Basic Financial Econometrics

Alois Geyer
WU Vienna University of Economics and Business
alois.geyer@wu.ac.at
http://www.wu.ac.at/~geyer

this version:
June 24, 2021

preliminary and incomplete

Contents

1 Financial Regression Analysis **1**

1.1 Regression analysis . . . 1
1.1.1 Least squares estimation . . . 2
1.1.2 Implications . . . 3
1.1.3 Interpretation . . . 4
1.2 Finite sample properties of least squares estimates . . . 6
1.2.1 Assumptions . . . 8
1.2.2 Properties . . . 11
1.2.3 Testing hypothesis . . . 13
1.2.4 Example 6: CAPM, beta-factors and multi-factor models . . . 15
1.2.5 Example 7: Interest rate parity . . . 19
1.2.6 Prediction . . . 21
1.3 Large sample properties of least squares estimates . . . 22
1.3.1 Consistency . . . 23
1.3.2 Asymptotic normality . . . 25
1.3.3 Time series data . . . 26
1.4 Maximum likelihood estimation . . . 28
1.5 LM, LR and Wald tests . . . 31
1.6 Specifications . . . 33
1.6.1 Log and other transformations . . . 33
1.6.2 Dummy variables . . . 34
1.6.3 Interactions . . . 35
1.6.4 Difference-in-differences . . . 36
1.6.5 Example 11: Hedonic price functions . . . 37
1.6.6 Example 12: House price changes induced by siting decisions . . . 38
1.6.7 Omitted and irrelevant regressors . . . 39
1.6.8 Selection of regressors . . . 41
1.7 Regression diagnostics . . . 43
1.7.1 Non-normality . . . 43
1.7.2 Heteroscedasticity . . . 44
1.7.3 Autocorrelation . . . 46
1.8 Generalized least squares . . . 51
1.8.1 Heteroscedasticity . . . 51
1.8.2 Autocorrelation . . . 52
1.8.3 Example 19: Long-horizon return regressions . . . 55
1.9 Endogeneity and instrumental variable estimation . . . 57
1.9.1 Endogeneity . . . 57
1.9.2 Instrumental variable estimation . . . 59

1.9.3 Selection of instruments and tests . . . 62
1.9.4 Example 21: Consumption based asset pricing . . . 65
1.10 Generalized method of moments . . . 69
1.10.1 OLS, IV and GMM . . . 71
1.10.2 Asset pricing and GMM . . . 72
1.10.3 Estimation and inference . . . 74
1.10.4 Example 24: Models for the short-term interest rate . . . 77
1.11 Models with binary dependent variables . . . 78
1.12 Sample selection . . . 82
1.13 Duration models . . . 84

2 Time Series Analysis **87**
2.1 Financial time series . . . 87
2.1.1 Descriptive statistics of returns . . . 88
2.1.2 Return distributions . . . 91
2.1.3 Abnormal returns and event studies . . . 94
2.1.4 Autocorrelation analysis of financial returns . . . 97
2.1.5 Stochastic process terminology . . . 100
2.2 ARMA models . . . 101
2.2.1 AR models . . . 101
2.2.2 MA models . . . 104
2.2.3 ARMA models . . . 105
2.2.4 Estimating ARMA models . . . 106
2.2.5 Diagnostic checking of ARMA models . . . 107
2.2.6 Example 35: ARMA models for FTSE and AMEX returns . . . 108
2.2.7 Forecasting with ARMA models . . . 110
2.2.8 Properties of ARMA forecast errors . . . 112
2.3 Non-stationary models . . . 115
2.3.1 Random-walk and ARIMA models . . . 115
2.3.2 Forecasting prices from returns . . . 118
2.3.3 Unit-root tests . . . 119
2.4 Diffusion models in discrete time . . . 124
2.4.1 Discrete time approximation . . . 126
2.4.2 Estimating parameters . . . 126
2.4.3 Probability statements about future prices . . . 129
2.5 GARCH models . . . 131
2.5.1 Estimating and diagnostic checking of GARCH models . . . 133
2.5.2 Example 49: ARMA-GARCH models for IBM and FTSE returns . . . 133
2.5.3 Forecasting with GARCH models . . . 135
2.5.4 Special GARCH models . . . 136

