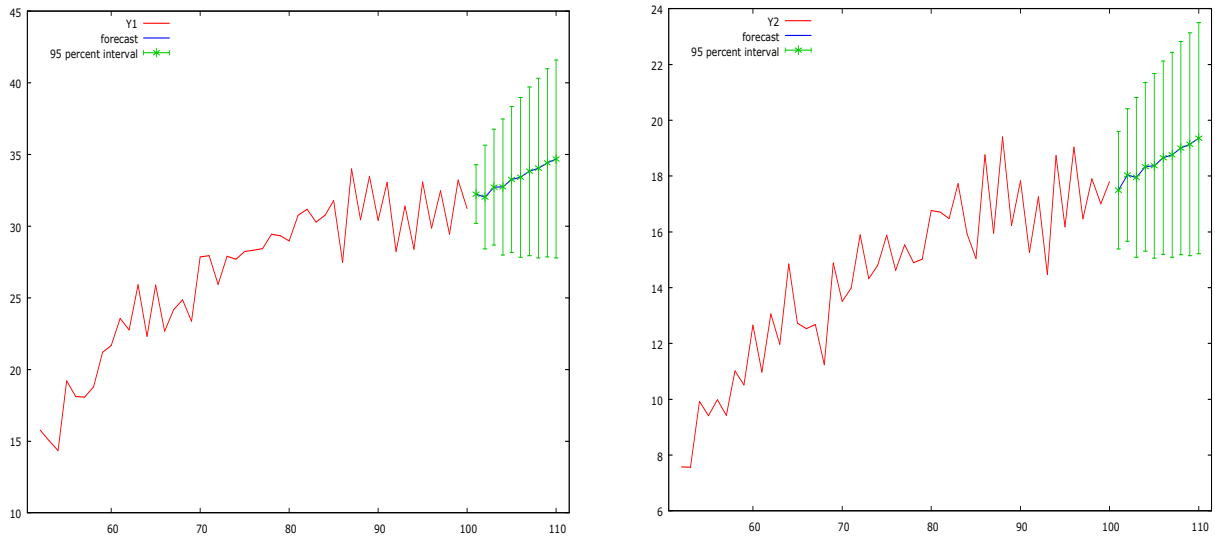


Figure 7.14. VECM 10 years forecast for $Y1$ and $Y2$.

The reader can verify that the assumption that the right model is Case3 practically does not change the forecast. ◀◀

A few final remarks:

1. If we have only two variables, both EG and Johansen tests are appropriate for testing cointegration. One can use the EG test in multidimensional case as well[11], but if there exist more than one cointegrating relation, EG method will suffer from the omitted variable bias. Thus, in multidimensional case use Johansen test. Note that the VEC model can also be created in the case where some components of the $\vec{Y}_t$ are $I(1)$ while others are $I(0)$. In this case, the situation is rather complicated, we send the interested reader to [H, p.652].

2. The most restricted model (Case 1) is unlikely to find general use because, at least, a constant will usually be included in the cointegration equation. The least restricted model (Case 5) allows for quadratic trends in the data which occurs quite rarely. The choice between Case 2 and Case 3 rests upon whether there is a need to allow for the possibility of linear trends in the data, a preliminary graphing of the data is often helpful in this respect. If Case 3 is preferred to Case 2, only then does Case 4 need to be considered since the data has to have a linear trend if we are to consider allowing a trend in the cointegration equation. As a rough guide, use case 2 if none of the series appear to have a trend. For trending series, use case 3 (default in gretl) if you believe all trends are stochastic; if you believe some of the series are trend stationary, use case 4. However, the simplest way that also allows one to „automate" selection is to choose the case according to the minimum of AIC of BIC of the model.

3. Multivariate VAR is a complicated model and the previous material was just an introduction to the relevant topics. Below we present some useful facts without proofs (see [L]).

- To estimate the coefficients of $\vec{Y}_t = \vec{\alpha} + \vec{\delta} t + \Theta_1 \vec{Y}_{t-1} + ... + \Theta_p \vec{Y}_{t-p} + \vec{\varepsilon}_t$ use (conditional) maximum likelihood method assuming that innovations $\vec{\varepsilon}_t$ have a multivariate normal distribution. This is equivallent to the least squares method applied to each equation separately.
- Maximum likelihood estimates are consistent even if the true innovations are non-Gaussian.
- Standard OLS t and F statistics applied to the coefficients of any single equation of the VAR are asymptotically valid.
- The goal of unit root tests is to find a parsimonious representation of the data that gives a reasonable approximation of the true process, as opposed to determining whether or not the true process is literally $I(1)$.
- If $\vec{Y}_t$ is cointegrated, a VAR estimated in levels is not misspecified but involves a loss of efficiency.
- Let $\vec{Y}_t$ have a unit root, but no cointegration. A VAR in levels is not subject to the spurious regression problem discussed above for single equation regressions.
- Even if there is no cointegration among the variables in $\vec{Y}_t$, equation-by-equation OLS estimation of VAR in levels delivers consistent estimates of the VAR parameters. Un-

[11] To test if $\alpha_1 Y_{1,t} + \alpha_2 Y_{2,t} + ... + \alpha_M Y_{M,t}$ is stationary for some $\alpha_1, ..., \alpha_M$.

like a univariate regression, differencing is not required to obtain consistent estimates. Nevertheless, the small sample properties of the estimator may be improved by estimating the VAR in differences.

- Suppose that some of the M variables are stationary while the other variables are each individually $I(1)$ and also cointegrated by, say, a single cointegration relation. In [H, p.651], one can find an explanation of how to construct a VEC model in this case. For example, if $M=4$, $Y_t^{(1)}$ is stationary and $\vec{Y}_t^{(2)}=(Y_t^{(2)},Y_t^{(3)},Y_t^{(4)})$ are cointegrated $I(1)$, then the VEC representation will be of the form

$$\begin{pmatrix} Y_t^{(1)} \\ \Delta\vec{Y}_t^{(2)} \end{pmatrix}=\begin{pmatrix} \mu_1 \\ \vec{\mu}_2 \end{pmatrix}+\begin{pmatrix} \vec{\pi}_1 \\ \vec{\pi}_2 \end{pmatrix}\vec{Y}_{t-1}^{(2)}+\begin{pmatrix} \gamma_{11}^{(1)} & \vec{\gamma}_{12}^{(1)} \\ \gamma_{21}^{(1)} & \vec{\gamma}_{22}^{(1)} \end{pmatrix}\begin{pmatrix} Y_{t-1}^{(1)} \\ \Delta\vec{Y}_{t-1}^{(2)} \end{pmatrix}+\ldots+\begin{pmatrix} \varepsilon_t^{(1)} \\ \vec{\varepsilon}_t^{(2)} \end{pmatrix},$$

where the 4×3 matrix $\begin{pmatrix} \vec{\pi}_1 \\ \vec{\pi}_2 \end{pmatrix}$ is restricted to be of the form $\vec{b}\vec{a}'$ where $\vec{b}$ is 4×1 and $\vec{a}'$ is 1×3.

7.3 exercise. Create a VEC model for the data from the cay.txt file.

7.4 exercise. Create a VEC model for the data from us-tbill.txt file.

Revision questions

7.1 Let the three-dimensional time series $\vec{Y}_t=(Y_{1t},Y_{2t},Y_{3t})^T$ is stationary. How would you test the claim that Y_3 is a Granger-cause of Y_1?

7.2 Let all the components of the three-dimensional VAR(1) process $\vec{Y}_t=(Y_{1t},Y_{2t},Y_{3t})^T$ are I(1) bound by one cointegration relationship. Write a respective VEC model.

7.3 What is the Johansen test used for?

8. Endogenous Right-Hand-Side Variables

Consider a simple regression model $Y_t = \alpha + \beta X_t + u_t$. We know that under the classical Gauss-Markov conditions, the OLS estimators of its coefficients are BLUE (Best Linear Unbiased Estimator). However, if it happens that X_t is correlated with u_t, then the OLS estimators become biased, inconsistent and inefficient. This situation often arises when one or more of the explanatory variables is *jointly determined* with the dependent variable, typically through an equilibrium mechanism (this is called a simultaneous equations model.) The leading method for estimating simultaneous equations models is the method of instrumental variables (IV) and we start its exposition in a one equation case.

8.1. One Equation

Consider a simple model written as $Y = \alpha + \beta X + u$, where we think that X and u are correlated: $\text{cov}(X,u) \neq 0$ (thus, X is an *endogenous* variable). In order to obtain consistent estimators of α and β, suppose that we have an observable variable Z that satisfies two assumptions: (1) Z is uncorrelated with u, that is, $\text{cov}(Z,u) = 0$ (we say that Z is *exogenous* variable), but (2) Z is correlated with X, that is, $\text{cov}(Z,X) \neq 0$ (we call Z an *instrumental variable* for X, or sometimes simply an *instrument* for X.) Recall that under the classical assumptions the usual $\hat{\beta}_{LS} = \dfrac{\widehat{\text{cov}}(X,Y)}{\widehat{\text{var}}X}$ is the solution of the following two moments equations : $\begin{cases} \sum\left(Y_t - (\alpha + \beta X_t)\right) = 0 & \sim E\varepsilon = 0 \\ \sum\left(Y_t - (\alpha + \beta X_t)\right) X_t = 0 & \sim \text{cov}(\varepsilon, X) = 0 \end{cases}$.

Now, as the second equation fails, we replace X in it by the instrument Z and obtain the (consistent[1]) instrumental variable estimator $\hat{\beta}_{IV} = \dfrac{\widehat{\text{cov}}(Z,Y)}{\widehat{\text{cov}}(Z,X)}$ (prove the formula). Quite often the lag X_{t-1} serves as a good instrument to X $(= X_t)$; X_{t-1} will be a „good“ instrument if the correlation between X_{t-1} and X_t is sufficiently high or, what is almost the same, the coefficient δ_1 in the regression $X_t = \delta_0 + \delta_1 X_{t-1} + v_t$ is significant.

In what follows, we shall use Y to denote endogenous variables and Z exogenous. Thus our previous model can be rewritten as $Y_1 = \alpha + \beta Y_2 + u$; we also assume that we know Z. The model can be generalized to $Y_1 = \alpha + \beta_1 Y_2 + \beta_2 Z_1 + u$; in order to use the IV method, we need another exogenous variable, call it Z_2, that does not appear in our equation. The last model can still be generalized to $Y_1 = \alpha + \beta_1 Y_2 + \beta_2 Z_1 + ... + \beta_k Z_{k-1} + u$, where we again assume that we have an instrument to Y_2, say Z_k. Now, to get the the IV estimators of βs, one has to sol-

[1] The IV estimator is biased in finite samples even if asymptotically consistent (the bias can be especially pronounced when instruments are weak, i.e., when $r^2_{X,Z}$ is small).

ve the system

$$\begin{cases} \sum\left(Y_t - (\alpha + \beta_1 Y_{2t} + \beta_2 Z_{1t} + ... + \beta_k Z_{k-1,t})\right) = 0 \\ \sum\left(Y_t - (\alpha + \beta_1 Y_{2t} + \beta_2 Z_{1t} + ... + \beta_k Z_{k-1,t})\right) Z_{1t} = 0 \\ .... \\ \sum\left(Y_t - (\alpha + \beta_1 Y_{2t} + \beta_2 Z_{1t} + ... + \beta_k Z_{k-1,t})\right) Z_{kt} = 0 \end{cases}$$

What if we have <u>two</u> instruments for Y_2, Z_k and Z_{k+1}? We can get two IV estimators of β's , and neither of these would, in general, be efficient. To find the best IV, we choose the linear combination of *all* exogenous variables that is best correlated with Y_2. This turns out to be given by $\hat{Y}_2 = \hat{\pi}_0 + \hat{\pi}_1 Z_1 + ... + \hat{\pi}_{k+1} Z_{k+1}$, where the $\hat{\pi}$s are the OLS estimates in respective model. We can use $\hat{Y}_2$ as an instrument to Y_2 or, <u>alternatively</u>, apply the following *two stage least squares* (2SLS) procedure: 1. Obtain the above mentioned estimator $\hat{Y}_2$ and 2. Replace Y_2 with $\hat{Y}_2$ in $Y_1 = \alpha + \beta_1 Y_2 + \beta_2 Z_1 + ... + \beta_k Z_{k-1} + u$ and once again apply OLS (gretl uses the second option). Note that 2SLS can also be used in models with more than one endogenous explanatory variable. For example, consider the model $Y_1 = \alpha + \beta_1 Y_2 + \beta_2 Y_3 + \beta_3 Z_1 + \beta_4 Z_2 + \beta_5 Z_3 + u$. To estimate β's we need at least two more exogenous variables Z_4 and Z_5 that do not appear in this equation but that are correlated with Y_2 and Y_3. On the first stage, we apply OLS and estimate $\hat{Y}_2 = \hat{\pi}_0^{(2)} + \hat{\pi}_1^{(2)} Z_1 + ... + \hat{\pi}_5^{(2)} Z_5$ and $\hat{Y}_3 = \hat{\pi}_0^{(3)} + \hat{\pi}_1^{(3)} Z_1 + ... + \hat{\pi}_5^{(3)} Z_5$, and, on the second stage, replace Y_2 and Y_3 with, respectively, $\hat{Y}_2$ and $\hat{Y}_3$ and estimate α and β's with OLS.