3 Vector time series models **138**
3.1 Vector-autoregressive models . . . 138
3.1.1 Formulation of VAR models . . . 138
3.1.2 Estimating and forecasting VAR models . . . 140
3.2 Cointegration and error correction models . . . 143
3.2.1 Cointegration . . . 143
3.2.2 Error correction model . . . 143
3.2.3 Example 53: The expectation hypothesis of the term structure . . . 145
3.2.4 The Engle-Granger procedure . . . 146
3.2.5 The Johansen procedure . . . 150
3.2.6 Cointegration among more than two series . . . 155
3.3 State space modeling and the Kalman filter . . . 157
3.3.1 The state space formulation . . . 157
3.3.2 The Kalman filter . . . 158
3.3.3 Example 60: The Cox-Ingersoll-Ross model of the term structure . . . 159

Bibliography **162**

I am grateful to many PhD students of the VGSF program, as well as doctoral and master students at WU for valuable comments which have helped to improve these lecture notes.

1 Financial Regression Analysis

1.1 Regression analysis

We start by reviewing key aspects of regression analysis. Its purpose is to relate a **dependent variable** $\boldsymbol{y}$ to one or more variables $\boldsymbol{X}$ which are assumed to affect $\boldsymbol{y}$. The relation is specified in terms of a systematic part which determines the **expected value** of $\boldsymbol{y}$ and a random part $\boldsymbol{\epsilon}$. For example, the systematic part could be a (theoretically derived) valuation relationship. The random part represents unsystematic deviations between observations and expectations (e.g. deviations from equilibrium). The relation between $\boldsymbol{y}$ and $\boldsymbol{X}$ depends on *unknown* parameters $\boldsymbol{\beta}$ which are used in the function that relates $\boldsymbol{X}$ to the expectation of $\boldsymbol{y}$.

Assumption AL (linearity): We consider the *linear* regression equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

$\boldsymbol{y}$ is the $n\times 1$ vector $(y_1, \ldots, y_n)'$ of observations of the dependent (or **endogenous**) variable, $\boldsymbol{\epsilon}$ is the vector of **errors** (also called **residuals**, **disturbances**, **innovations** or **shocks**), $\boldsymbol{\beta}$ is the $K\times 1$ vector of parameters, and the $n\times K$ matrix $\boldsymbol{X}$ of **regressors** (also called **explanatory variables** or **covariates**) is defined as follows:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}.$$

k is the number of regressors and $K{=}k{+}1$ is the dimension of $\boldsymbol{\beta}{=}(\beta_0, \beta_1, \ldots, \beta_k)'$, where β_0 is the **constant term** or **intercept**. A single row i of $\boldsymbol{X}$ will be denoted by the $K\times 1$ *column* vector $\boldsymbol{x}_i$. For a *single* observation the model equation is written as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \qquad (i = 1, \ldots, n).$$

We will frequently (mainly in the context of model specification and interpretation) use formulations like

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon,$$

where the symbols y, x_i and ϵ *represent* the variables in question. It is understood that such equations also hold for a single observation.

1.1.1 Least squares estimation

A main purpose of regression analysis is to draw conclusions about the population using a sample. The regression equation $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon}$ is assumed to hold in the population. The sample estimate of $\boldsymbol{\beta}$ is denoted by $\boldsymbol{b}$ and the estimate of $\boldsymbol{\epsilon}$ by $\boldsymbol{e}$. According to the least squares (LS) criterion, $\boldsymbol{b}$ should be chosen such that the sum of squared errors SSE is minimized

$$\mathrm{SSE}(\boldsymbol{b}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{b})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \longrightarrow \min .$$

A necessary condition for a minimum is derived from

$$\mathrm{SSE}(\boldsymbol{b}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b},$$

and is given by

$$\frac{\partial \mathrm{SSE}(\boldsymbol{b})}{\partial \boldsymbol{b}} = \boldsymbol{0} : \quad -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}.$$

Assumption AR (rank): We assume that $\boldsymbol{X}$ has full rank equal to K (i.e. the columns of $\boldsymbol{X}$ are linearly independent). If $\boldsymbol{X}$ has full rank, $\boldsymbol{X}'\boldsymbol{X}$ is positive definite and the **ordinary least squares (OLS)** estimates $\boldsymbol{b}$ are given by