8.1 example. In the 1960s, macroeconomists grew very interested in the relationship between unemployment and inflation. In 1958, A. W. Phillips had pointed out that British macroeconomic data showed a negative correlation between the two variables; unemployment tended to be low when inflation was high and vice versa. Econometricians quickly confirmed that U.S. data showed the same pattern, which was given the name *Phillips curve,* see Fig. 8.1.

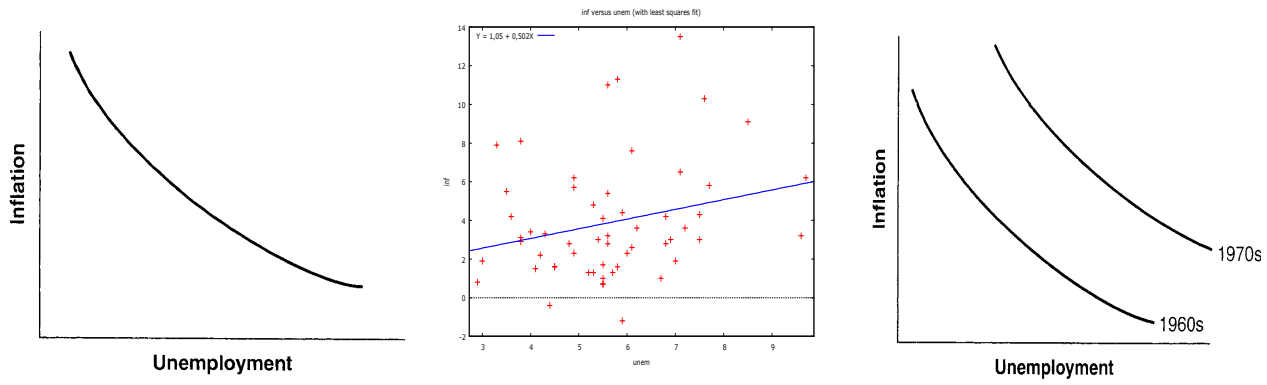


Figure 8.1. A theoretical Phillips curve (left), the regression line based on phillips.txt data (center) and model with a shift (right)

There was not a very solid theoretical basis for the correlation, but economists became very interested in deriving one. They produced a variety of models that suggested that economic policies that stimulated the economy, and thus reduced unemployment, would be inflationary, and those policies that cooled the economy off and increased unemployment would reduce inflation also.

In the early 1970s, however, inflation and unemployment rose simultaneously to levels not seen since the 1930s. As a result, data on unemployment and inflation rates fail to show the predicted pattern. From 1948 to 2000, there appears, in fact, to be a positive relationship between unemployment and inflation, as shown in Fig. 8.1, center. The graph presented a serious challenge to the theory of the Phillips curve, which suggested the two variables should not rise simultaneously.

Our further analysis is based on data from phillips.txt:

`metai`	years 1948 through 2003
`unem`	civilian unemployment rate, %
`inf`	percentage change in CPI
`inf_1`	inf(-1)
`unem_1`	unem(-1)

	metai	unem	inf	inf_1	unem_1
1948	1948	3.8	8.1		
1949	1949	5.9	-1.2	8.1	3.8
1950	1950	5.3	1.3	-1.2	5.9
1951	1951	3.3	7.9	1.3	5.3

Economists were reluctant to abandon the Phillips curve. They realized that a shift of the curve, as shown in Fig. 8.1, right, could explain the macroeconomic events of the 1970s. During the 1960s, the low Phillips curve would permit the economy to enjoy relatively low levels of unemployment and inflation, or to have very low levels of one in return for a modest rise in the other. When the curve shifted out, sometime around 1970, this become impossible; the economy would then have fairly high levels of both inflation and unemployment, and could have a low level of one only by tolerating very high levels of the other. If the Phillips curve moves over time, however, it will be difficult to estimate it econometrically. We locate an economic relationship in data by observing many points derived from that relationship, expecting that the error terms will tend to average out if we observe enough points. But this works only if the underlying relationship does not move. If the relationship changes over time, then each point is created by a different version of the regression equation and we never get more than one observation on each version. Only if we are willing to make some assumptions about how the relationship changes over time can we use those assumptions to estimate both the slope and intercept of the line, and the amount by which it has shifted.

To take into account possible shift, we introduce a dummy variable `D` for the years 1973 through 1983:

```
genr D = (metai>=1973)&&(metai<=1983)
ols inf const unem D
genr infh = $yhat
```

The graph in Fig. 8.2, left, presents `unem` on the x axis and `inf` with `infh` on the y axis. The OLS estimate of the slope $\hat{\beta}_1$ in $inf_t = \alpha + \beta_1 unem_t + \beta_2 D_t + u_t$ is now negative (-0,37), though not significant (p-value=0.10).

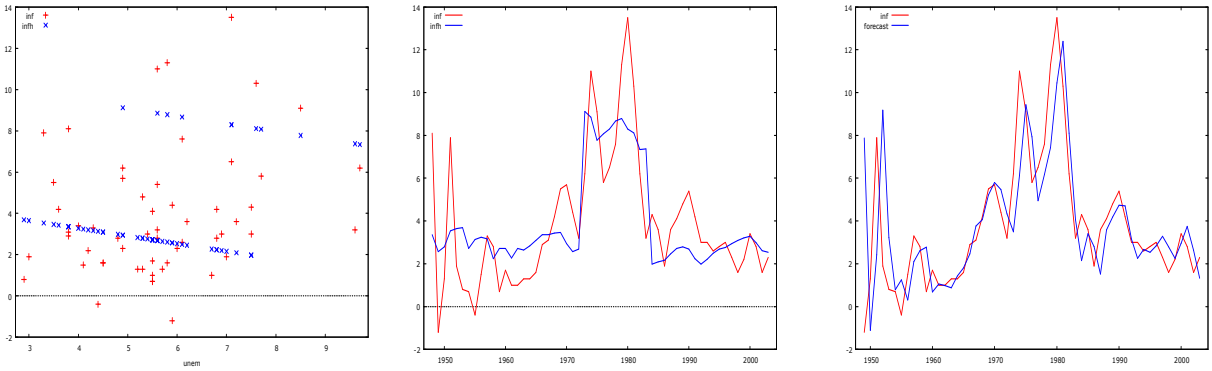


Figure 8.2. Scatter diagramm of `inf` and `infh` vs `unem` (left); time series plot (`inf` and `infh`) (center); time series plot (`inf` and its forecast from the equation for Δinf_t, see below) (right; note that the approximation, compared with the central graph, is much better)

The techniques discussed in the preceding sections are very good for demonstrating that curves have moved, and if they have, by how much. However, they are extremely unsatisfactory for one reason: they do not explain *why* the curve shifted. [redacted]

One of the theories suggests the following economic relationship: $inf_t - inf_t^{pred} = \alpha + \beta_1 unem_t + u_t$, where inf_t^{pred} is the prediction of inf_t made at moment $t-1$ (the model is called the *expectations-augmented Phillips curve*.) We take the simplest, the so-called naive prediction, given by the formula $inf_t^{pred} = inf_{t-1}$ and rewrite our equation as $\Delta inf_t = \alpha + \beta\, unem_t + u_t$. The OLS estimate is as follows:

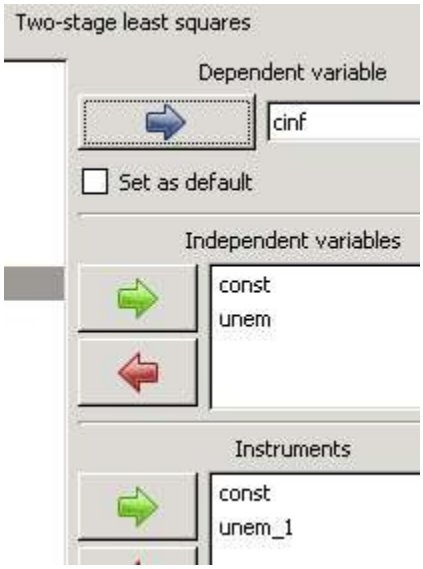
```
Model 2: OLS, using observations 1949-2003 (T = 55)
Dependent variable: d_inf

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        2,82820      1,22487      2,309    0,0249   **
  unem        -0,517649     0,209045    -2,476    0,0165   **
```

Thus, a one-point increase in `unem` lowers unanticipated inflation by over one-half of a point which is closer to „classical“ theory. (To plot the graph in Fig. 9.2, right, go to the Model 2 window and choose Analysis * Forecasts... * Produce forecast for `inf`).

Our final point is to doubt the assumption that `unem` and `u` are uncorrelated (this is where we shall use the IV technique). If we assume that they are correlated, OLS is not a reliable method and we need an instrument for $\texttt{unem}_t$. To test whether $\texttt{unem}_{t-1}$ could be used as an instrument, we regress $\texttt{unem}_t$ on $\texttt{unem}_{t-1}$:

```
Dependent variable: unem

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       1,48968      0,520203     2,864    0,0060    ***
  unem_1      0,742382     0,0892927    8,314    3,54e-011 ***
```

Since the regression coefficient is signicant, we can use $unem_{t-1}$ as an instrument. We shall apply 2SLS: go to Model * Instrumental variables * Two-stage Least Squares and fill in the boxes as shown on the right[2]. The expectations-augmented Phillips curve estimated by IV is




```
TSLS, using observations 1949-2003 (T = 55)
Dependent variable: d_inf
Instrumented: unem
Instruments: const unem_1

          coef    st.err       z p-value
  ------------------------------------------
  const    0,63     1,66     0,38    0,70
  unem    -0,13     0,29    -0,46    0,65


Hausman test -
Null hypothesis: OLS estimates are consistent
Asymptotic test statistic:
Chi-square(1) = 5,0715
with p-value = 0,0243223
```

The 2SLS estimator is less efficient than OLS when the explanatory variables are exogenous. Therefore, it is useful to have a test for endogeneity of an explanatory variable that shows whether 2SLS is necessary. It is common to use Hausman test to test for exogeneity[3]: in our case its p-value is less than 0.05, therefore unem is endogenous and it is correct to use IV. Note that the IV estimate of β (= - 0.13) is much lower in magnitude than the OLS estimate (= - 0.52), and $\hat{\beta}$ is not statistically different from zero. This means that the Phillips law in the present form is not applicable to our data. ◀◀

The first case where endogeneous variable on the right-hand-side emerges is a *measurement error in explanatory variable*. Assume that the right model is $Y = \alpha + \beta X + u$ but instead of X we observe $X^* = X + v$ where $Ev = 0$ and v does not depend on u. Thus our regression model is of the form $Y = \alpha + \beta X^* + (u - \beta v) = \alpha + \beta X^* + \varepsilon^*$, where X^* is endogeneous because $cor(X^*, \varepsilon^*) = E((X + v) \cdot (u - \beta v) = -\beta Ev^2 \neq 0$. Recall, we want to estimate β in $Y = \alpha + \beta X + u$, but since we do not have X, we replace it by endogenous X^*. Consequently, we have to look for an instrument for X^*.

[2] For a model to be identified, the *Instruments* list must be at least as long as that of *Independent variables*.
[3] In the Hausman test, the null hypothesis is that OLS estimators are consistent, or in other words, there is no need to use IV.

8.1 exercise. [redacted] Use the data in intdef.txt for this exercise. A simple equation relating the three-month T-Bill rate to the inflation rate (constructed from the consumer price index) is $i3_t = \alpha + \beta inf_t + u_t$.

(i) Estimate this equation by OLS, omitting the first time period for later comparisons. Report the results in the usual form.

(ii) Some economists feel that the consumer price index mismeasures the true rate of inflation, so that the OLS from part (i) suffers from measurement error bias. Reestimate the equation from part (i), using inf_{t-1} as an IV for inf_t. How does the IV estimate of β compare with the OLS estimate?

(iii) Now first difference the equation to obtain $\Delta i3_t = \beta \Delta inf_t + \Delta u_t$. Estimate this by OLS and compare the estimate of β with the previous estimates.

(iv) Can you use Δinf_{t-1} as an IV for Δinf_t in the differenced equation in part (iii)? Explain. (*Hint*: Are Δinf_t and Δinf_{t-1} sufficiently correlated?) ◀◀

8.2. System of Equations

Another important source of endogeneity is simultaneity. The reason that there are two equations in a supply and demand model is that there are two variables – Q for equilibrium quantity and P for equilibrium price – whose values the model explains:

$$\begin{cases} Q_t^D = \beta_0 + \beta_1 P_t + \beta_2 I_t + \varepsilon_t^D & \text{(demand equation)} \\ Q_t^S = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \varepsilon_t^S & \text{(supply equation)} \\ Q_t^D = Q_t^S = Q_t & \text{(equilibrium condition)} \end{cases}; \qquad (8.1)$$

the model can also contain some extra variables (in (8.1), it is I (the income of buyers) and W (the wage rate of seller's employees).) Note that (8.1) can be rewritten as

$$\begin{cases} Q_t = \beta_0 + \beta_1 P_t + \beta_2 I_t + \varepsilon_t^D & \text{(demand equation)} \\ Q_t = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \varepsilon_t^S & \text{(supply equation)} \end{cases}. \qquad (8.1^*)$$

A variable (such as Q and P) is *endogenous* to an economic model if its value is defined within the model. A variable (such as I and W) is *exogenous* to the model if its value is taken as given (i.e., is treated a fixed parameter) by the model (the market forces bring Q and P to equilibrium *together*, but market forces do not influence neither I nor W). These definitions are equivalent to the following ones: the right-hand-side variable of an equation is called endogeneous if it is *correlated* (and exogenous if it is *uncorrelated*) with the error term ε. Recall that the OLS estimates of the coefficients of an equation are BLUE only if certain

(Gauss-Markov) conditions are met, in particular, if all the right-hand-side variables are exogenous. But what happens if one of the right-hand-side variables is endogenous?

- If all the Gauss-Markov assumptions are true except the one of exogeneity then the OLS estimators of the coefficients become biased, inconsistent and inefficient.

Thus, we cannot apply the OLS to neither demand nor supply equation[4]. To cure the *structural* (or economic) system (8.1), solve the model for its endogenous variables – the new *reduced* (or econometric) system of the model will take the form of

$$\begin{cases} Q = \delta_0 + \delta_1 W + \delta_2 I + \varepsilon_Q \\ P = \pi_0 + \pi_1 W + \pi_2 I + \varepsilon_P \end{cases}. \tag{8.2}$$

Since W and I are exogenous, respective estimates are BLUE. They provide a simple description of the equilibrium of the model and of how it changes when the exogenous variables change. However, δ 's and π 's are not the slopes of the supply and demand lines. One possibility is to work backward from these values to slopes[5] but sometimes it is rather complicated or even impossible (in any case, it will not provide estimates of the standard errors of the β and γ parameters which are necessary to test hypothesis about them.) Therefore, to estimate the coefficients of the original equation, we apply a *two-stage least squares* procedure. Let K be the number of all the exogenous variables in the model (including a constant) and H_j the number of (unknown) coefficients in the *j* th structural equation. The necessary condition for the equation to be identified[6] (or estimable) is $K \geq H_j$[7]. We shall explain the procedure by means of example.

8.2 example. In system (8.1*), the list of exogenous variables consists of a constant, I, and W, therefore, $K = 3$. In the demand equation, we have three β 's, in the supply equation three γ 's, thus according to the order condition, we can proceed with both equations.

Stage 1. Using OLS, regress the endogenous variables on all of the exogenous variables (you have to estimate both equations in (8.2)).

Stage 2. Now estimate the structural equations (8.1) by OLS, replacing the endogenous variables with their predicted values, $\hat{Q}_t$ and $\hat{P}_t$, from Stage 1:

$$\begin{cases} \hat{Q}_t = \beta_0 + \beta_1 \hat{P}_t + \beta_2 I_t + \varepsilon_t^{(1)} \\ \hat{Q}_t = \gamma_0 + \gamma_1 \hat{P}_t + \gamma_2 W_t + \varepsilon_t^{(2)} \end{cases}$$

[4] The purpose of our analysis is to solve both equations in (8.1*), i.e., to estimate all β's and γ'. Since the rhs's contain endogenous variables, the least squares method would give us „bad" estimates.

[5] This called the *indirect least squares* method.

[6] This is called an order condition or counting rule. An equation is *identified* if we can find (consistent) estimates of all of its coefficients.

[7] This condition is only necessary (that is, if $K < H_j$, the equation is definitely not estimable). The necessary and sufficient condition is rather complicated but, in practice, one can safely use the order condition to establish identifiability.

It can be proved that the estimated $\hat{\beta}$'s and $\hat{\gamma}$'s from the second-stage regression are consistent estimators of the true β and γ parameters. They are biased, but the bias diminishes as the sample grows larger. ◀◀

The best way to memorize the above given identification rules is to learn the syntax of respective gretl command: the first equation of (8.1*) is solved with the help of the following command:

```
tsls Q 0 P I ; 0 W I
```

(here `Q 0 P I` denotes the equation while `0 W I` presents the list of exogenous variables.) The first right hand variable in the equation (a constant) is denoted by `0` and is explained by `0` behind the semicolon; another right hand variable, `P`, is endogenous but we have exogenous `W` which can be used as an instrument; finally, the right hand exogenous variable `I` is explained by `I` itself. Thus, we have enough intruments to solve the equation.

Note that in a similar system

$$\begin{cases} Q_t = \beta_0 + \beta_1 P_t + \varepsilon_t^D \\ Q_t = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \varepsilon_t^S \end{cases}$$

$K = 2$, $H_1 = 2$, and $H_2 = 3$, thus, the second equation is not identified (i.e., we <u>cannot</u> consistently estimate γ's from our data by <u>any</u> estimation method). Note that respective gretl command `tsls Q 0 P W; 0 W` will not work (the equation is not estimable) because we have three variables on the right-hand-side, but only two instruments.

Finally, in the system

$$\begin{cases} Q_t = \beta_0 + \beta_1 P_t + \varepsilon_t^D \\ Q_t = \gamma_0 + \gamma_1 P_t + \varepsilon_t^S \end{cases}$$

both equations are unidentified (i.e., if our data consists of the equilibrium data $(Q_1, P_1), ..., (Q_T, P_T)$ only, there is no way to estimate β's and γ's, see Fig. 8.3). Indeed, we can solve the system as

$$\begin{cases} Q_t = \delta_Q + \varepsilon_{Q,t} \\ P_t = \delta_P + \varepsilon_{P,t} \end{cases}$$

but we cannot restore <u>four</u> parameters $\beta_0, \beta_1, \gamma_0,$ and γ_1 from <u>two</u> parameters δ_Q and δ_P.

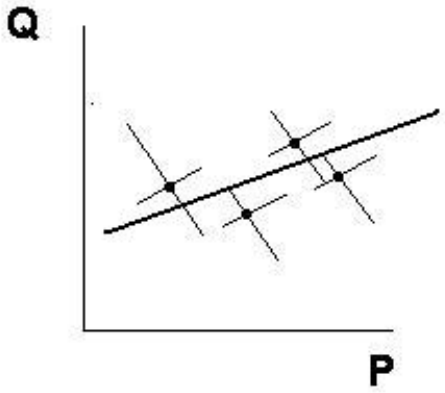


Figure 8.3. Both the demand and supply lines (i.e., the coefficients β 's and γ 's) are always the same; for every t, the shocks ε_t^D and ε_t^S move the demand and supply lines up or down; the bold regression line is neither demand nor supply line; both equations are unidentified

8.3 example. Consider the following IS-LM model:

$$\begin{cases} R_t = \beta_{11} + \beta_{12} M_t + \beta_{13} Y_t + \beta_{14} M_{t-1} + u_{1t} \\ Y_t = \beta_{21} + \beta_{22} R_t + \beta_{23} I_t + u_{2t} \end{cases}$$

where R denotes the interest rates, M denotes the money stock, Y is GDP, and I is investment expenditure. In this model, R and Y are endogenous variables and M, $M(-1)$, and I are exogenous variables (the first (LM) equation is exactly identified and the second one (IS) is overidentified[8]). The data for this example are in the file simult.xls and are annual time series data from 1969 to 1997 for the UK economy. [redacted] After importing the data, copy the following script to the GRETL script window

```
tsls R 0 M Y M(-1) ; 0 M M(-1) I
tsls Y 0 R I ; 0 M M(-1) I
```

and then run it.

```
? tsls R 0 M Y M(-1) ; 0 M M(-1) I


Model 1: TSLS, using observations 1970-1997 (T = 28)
Dependent variable: R
Instrumented: Y
Instruments: const M M_1 I

             coefficient   std. error       z      p-value
  ---------------------------------------------------------
  const       27,5275       11,1348        2,472    0,0134  **
  M            0,00187096    0,00185405    1,009    0,3129
  Y           -0,264700      0,224145     -1,181    0,2376
  M_1         -0,00172884    0,00175432   -0,9855   0,3244
```

[8] Overidentified means that there are more exogenous variables than coefficients to estimate.

```
Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 5,12934
  with p-value = 0,0235247
```

The coefficients of Y as well as M are insignificant, suggesting that the LM function is very flat.

```
? tsls Y 0 R I ; 0 M M(-1) I

Model 2: TSLS, using observations 1970-1997 (T = 28)
Dependent variable: Y
Instrumented: R
Instruments: const M M_1 I

             coefficient     std. error       z      p-value
  ------------------------------------------------------------
  const      98,7996        68,7067          1,438    0,1504
  R          -4,04296        3,13059        -1,291    0,1966
  I           0,000217952    0,000408970     0,5329   0,5941

Mean dependent var   79,16429   S.D. dependent var   14,07352
Sum squared resid    3191,649   S.E. of regression   11,29894
R-squared            0,596911   Adjusted R-squared   0,564664
F(2, 25)             19,97017   P-value(F)           6,57e-06
rho                  0,354478   Durbin-Watson        1,283596

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 23,9747
  with p-value = 9,76089e-007
```

Interpreting this model, we can say that income and the rate of interest are negatively related, according to the theoretical prediction, and income is quite sensitive to changes in the rate of interest. Also, a change in investment is would cause the function to shift to the right, again as theory suggests.

8.2 exercise. Repeat the analysis with the US data from US_all data for is-lm.xls.

8.3 exercise. Import the data file gasdemand.xls, which contains monthly observations from 1978:1 to 2002:08 on the following variables: `pricegas` (cents per gallon), `quantgas` (thousands of barrels per day), `persincome` (personal income, billions of dollars), and `carsales` (millions of cars per year). Consider the following supply and demand model for unleaded gasoline, where the first equation represents demand and the second represents supply:

$$\begin{cases} pricegas = \beta_0 + \beta_1 quantgas + \beta_2 persincome + \beta_3 carsales + \varepsilon_D & \text{(demand)} \\ pricegas = \gamma_0 + \gamma_1 quantgas + \varepsilon_S & \text{(supply)} \end{cases}$$

a. What variables in this system are endogenous and which are exogenous?

b. Is the demand (supply) equation overindentified, underidentified, or exactly identified?
c. Estimate the supply supply equation using (OLS and then) 2SLS. What value do you get for $\hat{\gamma}_1$? Does it take the sign you expect and, if so, is it statistically significantly different from zero?
d. Can you estimate the demand curve by 2SLS? If not, what variables might you add to the supply curve, and exclude from the demand curve, in order to provide instruments to estimate the demand curve?
e. Estimate the demand curve by OLS. What is your estimated value for β_1? This estimate is biased – is it biased upward or downward? What can you conclude about the slope of the demand curve from this information?

9. Simultaneous Equations

Most economic models involve more than one equation. We know how to estimate each equation in, say, supply and demand models, but given the links between the equations, we might reasonably ask whether it is possible to estimate the equations jointly (it turns out that it is possible.) Sometimes this improves the efficiency of the estimators, sometimes it has no effect on their efficiency, and sometimes it makes things worse.

There are two useful types of relationships between equations. First, there may be a relationship between the error terms of the model. Second, there may be relationships between the parameters of the model themselves, either because two parameters take the same value or because some more general mathematical relationship exists between two or more parameters.