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1}$$

The solution is a minimum since

$$\frac{\partial^2 \mathrm{SSE}(\boldsymbol{b})}{\partial \boldsymbol{b}^2} = 2\boldsymbol{X}'\boldsymbol{X}$$

is positive definite by assumption **AR**.
It is useful to express $\boldsymbol{X}'\boldsymbol{y}$ and $\boldsymbol{X}'\boldsymbol{X}$ in terms of the sums

$$\boldsymbol{X}'\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i \qquad \boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$$

to point out that the estimate is related to the covariance between the dependent variable and the regressors, and the covariance among regressors. In the special case of the simple regression model $y=b_0+b_1x+e$ with a single regressor the estimates b_1 and b_0 are given by

$$b_1 = \frac{s_{yx}}{s_x^2} = r_{yx}\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x},$$

where s_{yx} (r_{yx}) is the sample covariance (correlation) between y and x, s_y and s_x are the sample standard deviations of y and x, and $\bar{y}$ and $\bar{x}$ are their sample means.

1.1.2 Implications

By the first order condition the OLS estimates satisfy the **normal equation**

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} - \boldsymbol{X}'\boldsymbol{y} = -\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = -\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}, \tag{2}$$

which implies that each column of $\boldsymbol{X}$ is uncorrelated with (orthogonal to) $\boldsymbol{e}$.
If the first column of $\boldsymbol{X}$ is a column of ones denoted by $\boldsymbol{\iota}$, LS estimation has the following implications:

1. The residuals have zero mean since $\boldsymbol{\iota}'\boldsymbol{e}=0$ (from the normal equation).
2. This implies that the mean of the **fitted values** $\hat{y}_i=\boldsymbol{x}_i'\boldsymbol{b}$ is equal to the sample mean:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}.$$

3. The fitted values are equal to the mean of $\boldsymbol{y}$ if the regression equation is evaluated for the means of $\boldsymbol{X}$:

$$\bar{y} = b_0 + \sum_{j=1}^{k}\bar{x}_j b_j.$$

4. The fitted values and the residuals are orthogonal:

$$\hat{\boldsymbol{y}}'\boldsymbol{e} = \boldsymbol{0}.$$

5. The slope in a regression of $\boldsymbol{y}$ on $\boldsymbol{e}$ is always equal to one and the constant is equal to $\bar{y}$.[1]

The goodness of fit of a regression model can be measured by the **coefficient of determination** R^2 defined as

$$R^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{(\boldsymbol{y}-\bar{y})'(\boldsymbol{y}-\bar{y})} = 1 - \frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})}{(\boldsymbol{y}-\bar{y})'(\boldsymbol{y}-\bar{y})} = \frac{(\hat{\boldsymbol{y}}-\bar{y})'(\hat{\boldsymbol{y}}-\bar{y})}{(\boldsymbol{y}-\bar{y})'(\boldsymbol{y}-\bar{y})}.$$

This is the so-called *centered version* of R^2 which lies between 0 and 1 if the model contains an intercept. It is equal to the squared correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$. The three terms in the expression

$$(\boldsymbol{y}-\bar{y})'(\boldsymbol{y}-\bar{y}) = (\hat{\boldsymbol{y}}-\bar{y})'(\hat{\boldsymbol{y}}-\bar{y}) + (\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})$$

are called the total sum of squares (SST), the sum of squares from the regression (SSR), and the sum of squared errors (SSE). Based on this relation R^2 is frequently interpreted as

[1] By implication 3 the constant must be equal to $\bar{y}$ since the mean of $\boldsymbol{e}$ is zero. The slope is given by $(\boldsymbol{e}'\boldsymbol{e})^{-1}\boldsymbol{e}'\tilde{\boldsymbol{y}}$, where $\tilde{\boldsymbol{y}}=\boldsymbol{y}-\bar{y}$. The slope is equal to one since $\boldsymbol{e}'\tilde{\boldsymbol{y}}=\boldsymbol{e}'\boldsymbol{e}$. The latter identity holds since in the original regression $\boldsymbol{e}'\boldsymbol{y}=\boldsymbol{e}'\boldsymbol{X}\boldsymbol{b}+\boldsymbol{e}'\boldsymbol{e}$ and $\boldsymbol{e}'\boldsymbol{X}=\boldsymbol{0}'$. Finally, $\boldsymbol{e}'\boldsymbol{y}=\boldsymbol{e}'\tilde{\boldsymbol{y}}$ since $\boldsymbol{e}'\bar{\boldsymbol{y}}=0$.

the percentage of y's variance 'explained' by the regression. If the model does not contain an intercept, the centered R^2 may become negative. In that case the *uncentered* R^2 can be used:

$$\text{uncentered } R^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}}.$$

R^2 is zero if all regression coefficients except for the constant are zero ($\boldsymbol{b}$=(b_0 $\mathbf{0}$)$'$ and $\hat{y}$=b_0=$\bar{y}$). In this case the regression is a horizontal line. If R^2=1 all observations are located on the regression line (or hyperplane) (i.e. $\hat{y}_i$=y_i). R^2 is (only) a measure for the goodness of the *linear approximation* implied by the regression. Many other, more relevant aspects of a model's quality, are not taken into account by R^2. Such aspects will become more apparent as we proceed.