9.1. Seemingly Unrelated Regression (SUR)

In the example below, we consider five firms: GM, Chrysler, GE, Westinghouse, and U.S. Steel. The data in greene13_1.gdt consist of time series of 20 yearly observations for these firms and three variables:

I_{it}	gross investment,
F_{it}	market value of the firm at the end of the previous year,
C_{it}	value of the stock of plant and equipment at the end of the previous year.

where $i = 1,...,5$, $t = 1935,...,1954$. At any moment t we can write five equations:

$$\begin{cases} I_{1t} = \beta_{10} + \beta_{11} F_{1t} + \beta_{12} C_{1t} + \varepsilon_{1t} \\ ..... \\ I_{5t} = \beta_{50} + \beta_{51} F_{5t} + \beta_{52} C_{5t} + \varepsilon_{5t} \end{cases} . \qquad (9.1)$$

Each equation can be estimated individually taking $t = 1935,...,1954$ (each firm performs its own investment policy, therefore we can treat these equations as unrelated.) On the other hand, all the economic activities take place in the same economic envinroment, therefore the five shocks $\varepsilon_{1t},...,\varepsilon_{5t}$ *can* be correlated. The procedure which takes into account this correlation is called *SUR*. Generally, it differs from OLS except for the two cases:

1. The equations (9.1) are really uncorrelated, i.e., $\text{cov}(\varepsilon_{it}, \varepsilon_{js}) = 0$, $t \neq s$ (and also $\text{cov}(\varepsilon_{it}, \varepsilon_{jt})$ $(= \sigma_{ij})$ does not depend on t.)
2. All the equations in (9.1) have the same explanatory variables on the right-hand-side, i.e., $F_{1t} = ... = F_{5t} = F_t$ etc (this is not true in our case.)

Now, copy and paste the following script to GRETL script window:

```
open greene13_1.gdt
greene <- system       # greene will be the system name
```

```
 equation I_GM const F_GM C_GM
 equation I_CH const F_CH C_CH
 equation I_GE const F_GE C_GE
 equation I_WE const F_WE C_WE
 equation I_US const F_US C_US
end system                          # we shall apply two methods of estimation:
estimate greene method=sur          # for the whole system with sur
estimate greene method=ols          # for each equation individually with ols
```

For simplicity, we shall present SUR and OLS models for only I_1 (that is, for GM investment):

```
Equation system, greene
Estimator: Seemingly Unrelated Regressions

Equation 1: SUR, using observations 1935-1954 (T = 20)
Dependent variable: I_GM

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const      -162,364      89,4592      -1,815    0,0872   *
  F_GM          0,120493    0,0216291     5,571    3,38e-05 ***
  C_GM          0,382746    0,0327680    11,68     1,52e-09 ***

Mean dependent var   608,0200   S.D. dependent var   309,5746
Sum squared resid    144320,9   S.E. of regression   84,94730
R-squared            0,921330   Adjusted R-squared   0,912075

*****
*****

Equation system, greene
Estimator: Ordinary Least Squares

Equation 1: OLS, using observations 1935-1954 (T = 20)
Dependent variable: I_GM

             coefficient   std. error    t-ratio   p-value
  ----------------------------------------------------------
  const      -149,782      105,842       -1,415    0,1751
  F_GM          0,119281     0,0258342     4,617    0,0002   ***
  C_GM          0,371445     0,0370728    10,02     1,51e-08 ***

Mean dependent var   608,0200   S.D. dependent var   309,5746
Sum squared resid    143205,9   S.E. of regression   91,78167
R-squared            0,921354   Adjusted R-squared   0,912102

*****
*****
```

The difference between two estimating methods is hardly noticable, probably, because of small correlation between the errors. Generally, using SUR to jointly estimate the equations of the system, allowing for correlation between the errors of the equations, will improve the efficiency of estimation, but usually not much. ◀◀

There is one potential problem with with simultaneous equations, which is that it requires the Gauss-Markov assumptions to be true for all equations. Suppose that the Gauss-Markov as-

sumptions are true for one equation but not for another. For example, one equation might have an omitted right-hand-side variable or an endogeneous one. The estimating by SUR will generally no longer be unbiased or consistent for *any* of the equations. In such a case, OLS would remain unbiased and consistent for those equations for which the Gauss-Markov assumptions held. Estimating equation by equation has the advantage that, if there is a problem with one equation, the problem is limited to that equation and cannot spill over to the estimates of the parameters of the other equations.

9.1 exercise. In this problem we shall look at a multiple equation model of international trade flows, and show SUR provides more efficient estimates than least squares does. Open the data file intltrade.xls, which contains monthly data from January 1990 to December 2000 on the following variables:

`TRADEDEFCANADA`	US trade deficit with Canada
`TRADEDEFJAPAN`	US trade deficit with Japan
`EXRATECANADA`	Canadian dollars per US dollars
`EXRATEJAPAN`	yen per US dollar
`RATECANADA`	Canadian government bond interest rate
`RATEJAPAN`	Japanese government bond interest rate
`RATEUS`	US Treasury bill interest rate

a. Estimate the OLS regression $\log(TRADEDEFCANADA) = \beta_0 + \beta_1 \cdot EXRATECANADA + \beta_2 RATECANADA + \beta_3 RATEUS + \varepsilon_{Ct}$. Observe that we have chosen the semilogarithmic functional form. If the Canadian dollar depreciates from 1.1 per US dollar to 1.2 per US dollar, by what percent does the US trade deficit with Canada rise, fall, or not change? Does it move in the direction you would expect to move?

b. Estimate the OLS regression $\log(TRADEDEFJAPAN) = \gamma_0 + \gamma_1 EXRATEJAPAN + \gamma_2 RATEJAPAN + \gamma_3 RATEUS + \varepsilon_{Jt}$, also in the semilogarithmic functional form. Is it reasonable to exclude $EXRATECANADA$ and $RATECANADA$ from this equation? If Japanese interest rates rise from 3 to 4 percent, does the trade balance rise, fall, or not change? Is this the answer you expect?

c. Would you expect ε_{Ct} and ε_{Jt} to be positively or negatively correlated?

d. Estimate both equations simultaneously using SUR? Have the parameter estimates changed a lot, or only slightly? Look at the standard errors of the parameter estimates. Are they larger or smaller than the standard errors of the single-equation estimates? Which technique is more eficient? ◀◀

Up to now, we have not made use of the economic connection between the equations in the system. We have allowed for the error terms to correlated, and we have some economic ideas about why the errors would be correlated, but the errors might well be correlated by coincidence even if there was no economic link between the equations at all. We may do better if we can use economic theory to suggest direct links between the parameters of the equations of the system. For example, the condition $\beta_2 = \gamma_2 (= \delta)$ in 9.1 exercise is called a *cross-equation restriction*. If we use OLS to estimate individually both equations from the above system, we have to minimize $RSS = RSS_1(\beta_0, \beta_1, \delta, \beta_3) + RSS_2(\gamma_0, \gamma_1, \delta, \gamma_3)$. In the case where we use

SUR to estimate the parameters, we have to generalize RSS and to include the effects of the correlation of the error terms.

9.1 example. In this example we shall see the effect of imposing a cross-equation restriction on the efficiency of SUR estimates. Open the data file intltrade, which was described in 10.1 exercise.

a. Consider the equations

$$\begin{cases} \log(TRADEDEFCANADA) = \beta_0 + \beta_1\, EXRATECANADA + \beta_2\, RATECANADA + \beta_3 RATEUS + \varepsilon_C \\ \log(TRADEDEFJAPAN) \quad = \gamma_0 + \gamma_1\, EXRATEJAPAN \quad + \gamma_2\, RATEJAPAN \quad + \gamma_3 RATEUS + \varepsilon_J \end{cases}$$

If a 1 percentage point increase in the interest rate of a foreign country has the same effect on trade for all countries, then what should be true about the model's parameters?

b. Estimate each equation separately with OLS. What estimated values do you get for β_2 and γ_2? How far apart are the two values? What are their standard errors? Does it look plausible that they might have the same true value? *Hint*. Import the file, add logs of the left-hand-side variables, go to Model * Simultaneous equations…, fill in the box with expressions similar to `equation I_GM const F_GM C_GM`, and choose Estimator OLS.

c. Estimate the two equations by SUR, and test the null hypothesis $\beta_2 = \gamma_2$. Do you reject or fail to reject this hypothesis? *Hint*. Repeat the above procedure but replace OLS with SUR; go to Tests * Linear restrictions and type in `b[1,3]-b[2,3]=0`. The bottom lines

```
F test for the specified restrictions:
F(1,256) = 1.15328 [0.2839]
```

mean that there is no ground to reject the null.

d. What value do you get for their common value? If the interest rate of a foreign country rises by 1 percent, how much does the U.S. trade deficit with that country fall?

9.2 example. To estimate a VAR(1) model in two variables one can use either the command

```
var 1 Y X
```

or

```
system method=sur
equation X const  X(-1) Y(-1)
equation Y const X(-1) Y(-1)
end system
```

9.2. Multiple Equations with Endogeneous Right-hand-side Variables

SUR is a useful technique for models that can be estimated by least squares. However, it cannot be used if the Gauss-Markov assumptions are not satisfied. In particular, if the equations contain endogenous right-hand-side variables, SUR will be biased and inconsistent. Any time we have two equations solving for the values of two variables, such as the supply and demand model, there will be endogenous right-hand-side variables and SUR will not be appropriate.
Fortunately, we can simultaneously estimate equations by two-stage least squares in exactly the same way that we simultaneously estimate them by ordinary least squares. Doing so requires a three-step process:

1. Regress each endogenous variable on all exogenous variables in the system of equations, and calculate predicted values for the endogenous variables.
2. Estimate the structural equations by least squares, replacing the endogenous right-hand-side variables with their predicted values from step 1.
3. Calculate the estimated variances and covariances of the residuals from step 2, and reestimate the structural equations using the SUR method.

This technique, known as *three-stage least squares,* is the instrumental variables equivalent to SUR. It has the same general relationship to two-stage least squares that SUR has to OLS. Its advantage is that it will be more efficient than two-stage least squares for large samples, as long as the right-hand-side variables of the equations are not the same in all equations. It is not unbiased—but two-stage least squares is not unbiased either, so that is not a disadvantage of three-stage least squares. Its main disadvantage is that, as with SUR, simultaneous estimation permits a violation of the Gauss-Markov assumptions in one of the equations to spread to the other equations.

****

We have already discussed the problem of the endogenous right-hand-side variables and we presented three methods to deal with it. The first one was a rather cumbersome indirect least squares method, the other were two- and three-stages least square methods. Now we shall briefly describe two more methods presented in GRETL. The single equation or limited-information methods (specifically, the *limited information maximum likelihood* method) was introduced in 1949 and was popular until the advent of 2SLS. Computationally it is rather complicated but if the equation under consideration is exactly identified, then LIML and 2SLS give identical estimates.

To estimate the coefficients of an equation, LIML uses the information of that equation only. In contrast, in system or *full-information* methods we use information on the restrictions on all equations.

9.3 example. In 1950, L. Klein proposed the dynamic model of macroeconomics which was later called Klein Model 1. It is described by the following system:

$$\begin{cases} C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t} & \text{(consumption)} \\ I_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} & \text{(investment)} \\ W_t^p = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} & \text{(demand for labor)} \\ X_t = C_t + I_t + G_t & \text{(equilibrium demand)} \\ P_t = X_t - T_t - W_t^p & \text{(private sector revenue)} \\ K_t = K_{t-1} + I_t & \text{(capital)}, \end{cases}$$

where

C	consumption expenditure
I	investment expenditure
G	government expenditure
P	profits
W^p	private wage bill
W^g	government wage bill
K	capital stock
T	taxes
X	income after tax
A	time trend

In the preceding model, the left-hand-side variables *C, I, W, Y, P,* and *K* are treated as jointly dependent, or endogenous, variables, G, T, W^g, and A as exogenous, and the variables P_{t-1}, K_{t-1}, and X_{t-1} are treated as predetermined. In all, there are six equations (including the three identities) to study the interdependence of six endogenous variables.

Note that because of the interdependence among the endogenous variables, in general they are not independent of the stochastic disturbance terms, which therefore makes it inappropriate to apply the method of OLS to an individual equation in the system (the estimators thus obtained are inconsistent, they do not converge to their true population values even when the sample size is very large.)

When estimating, we shall apply all the methods we know.

```
open klein.gdt

genr W = Wp + Wg
genr A = t + (1918 - 1931)
genr K1 = K(-1)

# set the model up as a system
"Klein Model 1" <- system
 equation C 0 P P(-1) W
 equation I 0 P P(-1) K1
 equation Wp 0 X X(-1) A
 identity P = X - T - Wp
 identity W = Wp + Wg
 identity X = C + I + G
 identity K = K1 + I
 endog C I Wp P W X K
end system
```

```
# and estimate it in various ways
estimate "Klein Model 1" method=ols
estimate "Klein Model 1" method=tsls
estimate "Klein Model 1" method=3sls
estimate "Klein Model 1" method=fiml --verbose
estimate "Klein Model 1" method=liml
```

9.2 exercise. Consider the following **IS-LM** model:

$$\begin{cases} R_t = \beta_{11} + \beta_{12} M_t + \beta_{13} Y_t + \beta_{14} M_{t-1} + u_{1t} & \text{(LM relationship)} \\ Y_t = \beta_{21} + \beta_{22} R_t + \beta_{23} I_t + u_{2t} & \text{(IS relationship)}, \end{cases}$$

where R denotes the interest rate, M denotes the money stock, Y is GDP, and I is investment expenditure. In this model, R and Y are the endogenous variables and M and I are the exogenous variables. The first equation is exactly identified and the second one is overidentified (check). The annual data (from 1969 through 1997) are in the file simult.xls. Use appropriate methods to estimate the system. *Hint*. Use Model * Simultaneous equations... * choose tsls (or other relevant option), fill in the box with

```
equation R 0 M Y M(-1)
equation Y 0 R I
instr 0 M M(-1) I   # or: endog R Y
```

and click OK. Draw relevant graphs through Analysis * Forecasts.

9.3 exercise. In 9.3_full.xls, monthly data are provided on commercial banks' loans to bussiness firms in the United States for 1979:01-1984:12.

	N	Q	R	RD	X	RS	y
1979:01	1	251.8	11.75	9.25	150.8	9.35	994.3
1979:02	2	255.6	11.75	9.26	151.5	9.32	1002.5
1979:03	3	259.8	11.75	9.37	152.0	9.48	994.0
.....							

The following demand-supply model has been estimated. [redacted]

Demand for loan by business firms: $Q = \beta_0 + \beta_1 R + \beta_2 RD + \beta_3 X + u$

and supply by banks of commercial loans: $Q = \alpha_0 + \alpha_1 R + \alpha_2 RS + \alpha_3 y + v$

where

Q	total commercial loans (billions of dollars)
R	average prime rate charged by banks
RS	3-month Treasure bill rate (represents an alternative rate of return for banks)
RD	AAA corporate bond rate (represents the price of alternative financing to firms)
X	industrial production index and represents firms' expectation about future economic activity
y	total bank deposits (represents a scale variable) (billions of dollars)

Draw the necessary graphs. Is it true that both equations are overidentified? Which methods to estimate the system you can choose? R_t is expected to have a negative sign in the demand function and a positive sign in the supply function. The coefficient of RS_t is expected to be negative. The coefficients of RD_t, X_t, and y_t are expected to be positive.

[redacted]

****

Here is a summary of this chapter: the set of economic variables $Y_1,...,Y_k$ is determined through a market equilibrium mechanism and we want to analyze the structure of relationships that determines the equilibrium. Suppose that $\vec{Y} = (Y_1,...,Y_n)$ is a vector consisting of *n* economic variables, among which there exist *n* relationships that determine the equilibrium levels of the variables. We also suppose that there exist *x* variables $\vec{Z} = (Z_1,...,Z_x)$ that are independent of the economic relations but affect the equilibrium. The variables $\vec{Y}$ are called endogenous variables, and the $\vec{Z}$ are called exogenous variables. If we assume linear relationships among them, we have an expression such as

$$\vec{Y} = B\vec{Y} + \Gamma\vec{Z} + \vec{u} \qquad (*)$$

where B and Γ are matrices with constant coefficients and $\vec{u}$ is a vector of disturbances or errors. (*) is called the linear *structural* equation system and is a system of simultaneous equations. By solving the equations formally, we get the so-called *reduced* form

$$\vec{Y} = \Pi\vec{Z} + \vec{v} \qquad (**)$$

where $\Pi = (I - B)^{-1}\Gamma, \vec{v} = (I - B)^{-1}\vec{u}$. The relation of $\vec{Y}$ to $\vec{Z}$ is determined through the reduced form (**), and if we have enough data on $\vec{Y}$ and $\vec{Z}$, we can estimate Π. The problem of identification is to decide whether we can determine the unknown parameters in B and Γ uniquely from the parameters in the reduced form. A necessary condition for the parameters in one of the equations in (*) to be identifiable is that the number of unknown variables in the equation not be greater than $x + 1$. If it is exactly equal to $x + 1$, the equation is said to be *just identified*, and if it is less than $x + 1$, the equation is said to be *overidentified*.

If all the equations in the system are just identified, for arbitrary Π there exist unique B and Γ that satisfy $\Pi = (I - B)^{-1}\Gamma$. Therefore, if we denote the least squares estimator of Π by $\hat{\Pi}$, we can estimate B and Γ from the equation $(I - \hat{B})\hat{\Pi} = \hat{\Gamma}$. This procedure is called the *indirect least squares* method and is equivalent to the maximum likelihood method if we assume normality for $\vec{u}$.

When some of the equations are overidentified, the estimation problem becomes complicated. Three kinds of procedures have been proposed: (1) full system methods, (2) single equation

methods, and (3) subsystem methods. In full system methods all the parameters are considered simultaneously, and if normality is assumed, the maximum likelihood estimator can be obtained by minimizing $|(\vec{Y}-\Pi\vec{Z})(\vec{Y}-\Pi\vec{Z})'|$. Since it is usually difficult to compute the maximum likelihood estimator, a simpler, but asymptotically equivalent, *three stage least squares* method has been proposed. The single equation methods and the subsystem methods take into consideration only the information about the parameters in one equation or in a subset of the equations, and estimate the parameters in each equation separately. There is a single equation method, called the *limited information maximum likelihood* method, based on the maximum likelihood approach, and also a *two-stage least squares* method, which estimates Π first by least squares, computes $\hat{\vec{Y}}=\hat{\Pi}\vec{Z}$, and then applies the least squares method to the model $\vec{Y}=B\hat{\vec{Y}}+\Gamma\vec{Z}+\hat{u}$. These two and also some others are asymptotically equivalent. Among asymptotically equivalent classes of estimators corresponding to different information structures it has been established that the maximum likelihood estimators have asymptotically higher-order efficiency than other estimators, and Monte Carlo and numerical studies show that they are in most cases better than others if properly adjusted for the biases.

In many simultaneous equation models which have been applied to actual macroeconomic data, the values of endogenous variables obtained in the past appear on the right-hand sides of equations (*). Such variables are called *lagged variables*, and they can be treated, at least in the asymptotic theory of inference, as though they were exogenous. Hence exogenous variables and lagged endogenous variables are jointly called *predetermined variables*. When many lagged variables appear over many time periods and when some structure among the coefficients of those lagged variables can be assumed, such a model is called a *distributed lag* model.

10. Panel Data Analysis

Panel data combines the cross section and time series data. Here we have a cross section, but we observe the cross section over time. If the *same* people or states or units, sampled in the cross section, are then re-sampled at a *different* time we call this a panel data set (usually there will be many cross sectional units and several time periods). For example, in order to estimate the production function, we can use the model

$$Y_{it} = \alpha_i + \beta_1 X_{it}^{(1)} + \beta_2 X_{it}^{(2)} + u_{it},\ i = 1,...,I,\ t = 1,...,T$$

where Y_{it} is the output and $X_{it}^{(1)}, X_{it}^{(2)}$ are the inputs for the ith firm in the tth period; α_i capture firm specific inputs (e.g., managerial skills) assumed to be constant over time (this model later will be called a fixed effects model).

The famous Grunfeld panel data (see gg.txt) consists of 5 large US manufacturing firms over 20 years (100 observations):

invest	value	capital	firm	year
317.60	3078.5	2.8	General_Motors	1935
391.80	4661.7	52.6	General_Motors	1936
410.60	5387.1	156.9	General_Motors	1937
257.70	2792.2	209.2	General_Motors	1938
330.80	4313.2	203.4	General_Motors	1939
461.20	4643.9	207.2	General_Motors	1940
512.00	4551.2	255.2	General_Motors	1941
448.00	3244.1	303.7	General_Motors	1942
499.60	4053.7	264.1	General_Motors	1943
547.50	4379.3	201.6	General_Motors	1944
561.20	4840.9	265.0	General_Motors	1945
688.10	4900.9	402.2	General_Motors	1946
568.90	3526.5	761.5	General_Motors	1947
529.20	3254.7	922.4	General_Motors	1948
555.10	3700.2	1020.1	General_Motors	1949
642.90	3755.6	1099.0	General_Motors	1950
755.90	4833.0	1207.7	General_Motors	1951
891.20	4924.9	1430.5	General_Motors	1952
1304.40	6241.7	1777.3	General_Motors	1953
1486.70	5593.6	2226.3	General_Motors	1954
40.29	417.5	10.5	Chrysler	1935
72.76	837.8	10.2	Chrysler	1936
.....	.....	.....	.....	.....
90.08	1193.5	174.8	Westinghouse	1953
68.60	1188.9	213.5	Westinghouse	1954
209.90	1362.4	53.8	US_Steel	1935
355.30	1807.1	50.5	US_Steel	1936
469.90	2676.3	118.1	US_Steel	1937
262.30	1801.9	260.2	US_Steel	1938
230.40	1957.3	312.7	US_Steel	1939
261.60	2202.9	254.2	US_Steel	1940
472.80	2380.5	261.4	US_Steel	1941
445.60	2168.6	298.7	US_Steel	1942
361.60	1985.1	301.8	US_Steel	1943
288.20	1813.9	279.1	US_Steel	1944

```
  258.70 1850.2   213.8          US_Steel 1945
  420.30 2067.7   232.6          US_Steel 1946
  420.50 1796.7   264.8          US_Steel 1947
  494.50 1625.8   306.9          US_Steel 1948
  405.10 1667.0   351.1          US_Steel 1949
  418.80 1677.4   357.8          US_Steel 1950
  588.20 2289.5   342.1          US_Steel 1951
  645.20 2159.4   444.2          US_Steel 1952
  641.00 2031.3   623.6          US_Steel 1953
  459.30 2115.5   669.7          US_Steel 1954
```

Here `invest` = I_{it} denotes real gross investment for firm i in year t, `value` = F_{it} is the real value of the firm (shares outstanding), and `capital` = C_{it} is the real value of the capital stock. When importing gg.txt in GRETL, you must choose the panel structure of dataset. Note that the names of the firms in GRETL will be replaced by their numbers.

The above presentation is termed a **stacked time series** (one time series is stacked above another). If one cross-section is above another, this is called a **stacked cross section**[1]. If we remove the firm and year attributes and do not make any distinction between cross section and time series, this is called a **pooled** data organization.

10.1. Panel Data Models

Basically, there are three models for panel data – pooled OLS, fixed and random effects models. We want to explain I in terms of F and C and thus to estimate the following investment equation:

$$I_{it} = \alpha + \beta_1 F_{it} + \beta_2 C_{it} + u_{it},\ i = 1,...,5,\ t = 1935,...,1954\,.$$

The simplest is the **pooled model** where

$$I_s = \alpha + \beta_1 F_s + \beta_2 C_s + u_s,\ s = 1,...,100\,.$$

(i.e., the same intercept α and the same slopes β_1 and β_2 for all observations). To create it, go to Model| Ordinary Least Squares… etc

```
Model 1: Pooled OLS, using 100 observations
Included 5 cross-sectional units
Time-series length = 20
Dependent variable: invest

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -48.0297      21.4802      -2.236     0.0276    **
  value        0.105085     0.0113778    9.236     5.99e-015 ***
  capital      0.305366     0.0435078    7.019     3.06e-010 ***
```

[1] No matter whether your original data is organized as stacked cross sections or stacked time series, GRETL always stores and displays the panel data as stacked time series.