1.1.3 Interpretation

The coefficients $\boldsymbol{b}$ can be interpreted on the basis of the fitted values[2]

$$\hat{y} = b_0 + x_1 b_1 + \cdots + x_k b_k.$$

b_j is the change in $\hat{y}$ (or, the *expected* change in y) if x_j changes by one unit ceteris paribus (c.p.), i.e. holding *all other* regressors fixed. In general the change in the expected value is

$$\Delta\hat{y} = \Delta x_1 b_1 + \cdots + \Delta x_k b_k,$$

which implies that the effects of simultaneously changing several regressors can be added up.

This interpretation is based on the **Frisch-Waugh theorem**. Suppose we partition the regressors in two groups $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ to save the residuals $\boldsymbol{e}_1$. Next we regress each column of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$ and save the residuals of these regressions in the matrix $\boldsymbol{E}_2$. According to the Frisch-Waugh theorem the coefficients from the regression of $\boldsymbol{e}_1$ on $\boldsymbol{E}_2$ are *equal* to the subset of coefficients from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ that corresponds to $\boldsymbol{X}_2$. In more general terms, the theorem implies that *partial* effects can be obtained *directly* from a multiple regression. It is not necessary to first construct orthogonal variables.

To illustrate the theorem we consider the regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + e.$$

To obtain the coefficient of x_2 such that the effect of x_1 (and the intercept) is held constant, we first run the two simple regressions

$$y = c_y + b_{y1} x_1 + e_{y1} \qquad x_2 = c_{x2} + b_{21} x_1 + e_{21}.$$

e_{y1} and e_{21} represent those parts of y and x_2 which do not depend on x_1. Subsequently, we run a regression using these residuals to obtain the coefficient b_2:

$$(y - c_y - b_{y1} x_1) = b_2(x_2 - c_{x2} - b_{21} x_1) + u \qquad e_{y1} = b_2 e_{21} + u.$$

[2] An analogous interpretation holds for $\boldsymbol{\beta}$ in the population.

In general, this procedure is also referred to as 'controlling for' or 'partialling out' the effect of $\boldsymbol{X}_1$. Simply speaking, if we want to isolate the effects of $\boldsymbol{X}_2$ on $\boldsymbol{y}$ we have to 'remove' the effects of $\boldsymbol{X}_1$ from the entire regression equation.[3] However, according to the Frisch-Waugh theorem it is not necessary to run this sequence of regressions in practice. Running a (multiple) regression of $\boldsymbol{y}$ on all regressors $\boldsymbol{X}$ 'automatically' controls for the effects of each regressor on all other regressors. A special case is an orthogonal regression, where all regressors are uncorrelated (i.e. $\boldsymbol{X}'\boldsymbol{X}$ is a diagonal matrix). In this case the coefficients from the multiple regression are identical to those obtained from K simple regressions using one column of $\boldsymbol{X}$ at a time.

Example 1: We use the real investment data from Table 3.1 in Greene (2003) to estimate a multiple regression model. The dependent variable is real investment (in trillion US$; denoted by y). The explanatory variables are real GNP (in trillion US$; g), the (nominal) interest rate r and the inflation rate i (both measured as percentages). The (rounded) estimated coefficients are

$$\boldsymbol{b} = (-0.0726 \;\; 0.236 \;\; -0.00356 \;\; -0.000276)',$$

where the first element is the constant term. The coefficient -0.00356 can be interpreted as follows: if the interest rate goes up by one percentage point and the other regressors do not change, real investment is expected to drop by about 3.56 billion US$. SST=0.0164, SSR=0.0127 and SSE=0.00364. The corresponding R^2 equals 0.78 (SSR/SST), which means that about 78% of the variance in real investment can be explained by the regressors. Further details can be found in the file `investment.xls`.