Mean dependent var	248.9570	S.D. dependent var	267.8654
Sum squared resid	1570884	S.E. of regression	127.2583
R-squared	0.778856	Adjusted R-squared	0.774296
F(2, 97)	170.8140	P-value(F)	1.65e-32
Log-likelihood	-624.9928	Akaike criterion	1255.986
Schwarz criterion	1263.801	Hannan-Quinn	1259.149
rho	0.936605	Durbin-Watson	0.218933

The model is fairly good (correct signs of the coefficients, high R-squared), however note i) the heteroscedasticity of the residuals (go to the model window, select Save| Residuals, then go to the GRETL window, right-click on `uhat1` and choose Panel plot…| single graph: groups in sequence| OK, see Fig. 10.1, left) and ii) low Durbin-Watson statistics (the residuals must form WN, but this will be only when $DW \approx 2$; now they are probably serially correlated). Recall that often the cause of the serial correlation is not the autoregressive structure of the

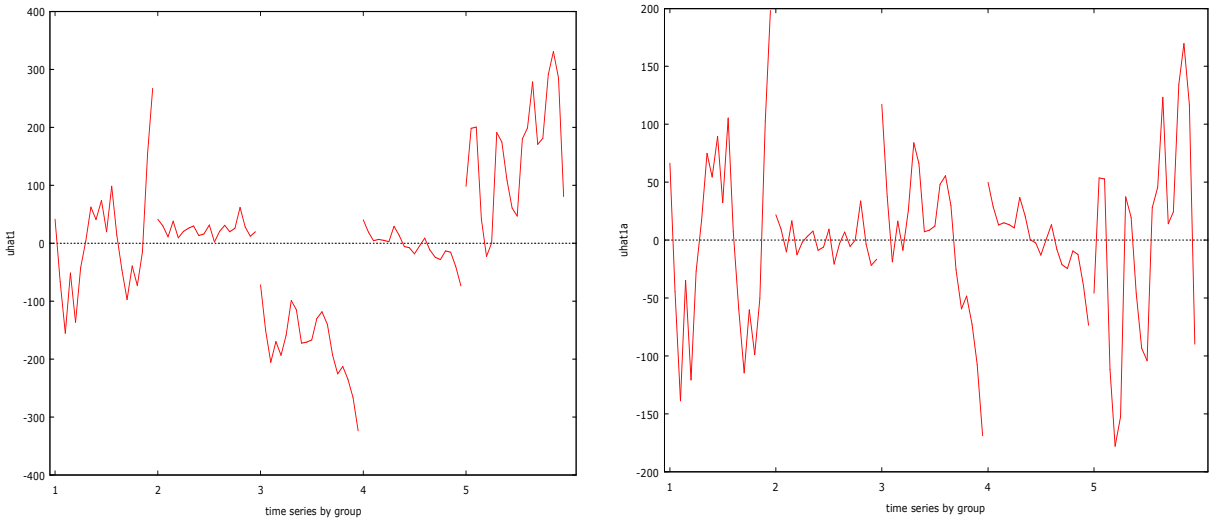


Figure 10.1. Residuals `uhat1` of the pooled model (left, note different levels of the residuals) and the residuals `uhat2` of the fixed effects model (right, note more homoscedastic behavior)

time series but wrong specification of the model. To be more concrete, it is quite possible that every firm has the same slopes β_i but different intercepts. We shall express this in different words – in another, **fixed effects** (FE), model we decompose the error term u_{it} into a unit-specific (and time invariant) component α_i and observation-specific error ε_{it}:

$$I_{it} = \beta_1 F_{it} + \beta_2 C_{it} + \alpha_i + \varepsilon_{it},\ i = 1,...,5,\ t = 1,...,20\,.$$

The α_i s are then treated as fixed parameters[2] which must be estimated. However, these individual intercepts are typically not of much inherent interest and also their estimated values are difficult to judge because there is often little data being used to estimate them (the time series are usually short). Instead, we are usually more interested in the slope coefficients.

[2] α_i may be treated as the mean of the error in the i th unit.

Note that the model can be written differently, in a more explicit way:

$$\begin{cases} I_{1,1} = \alpha + \alpha_1 + \beta_1 F_{1,1} + \beta_2 C_{1,1} + e_{1,1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I_{1,20} = \alpha + \alpha_1 + \beta_1 F_{1,20} + \beta_2 C_{1,20} + e_{1,20} \\ ********** \\ I_{2,1} = \alpha + \alpha_2 + \beta_1 F_{2,1} + \beta_2 C_{2,1} + e_{2,1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I_{2,20} = \alpha + \alpha_2 + \beta_1 F_{2,20} + \beta_2 C_{2,20} + e_{2,20} \\ ********** \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ ********** \\ I_{5,1} = \alpha + \alpha_5 + \beta_1 F_{5,1} + \beta_2 C_{5,1} + e_{5,1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I_{5,20} = \alpha + \alpha_5 + \beta_1 F_{5,20} + \beta_2 C_{5,20} + e_{5,20} \end{cases}$$

To create the model, go to Model| Panel| Fixed or random effects..., fill in the dialog boxes and choose Fixed effects:

```
Model 2: Fixed-effects, using 100 observations
Included 5 cross-sectional units
Time-series length = 20
Dependent variable: invest

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -62.5944      29.4419      -2.126     0.0361    **
  value        0.105980     0.0158910    6.669     1.82e-09  ***
  capital      0.346660     0.0241612   14.35      2.63e-025 ***

Mean dependent var   248.9570   S.D. dependent var   267.8654
Sum squared resid    444288.4   S.E. of regression   69.11798
R-squared            0.937454   Adjusted R-squared   0.933419
F(6, 93)             232.3194   P-value(F)           1.07e-53
Log-likelihood      -561.8468   Akaike criterion     1137.694
Schwarz criterion    1155.930   Hannan-Quinn         1145.074
rho                  0.606057   Durbin-Watson        0.774518

Test for differing group intercepts -
  Null hypothesis: The groups have a common intercept
  Test statistic: F(4, 93) = 58.9557
  with p-value = P(F(4, 93) > 58.9557) = 1.07556e-024
```

The residuals of the model now are more homogeneous (see Fig. 10.1, right), the Akaike statistics is lower (remember that now we have four more parameters), and the last lines of the model table tell us that the hypothesis $H_0 : all\ \alpha_i's\ are\ equal$ must be rejected; thus the $F-$test (what it exactly does?) tells us that this FE model is preferred to the previous, pooled OLS model.

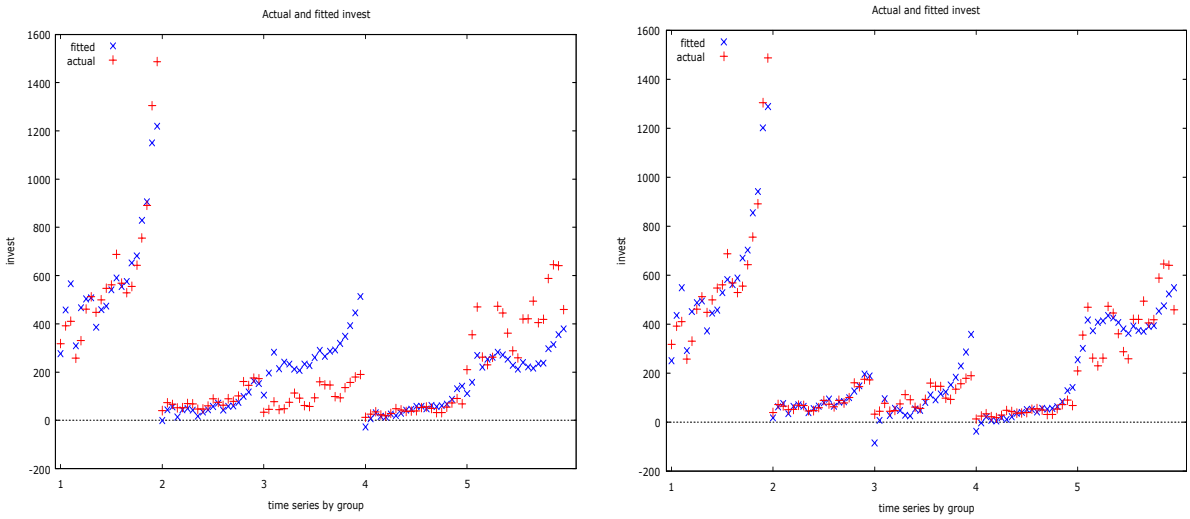


Figure 10.2. The fixed effects model (right) fits `invest` better that the pooled model (left)

To display the intercepts, in the Model 2 window go to Save| Per-unit constants, then in GRETL's window right-click on `ahat2`, choose Panel plot etc (note that the unit-specific constants vary a lot, between -250 and 100).

For the **random effects** (RE) model we write $u_{it} = v_i + \varepsilon_{it}$, so the model becomes

$$I_{it} = \beta_1 F_{it} + \beta_2 C_{it} + v_i + \varepsilon_{it},\ i = 1,...,5,\ t = 1,...,20. \qquad (10.1)$$

In contrast to the FE model , the v_is are now treated not as fixed parameters but as random drawings from a given probability distribution (thus, v_i have no particular meaning, they are just realizations of a certain random variable).

To create this model, go to Model| Panel| Fixed or random effects..., fill in the dialog boxes and choose Random effects:

```
Model 3: Random-effects (GLS), using 100 observations
Included 5 cross-sectional units
Time-series length = 20
Dependent variable: invest

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -60.2905      54.4839      -1.107     0.2712
  value        0.104886     0.0147972    7.088     2.20e-010 ***
  capital      0.346016     0.0242535   14.27      1.48e-025 ***

Log-likelihood      -625.6905   Akaike criterion      1257.381
Schwarz criterion    1265.196   Hannan-Quinn          1260.544

'Within' variance = 4777.3   # = σ̂²_ε
'Between' variance = 11191   # = σ̂²_v
theta used for quasi-demeaning = 0.853903
```

```
Breusch-Pagan test -
  Null hypothesis: Variance of the unit-specific error = 0
  Asymptotic test statistic: Chi-square(1) = 453.822
  with p-value = 1.06252e-100

Hausman test -
  Null hypothesis: GLS estimates are consistent
  Asymptotic test statistic: Chi-square(2) = 3.30507
  with p-value = 0.191563
```

This model table contains the outputs of several tests.

1. In the Breusch–Pagan test, the null hypothesis is that the variance of v_i in equation (10.1) equals zero: if this hypothesis is not rejected, then we conclude that the simple pooled model is adequate (in our case, we conclude that pooled model is inadequate).
2. The same and even more concrete conclusion can be infered from the theta value – if $\theta \approx 1$, then the FE estimator is optimal; if $\theta \approx 0$, then the pooled model is optimal. Thus, in our case we choose FE model.
3. The Hausman test (see PE.I, Lecture Notes, p. 4-55 or PE.II, Lecture Notes, p.8-5) probes the null hypothesis that RE model is preferrable to that of the fixed effects (we see that in our case there is no ground to discard RE model). However, note that Akaike's criterion in RE case exceeds that of FE. Also, if you plotted the graphs of the fitted values and residuals of the RE model (do this), they appear to be inferior to those of the FE model. Thus, at the moment we stick to the FE model.

There are several means to augment the model.

1. Include in the FE model time dummies `dt_1`,..., `dt_20` (it is quite possible that investment policy changes over time).
2. Define unit dummies `du_1`,..., `du_5` and create a model with these variables in the intercept (this is exactly the same as FE model).
3. Add also these dummies to the slopes of value and capital[3] (this is the same as separate models for each firm).

It appears that the „best" model is the following:

```
Model 8: Pooled OLS, using 100 observations
Dependent variable: invest

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -30.7689      29.4704      -1.044    0.2993
  value        0.156761     0.0211516    7.411    6.89e-011 ***
  capital      0.423923     0.0945327    4.484    2.17e-05  ***
  c_du1       -0.0428438    0.0978435   -0.4379   0.6625
  c_du2       -0.111477     0.165965    -0.6717   0.5035
  c_du3       -0.268581     0.110130    -2.439    0.0167    **
  c_du4       -0.411074     0.343709    -1.196    0.2349
  v_du1       -0.0653505    0.0180792   -3.615    0.0005    ***
  v_du2       -0.0444975    0.0442948   -1.005    0.3178
  v_du3       -0.120664     0.0207684   -5.810    9.55e-08  ***
  v_du4       -0.0522707    0.0619619   -0.8436   0.4012
Mean dependent var   248.9570   S.D. dependent var   267.8654
```

[3] Add new variables `c_du1 = capital * du_1` etc

Sum squared resid	352185.1	S.E. of regression	62.90577
R-squared	0.950420	Adjusted R-squared	0.944850
F(10, 89)	170.6096	P-value(F)	1.45e-53
Log-likelihood	-550.2310	Akaike criterion	1122.462
Schwarz criterion	1151.119	Hannan-Quinn	1134.060
rho	0.584909	Durbin-Watson	0.834968

Recall that we usually create panel data models which assume different intercepts and equal slopes for all units; however, as you have just seen, sometimes it is better to take the same constant and individual slopes.