Exercise 1: Use the quarterly data in Table F5.1 from Greene's website http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm (see file `Table F5.1.xls`) to estimate a regression of real investment (realinvs) on a constant, real GDP, the nominal interest rate (tbilrate; 90 day treasury bill rate) and the inflation rate (infl). Check the validity of the five OLS implications mentioned on p.3.

Apply the Frisch-Waugh theorem and show how the coefficients of the constant term and tbilrate can be obtained by controlling for the effects of the nominal interest rate and inflation.

[3] As a matter of fact, the effects of $\boldsymbol{X}_1$, or any other set of regressors we want to control for, *need not* be removed from $\boldsymbol{y}$. It can be shown that the coefficients associated with $\boldsymbol{X}_2$ can also be obtained from a regression of $\boldsymbol{y}$ on $\boldsymbol{E}_2$. Because of implication 5 the covariance between $\boldsymbol{e}_1$ and the columns of $\boldsymbol{E}_2$ is identical to the covariance between $\boldsymbol{y}$ and $\boldsymbol{E}_2$.

1.2 Finite sample properties of least squares estimates[4]

Review 1: For any constants a and b and random variables Y and X the following relations hold:

$$\mathrm{E}[a+Y]=a+\mathrm{E}[Y] \quad \mathrm{E}[aY]=a\mathrm{E}[Y] \quad \mathrm{V}[a+Y]=\mathrm{V}[Y] \quad \mathrm{V}[aY]=a^2\mathrm{V}[Y].$$

$$\mathrm{E}[aX+bY]=a\mathrm{E}[X]+b\mathrm{E}[Y] \quad \mathrm{V}[aX+bY]=a^2\mathrm{V}[X]+b^2\mathrm{V}[Y]+2ab\mathrm{cov}[XY].$$

Jensen's inequality: $\mathrm{E}[f(X)]\geq f(\mathrm{E}[X])$ for any convex function $f(X)$.

For a constant a and random variables W, X, Y, Z the following relations hold:

$$\text{if } Y=aZ: \quad \mathrm{cov}[X,Y]=a\mathrm{cov}[X,Z].$$

$$\text{if } Y=W+Z: \quad \mathrm{cov}[X,Y]=\mathrm{cov}[X,W]+\mathrm{cov}[X,Z].$$

$$\mathrm{cov}[X,Y]=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y] \qquad \mathrm{cov}[Y,a]=0.$$

If $\boldsymbol{X}$ is a $n\times 1$ vector of random variables $\mathrm{V}[\boldsymbol{X}]=\mathrm{cov}[\boldsymbol{X}]=\boldsymbol{\Sigma}=\mathrm{E}[(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])']$ is a $n\times n$ matrix. Its diagonal elements are the variances of the elements of $\boldsymbol{X}$. Using $\boldsymbol{\mu}=\mathrm{E}[\boldsymbol{X}]$ we can write $\boldsymbol{\Sigma}=\mathrm{E}[\boldsymbol{X}\boldsymbol{X}']-\boldsymbol{\mu}\boldsymbol{\mu}'$.

If $\boldsymbol{b}$ is a $n\times 1$ vector and $\boldsymbol{A}$ is a $n\times n$ matrix of constants, the following relations hold:

$$\mathrm{E}[\boldsymbol{b}'\boldsymbol{X}]=\boldsymbol{b}'\boldsymbol{\mu} \quad \mathrm{V}[\boldsymbol{b}'\boldsymbol{X}]=\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b} \quad \mathrm{E}[\boldsymbol{A}\boldsymbol{X}]=\boldsymbol{A}\boldsymbol{\mu} \quad \mathrm{V}[\boldsymbol{A}\boldsymbol{X}]=\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'.$$

Review 2: The conditional and unconditional moments of two random variables Y and X are related as follows:

Law of iterated expectations:[5] $\mathrm{E}[Y]=\mathrm{E}_{\mathrm{x}}[\mathrm{E}[Y|X]]$

Functions of the conditioning variable:[6] $\mathrm{E}[f(X)Y|X]=f(X)\mathrm{E}[Y|X]$

If $\mathrm{E}[Y|X]$ is a linear function of X: $\mathrm{E}[Y|X]=\mathrm{E}[Y]+\dfrac{\mathrm{cov}[Y,X]}{\mathrm{V}[X]}(X-\mathrm{E}[X])$

Variance decomposition: $\mathrm{V}[Y]=\mathrm{E}_{\mathrm{x}}[\mathrm{V}[Y|X]]+\mathrm{V}_{\mathrm{x}}[\mathrm{E}[Y|X]]$

Conditional variance: $\mathrm{V}[Y|X]=\mathrm{E}[(Y-\mathrm{E}[Y|X])^2|X]=\mathrm{E}[Y^2|X]-(\mathrm{E}[Y|X])^2$.