The common problem with the panel data is to decide which panel method should one use, FE or RE? We have already discussed some tests which help to decide upon the model (and the best approach is to make use of these tests). However, some heuristic considerations are also sometimes helpful. If the panel comprises observations on a fixed and relatively small set of units of interest (say, the member states of the European Union), there is a presumption in favor of fixed effects. If it comprises observations on a large number of randomly selected individuals (as in many epidemiological and other longitudinal studies), there is a presumption in favor of random effects.

Note that some panel data sets contain variables whose values are specific to the cross-sectional unit but which do not vary over time (for example, the sex of an individual). If you want to include such variables in the model, the fixed effects option is simply not available (when the FE approach is implemented using dummy variables, the problem is that the time-invariant variables are perfectly collinear with the per-unit dummies). In this case, you should use RE approach (but test first that the individual effects v_i are not correlated with some of the explanatory variables - this is what the Hausman test probes).

10.2. Autoregressive Panel Models

Panel data models are after all regression models, therefore we can analyze dynamic regression models as well, for example

$$Y_{it} = \alpha + \gamma Y_{i,t-1} + \beta X_{it} + (v_i + \varepsilon_{it}),\ i = 1,...,N,\ t = 1,...,T\ . \qquad (10.2)$$

However, some specific problems arise when a lag of the dependent variable is included among the regressors in a panel model (this is connected with the fact that $Y_{i,t-1}$ is bound to be correlated with the error term v_i and, in this case, the OLS estimators of the coefficients will be biased and inconsistent). One strategy for handling this problem is to take first differences of (10.2) for sweeping out the group effects:

$$\Delta Y_{it} = \gamma \Delta Y_{i,t-1} + \beta \Delta X_{it} + (\varepsilon_{it} - \varepsilon_{i,t-1})\ .$$

To remove the still existing correlation between $\Delta Y_{i,t-1}$ and $\varepsilon_{it} - \varepsilon_{i,t-1}$, it is suggested to use an instrument[4] for $\Delta Y_{i,t-1}$ (it can be $Y_{i,t-2}$ or $\Delta Y_{i,t-2}$) .

[4] In GRETL, this is accomplished by the two stage least squares procedure `tsls`.

10.1 example. We shall analyze the panel data set penngrow.txt (or penngrow.gdt from Open Datal Sample files) – it contains 94 states in 6 time moments:

```
   country year        Y       X
1       53 1960 7.987185      NA
2       53 1965 8.403801 1.62254
3       53 1970 8.896177 1.76904
4       53 1975 9.033126 1.78886
5       53 1980 9.217117 1.78772
6       53 1985 9.373394 1.73075
7       70 1960 8.544225      NA
8       70 1965 8.720297 1.46819
9       70 1970 8.923191 1.57635
.............................
```

where

`country`	country code
`year`	year of observation
`Y`	log of real per capita GDP
`X`	log(savings)-log(population growth) (5-year average)

The well-known convergence theory claims that the lower the starting level of real per capita GDP, relative to the long run of steady state position, the faster is the growth rate which means that developing countries will finally catch with the richer ones.

Testing the growth convergence hypothesis, in this context, resolves largely around the coefficient γ in (10.2) (the rate of convergence equals $1/\gamma$, $0<\gamma<1$). If (the estimate of) γ is much less than 1, the implication is that on average countries with low initial values are growing faster than those with high initial values and is therefore evidence of convergence. Whereas, if this coefficient is close to one, perhaps even slightly larger than one, the implication is that initial values have little or no effect or even a perverse one on subsequent growth (such a finding is therefore evidence against the neoclassical theory which implies convergence). For example, if $\gamma=0.9$, convergence to within 90 per cent of final equilibrium occurs only in 22 periods, which, given quinquennial data, implies 110 years! Similarly, 0.7 requires 32 years, while 0.2 requires only 7 years.

The estimates of γ for the level model presented below using crosscountry quinquennial data are generally in excess of 0.7 no matter what econometric procedure is employed, but vary over a wide range depending on the method, 0.7 to 0.98. But for the differenced model, many estimates of γ are much smaller in the vicinity of 0.5. It is apparent that, for all practical purposes, coefficients in excess of 0.7 represent negligible convergence, since it would take more than a generation to achieve 90 per cent of equilibrium real per capita GDP.

Despite the fact that the FE model is unsatisfactory (because of the correlation problem), we shall use it as a reference point. The (10.2) model can be created from the meniu bar but it still simplier to run the line

```
panel Y 0 Y(-1) X
```

from the script window (note that we obtain $\gamma=0.72$):

```
Model 1: Fixed-effects, using 470 observations
Included 94 cross-sectional units
Time-series length = 5
Dependent variable: Y

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       2.12389      0.178932     11.87     8.50e-028 ***
  Y_1         0.720369     0.0236765    30.43     3.36e-103 ***
  X           0.165589     0.0193324     8.565    2.85e-016 ***

Log-likelihood       386.8309   Akaike criterion    -581.6618
```

The second (still unsatisfactory) model is that of RE (note that now $\gamma = 0.93$):

```
panel Y 0 Y(-1) X --random-effects

Model 2: Random-effects (GLS), using 470 observations

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       0.527612     0.0755693     6.982    1.01e-011 ***
  Y_1         0.932259     0.0104997    88.79     1.40e-294 ***
  X           0.138248     0.0122326    11.30     2.40e-026 ***

Log-likelihood       259.5798   Akaike criterion    -513.1595
```

Thus none could be characterized as evidence of reasonably rapid convergence.

Parameter estimates for the model in first differences can be obtained through the following lines:

```
diff Y X # take differences of all variables
tsls d_Y d_Y(-1) d_X ; 0 d_X Y(-2)
tsls d_Y d_Y(-1) d_X ; 0 d_X d_Y(-2)
```

The output is presented below.

```
tsls d_Y d_Y(-1) d_X ; 0 d_X Y(-2)  # use Y(-2) as instrument


Model 3: TSLS, using 376 observations
Dependent variable: d_Y
Instrumented: d_Y_1
Instruments: const d_X Y_2

             coefficient   std. error      z       p-value
  ---------------------------------------------------------
  d_Y_1       0.686966     0.0644896    10.65    1.70e-026 ***
  d_X         0.148424     0.0211451     7.019   2.23e-012 ***

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

? tsls d_Y d_Y(-1) d_X ; 0 d_X d_Y(-2)  # use d_Y(-2) as instrument
```

```
Model 4: TSLS, using 282 observations
Dependent variable: d_Y
Instrumented: d_Y_1
Instruments: const d_X d_Y_2

             coefficient   std. error      z      p-value
  ---------------------------------------------------------
  d_Y_1        0.584359     0.0717879    8.140   3.95e-016 ***
  d_X          0.134747     0.0218480    6.167   6.94e-010 ***
```

Now the (presumably, more correct) γs are lower thus implying more rapid convergence to equilibrium. In any case, any substantive study on growth rate convergence, should include additional explanatory variables such as, for example, the stock of human capital, infrastructure investment, and so forth.

Revision questions

10.1 aret
10.2 aertb
10.3 weqrb w
10.4 How to test that FE model is preferred to pooled?
10.5 Wė yw
10.6 Wt q

The Rules of Summation

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

$$\sum_{i=1}^{n} a = na$$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} (a + bx_i) = na + b\sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{3} f(x_i, y_j) &= \sum_{i=1}^{2} [f(x_i, y_1) + f(x_i, y_2) + f(x_i, y_3)] \\ &= f(x_1, y_1) + f(x_1, y_2) + f(x_1, y_3) \\ &\quad + f(x_2, y_1) + f(x_2, y_2) + f(x_2, y_3)\end{aligned}$$

Expected Values & Variances

$$\begin{aligned}E(X) &= x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n) \\ &= \sum_{i=1}^{n} x_i f(x_i) = \sum_{x} x f(x)\end{aligned}$$

$$E[g(X)] = \sum_{x} g(x) f(x)$$

$$\begin{aligned}E[g_1(X) + g_2(X)] &= \sum_{x} [g_1(x) + g_2(x)] f(x) \\ &= \sum_{x} g_1(x) f(x) + \sum_{x} g_2(x) f(x) \\ &= E[g_1(X)] + E[g_2(X)]\end{aligned}$$

$E(c) = c$
$E(cX) = cE(X)$
$E(a + cX) = a + cE(X)$
$\text{var}(X) = \sigma^2 = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$
$\text{var}(a + cX) = E[(a + cX) - E(a + cX)]^2 = c^2\text{var}(X)$

Marginal and Conditional Distributions

$$f(x) = \sum_{y} f(x, y) \quad \text{for each value } X \text{ can take}$$

$$f(y) = \sum_{x} f(x, y) \quad \text{for each value } Y \text{ can take}$$

$$f(x|y) = P[X = x | Y = y] = \frac{f(x, y)}{f(y)}$$

If X and Y are independent random variables, then $f(x,y) = f(x)f(y)$ for each and every pair of values x and y. The converse is also true.

If X and Y are independent random variables, then the conditional probability density function of X given that

$$Y = y \text{ is } f(x|y) = \frac{f(x, y)}{f(y)} = \frac{f(x)f(y)}{f(y)} = f(x)$$

for each and every pair of values x and y. The converse is also true.

Expectations, Variances & Covariances

$$\begin{aligned}\text{cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= \sum_{x}\sum_{y} [x - E(X)][y - E(Y)] f(x, y)\end{aligned}$$

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

$E(c_1X + c_2Y) = c_1E(X) + c_2E(Y)$
$E(X + Y) = E(X) + E(Y)$

$$\begin{aligned}\text{var}(aX + bY + cZ) &= a^2\text{var}(X) + b^2\text{var}(Y) + c^2\text{var}(Z) \\ &\quad + 2ab\text{cov}(X,Y) + 2ac\text{cov}(X,Z) + 2bc\text{cov}(Y,Z)\end{aligned}$$

If X, Y, and Z are independent, or uncorrelated, random variables, then the covariance terms are zero and:

$$\begin{aligned}\text{var}(aX + bY + cZ) &= a^2\text{var}(X) \\ &\quad + b^2\text{var}(Y) + c^2\text{var}(Z)\end{aligned}$$

Normal Probabilities

If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
If $X \sim N(\mu, \sigma^2)$ and a is a constant, then

$$P(X \geq a) = P\left(Z \geq \frac{a - \mu}{\sigma}\right)$$

If $X \sim N(\mu, \sigma^2)$ and a and b are constants, then

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)$$

Assumptions of the Simple Linear Regression Model

SR1	The value of y, for each value of x, is $y = \beta_1 + \beta_2 x + e$
SR2	The average value of the random error e is $E(e) = 0$ since we assume that $E(y) = \beta_1 + \beta_2 x$
SR3	The variance of the random error e is $\text{var}(e) = \sigma^2 = \text{var}(y)$
SR4	The covariance between any pair of random errors, e_i and e_j is $\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$
SR5	The variable x is not random and must take at least two different values.
SR6	(*optional*) The values of e are *normally distributed* about their mean $e \sim N(0, \sigma^2)$

Least Squares Estimation

If b_1 and b_2 are the least squares estimates, then

$$\hat{y}_i = b_1 + b_2 x_i$$
$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

The Normal Equations

$$Nb_1 + \Sigma x_i b_2 = \Sigma y_i$$
$$\Sigma x_i b_1 + \Sigma x_i^2 b_2 = \Sigma x_i y_i$$

Least Squares Estimators

$$b_2 = \frac{\Sigma (x_i - \overline{x})(y_i - \overline{y})}{\Sigma (x_i - \overline{x})^2}$$
$$b_1 = \overline{y} - b_2 \overline{x}$$

Elasticity

$$\eta = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x}\cdot\frac{x}{y}$$

$$\eta = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x}\cdot\frac{x}{E(y)} = \beta_2\cdot\frac{x}{E(y)}$$

Least Squares Expressions Useful for Theory

$$b_2 = \beta_2 + \Sigma w_i e_i$$

$$w_i = \frac{x_i - \overline{x}}{\Sigma(x_i - \overline{x})^2}$$

$$\Sigma w_i = 0, \quad \Sigma w_i x_i = 1, \quad \Sigma w_i^2 = 1/\Sigma(x_i - \overline{x})^2$$

Properties of the Least Squares Estimators

$$\text{var}(b_1) = \sigma^2\left[\frac{\Sigma x_i^2}{N\Sigma(x_i - \overline{x})^2}\right] \quad \text{var}(b_2) = \frac{\sigma^2}{\Sigma(x_i - \overline{x})^2}$$

$$\text{cov}(b_1, b_2) = \sigma^2\left[\frac{-\overline{x}}{\Sigma(x_i - \overline{x})^2}\right]$$

Gauss-Markov Theorem: Under the assumptions SR1–SR5 of the linear regression model the estimators b_1 and b_2 have the *smallest variance of all linear and unbiased estimators* of β_1 and β_2. They are the Best Linear Unbiased Estimators (BLUE) of β_1 and β_2.

If we make the normality assumption, assumption SR6, about the error term, then the least squares estimators are normally distributed.

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2\Sigma x_i^2}{N\Sigma(x_i - \overline{x})^2}\right), b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\Sigma(x_i - \overline{x})^2}\right)$$

Estimated Error Variance

$$\hat{\sigma}^2 = \frac{\Sigma\hat{e}_i^2}{N-2}$$

Estimator Standard Errors

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}(b_1)}}, \quad \text{se}(b_2) = \sqrt{\widehat{\text{var}(b_2)}}$$

t-distribution

If assumptions SR1–SR6 of the simple linear regression model hold, then

$$t = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-2)}, \quad k = 1, 2$$

Interval Estimates

$$P[b_2 - t_c\text{se}(b_2) \le \beta_2 \le b_2 + t_c\text{se}(b_2)] = 1 - \alpha$$

Hypothesis Testing

Components of Hypothesis Tests

1. A *null* hypothesis, H_0
2. An *alternative* hypothesis, H_1
3. A test *statistic*
4. A *rejection* region
5. A conclusion

If the null hypothesis $H_0 : \beta_2 = c$ is ***true***, then

$$t = \frac{b_2 - c}{\text{se}(b_2)} \sim t_{(N-2)}$$

Rejection rule for a two-tail test: If the value of the test statistic falls in the rejection region, either tail of the *t*-distribution, then we reject the null hypothesis and accept the alternative.

Type I error: The null hypothesis is *true* and we decide to *reject* it.

Type II error: The null hypothesis is *false* and we decide *not* to reject it.

***p*-value rejection rule:** When the *p*-value of a hypothesis test is *smaller* than the chosen value of α, then the test procedure leads to *rejection* of the null hypothesis.

Prediction

$$y_0 = \beta_1 + \beta_2 x_0 + e_0, \ \hat{y}_0 = b_1 + b_2 x_0, \ f = \hat{y}_0 - y_0$$

$$\widehat{\text{var}(f)} = \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}\right], \text{se}(f) = \sqrt{\widehat{\text{var}(f)}}$$

A $(1 - \alpha) \times 100\%$ confidence interval, or prediction interval, for y_0

$$\hat{y}_0 \pm t_c\text{se}(f)$$

Goodness of Fit

$$\Sigma(y_i - \overline{y})^2 = \Sigma(\hat{y}_i - \overline{y})^2 + \Sigma\hat{e}_i^2$$

$$SST = SSR + SSE$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = (\text{corr}(y, \hat{y}))^2$$

Log–Linear Model

$$\ln(y) = \beta_1 + \beta_2 x + e, \ \widehat{\ln(y)} = b_1 + b_2 x$$

$100 \times \beta_2 \approx \%$ change in y given a one-unit change in x.

$$\hat{y}_n = \exp(b_1 + b_2 x)$$

$$\hat{y}_c = \exp(b_1 + b_2 x)\exp(\hat{\sigma}^2/2)$$

Prediction interval:

$$\exp\left[\widehat{\ln(y)} - t_c\text{se}(f)\right], \ \exp\left[\widehat{\ln(y)} + t_c\text{se}(f)\right]$$

Generalized goodness-of-fit measure $R_g^2 = (\text{corr}(y, \hat{y}_n))^2$

Assumptions of the Multiple Regression Model

MR1 $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$

MR2 $E(y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \Leftrightarrow E(e_i) = 0.$

MR3 $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$

MR4 $\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0$

MR5 The values of x_{ik} are not random and are not exact linear functions of the other explanatory variables.

MR6 $y_i \sim N[(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}), \sigma^2]$
$\Leftrightarrow e_i \sim N(0, \sigma^2)$

Least Squares Estimates in MR Model

Least squares estimates $b_1, b_2, \ldots, b_K$ minimize $S(\beta_1, \beta_2, \ldots, \beta_K) = \Sigma(y_i - \beta_1 - \beta_2 x_{i2} - \cdots - \beta_K x_{iK})^2$

Estimated Error Variance and Estimator Standard Errors

$$\hat{\sigma}^2 = \frac{\Sigma\hat{e}_i^2}{N - K} \quad \text{se}(b_k) = \sqrt{\widehat{\text{var}(b_k)}}$$

Hypothesis Tests and Interval Estimates for Single Parameters

Use t-distribution $\quad t = \dfrac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-K)}$

t-test for More than One Parameter

$$H_0 : \beta_2 + c\beta_3 = a$$

When H_0 is true $\quad t = \dfrac{b_2 + cb_3 - a}{\text{se}(b_2 + cb_3)} \sim t_{(N-K)}$

$$\text{se}(b_2 + cb_3) = \sqrt{\widehat{\text{var}(b_2)} + c^2\widehat{\text{var}(b_3)} + 2c \times \widehat{\text{cov}(b_2, b_3)}}$$

Joint F-tests

To test J joint hypotheses,

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

To test the overall significance of the model the null and alternative hypotheses and F statistic are

$$H_0 : \beta_2 = 0,\ \beta_3 = 0, \ldots, \beta_K = 0$$
$$H_1 : \textit{at least one} \text{ of the } \beta_k \text{ is nonzero}$$
$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)}$$

RESET: A Specification Test

$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \qquad \hat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$

$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 \hat{y}_i^2 + e_i, \quad H_0 : \gamma_1 = 0$

$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + e_i, \quad H_0 : \gamma_1 = \gamma_2 = 0$

Model Selection

$\text{AIC} = \ln(SSE/N) + 2K/N$

$\text{SC} = \ln(SSE/N) + K\ln(N)/N$

Collinearity and Omitted Variables

$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$

$$\text{var}(b_2) = \frac{\sigma^2}{(1 - r_{23}^2)\sum (x_{i2} - \bar{x}_2)^2}$$

When x_3 is omitted, $\text{bias}(b_2^*) = E(b_2^*) - \beta_2 = \beta_3 \dfrac{\widehat{\text{cov}(x_2, x_3)}}{\widehat{\text{var}(x_2)}}$

Heteroskedasticity

$$\text{var}(y_i) = \text{var}(e_i) = \sigma_i^2$$

General variance function

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

Breusch-Pagan and White Tests for H_0: $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$

When H_0 is true $\quad \chi^2 = N \times R^2 \sim \chi^2_{(S-1)}$

Goldfeld-Quandt test for $H_0 : \sigma_M^2 = \sigma_R^2$ versus $H_1 : \sigma_M^2 \neq \sigma_R^2$

When H_0 is true $\quad F = \hat{\sigma}_M^2/\hat{\sigma}_R^2 \sim F_{(N_M - K_M, N_R - K_R)}$

Transformed model for $\text{var}(e_i) = \sigma_i^2 = \sigma^2 x_i$

$$y_i/\sqrt{x_i} = \beta_1(1/\sqrt{x_i}) + \beta_2(x_i/\sqrt{x_i}) + e_i/\sqrt{x_i}$$

Estimating the variance function

$$\ln(\hat{e}_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$$

Grouped data

$$\text{var}(e_i) = \sigma_i^2 = \begin{cases} \sigma_M^2 & i = 1, 2, \ldots, N_M \\ \sigma_R^2 & i = 1, 2, \ldots, N_R \end{cases}$$

Transformed model for feasible generalized least squares

$$y_i \Big/ \sqrt{\hat{\sigma}_i} = \beta_1 \left(1 \Big/ \sqrt{\hat{\sigma}_i}\right) + \beta_2 \left(x_i \Big/ \sqrt{\hat{\sigma}_i}\right) + e_i \Big/ \sqrt{\hat{\sigma}_i}$$

Regression with Stationary Time Series Variables

Finite distributed lag model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_q x_{t-q} + v_t$$

Correlogram

$$r_k = \sum (y_t - \bar{y})(y_{t-k} - \bar{y}) / \sum (y_t - \bar{y})^2$$

For $H_0 : \rho_k = 0, \quad z = \sqrt{T} r_k \sim N(0,1)$

LM test

$y_t = \beta_1 + \beta_2 x_t + \rho \hat{e}_{t-1} + \hat{v}_t \quad$ Test $H_0 : \rho = 0$ with t-test

$\hat{e}_t = \gamma_1 + \gamma_2 x_t + \rho \hat{e}_{t-1} + \hat{v}_t \quad$ Test using $LM = T \times R^2$

AR(1) error $\quad y_t = \beta_1 + \beta_2 x_t + e_t \quad e_t = \rho e_{t-1} + v_t$

Nonlinear least squares estimation

$$y_t = \beta_1(1 - \rho) + \beta_2 x_t + \rho y_{t-1} - \beta_2 \rho x_{t-1} + v_t$$

ARDL(p, q) model

$$y_t = \delta + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + v_t$$

AR(p) forecasting model

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + v_t$$

Exponential smoothing $\quad \hat{y}_t = \alpha y_{t-1} + (1 - \alpha)\hat{y}_{t-1}$

Multiplier analysis

$$\delta_0 + \delta_1 L + \delta_2 L^2 + \cdots + \delta_q L^q = (1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_p L^p) \times (\beta_0 + \beta_1 L + \beta_2 L^2 + \cdots)$$

Unit Roots and Cointegration

Unit Root Test for Stationarity: Null hypothesis:

$$H_0 : \gamma = 0$$

Dickey-Fuller Test 1 (no constant and no trend):

$$\Delta y_t = \gamma y_{t-1} + v_t$$

Dickey-Fuller Test 2 (with constant but no trend):

$$\Delta y_t = \alpha + \gamma y_{t-1} + v_t$$

Dickey-Fuller Test 3 (with constant and with trend):

$$\Delta y_t = \alpha + \gamma y_{t-1} + \lambda t + v_t$$

Augmented Dickey-Fuller Tests:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{s=1}^{m} a_s \Delta y_{t-s} + v_t$$

Test for cointegration

$$\Delta \hat{e}_t = \gamma \hat{e}_{t-1} + v_t$$

Random walk: $\quad y_t = y_{t-1} + v_t$

Random walk with drift: $\quad y_t = \alpha + y_{t-1} + v_t$

Random walk model with drift and time trend:

$$y_t = \alpha + \delta t + y_{t-1} + v_t$$

Panel Data

Pooled least squares regression

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it}$$

Cluster robust standard errors $\quad \text{cov}(e_{it}, e_{is}) = \psi_{ts}$

Fixed effects model

$y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it} \qquad \beta_{1i}$ not random

$y_{it} - \bar{y}_i = \beta_2(x_{2it} - \bar{x}_{2i}) + \beta_3(x_{3it} - \bar{x}_{3i}) + (e_{it} - \bar{e}_i)$

Random effects model

$y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it} \qquad \beta_{it} = \overline{\beta}_1 + u_i$ random

$y_{it} - \alpha\bar{y}_i = \overline{\beta}_1(1 - \alpha) + \beta_2(x_{2it} - \alpha\bar{x}_{2i}) + \beta_3(x_{3it} - \alpha\bar{x}_{3i}) + v_{it}^*$

$$\alpha = 1 - \sigma_e \Big/ \sqrt{T\sigma_u^2 + \sigma_e^2}$$

Hausman test

$$t = (b_{FE,k} - b_{RE,k}) \Big/ \left[\widehat{\text{var}(b_{FE,k})} - \widehat{\text{var}(b_{RE,k})}\right]^{1/2}$$

REFERENCES

[CK] Commandeur J.J.F., Koopman S.J. - An Introduction to State Space Time Series Analysis, Oxford University Press, 2007.

[Ham] Hamilton J.D. - Time Series Analysis, Princeton University Press,1994

[HBFKD] Heij, de Boer, Franses, Kloek, van Dijk - Econometric Methods with Applications in Business and Economics, Oxford University Press, 2004.

[Har] Harvey A.C. - Time Series Models, MIT Press, 1993

[L] Ludvigson S., https://files.nyu.edu/bpn207/public/Teaching/2008/EconometricsII-Part1/handout7.pdf

[MWH] Makridakis S., Wheelwright S.C., Hyndman R.J. - Forecasting – Methods and Applications, 3ed., Wiley, 1998