Review 3:[7] A set of n observations y_i (i=1,...,n) of a random variable Y is a **random sample** if the observations are drawn *independently* from the *same* population with probability density $f(y_i,\boldsymbol{\theta})$. A random sample is said to be **independent**, **identically distributed** (**i.i.d.**) which is denoted by $y_i\sim$i.i.d.

A **cross section** is a sample of several units (e.g. firms or households) observed at a specific point in time (or time interval). A **time series** is a chronologically ordered sequence of data usually observed at regular time intervals (e.g. days or months). **Panel data** is constructed by stacking time series of several cross sections (e.g. monthly consumption and income of several households).

We consider a parameter θ and its **estimator** $\hat{\theta}$ derived from a random sample of size n. Estimators are rules for calculating estimates from a sample. For simplicity $\hat{\theta}$ both

[4] Most of this section is based on Greene (2003), sections 2.3 and 4.3 to 4.7, and Hayashi (2000), sections 1.1 and 1.3.

[5] $\mathrm{E}[Y|X]$ is a function of X. The notation E_{x} indicates expectation over values of X.

[6] See equation 7-60 in Papoulis (1984, p.165).

[7] Greene (2003); sections C.1 to C.5.

denotes the estimated value from a specific sample and the estimator (the function used to derive the estimate). $\hat{\theta}$ is a random variable since it depends on the (random) sample. The **sampling distribution** describes the probability distribution of $\hat{\theta}$ across possible samples.

Unbiasedness: $\hat{\theta}$ is **unbiased**, if $\mathrm{E}[\hat{\theta}]=\theta$. The expectation is formed with respect to the sampling distribution of $\hat{\theta}$. The **bias** is $\mathrm{E}[\hat{\theta}]-\theta$.

Examples: The sample mean and the sample median are unbiased estimators. The unadjusted sample variance

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

is a biased estimator of σ^2, whereas $s^2=n\tilde{s}^2/(n-1)$ is unbiased.

Mean squared error: The mean squared error (MSE) of $\hat{\theta}$ is the sum of the variance and the squared bias:

$$\mathrm{MSE}[\hat{\theta}] = \mathrm{E}[(\hat{\theta}-\theta)^2] = \mathrm{V}[\hat{\theta}] + (\mathrm{E}[(\hat{\theta}-\theta)])^2.$$

Example: The MSE of the unbiased estimator s^2 is larger than the MSE of $\tilde{s}^2$.

Efficiency: $\hat{\theta}$ is **efficient** if it is unbiased, and its sampling variance is lower than the variance of any other[8] to another estimator unbiased estimator $\hat{\theta}'$:

$$\mathrm{V}[\hat{\theta}] < \mathrm{V}[\hat{\theta}'].$$

[8] If the condition holds for *another* estimator one could use the term 'relative efficiency'.

1.2.1 Assumptions

The sample estimates $\boldsymbol{b}$ and $\boldsymbol{e}$ can be used to draw conclusions about the population. An important question relates to the finite sample properties of the OLS estimates. Exact (or finite sample) inference as opposed to asymptotic (large sample) inference is valid for any sample size n and is based on further assumptions (in addition to **AL** and **AR**) mentioned and discussed below.

To derive the finite sample properties of the OLS estimate we rewrite $\boldsymbol{b}$ in (1) as follows:

$$\begin{aligned} \boldsymbol{b} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} = \boldsymbol{\beta}+\boldsymbol{H}\boldsymbol{\epsilon}. \end{aligned} \tag{3}$$

We consider the statistical properties of $\boldsymbol{b}$ (in particular $\mathrm{E}[\boldsymbol{b}]$, $\mathrm{V}[\boldsymbol{b}]$, and its distribution). This is equivalent to investigate the **sampling error** $\boldsymbol{b}-\boldsymbol{\beta}$. From (see Review 2)

$$\mathrm{E}[\boldsymbol{b}] = \boldsymbol{\beta}+\mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\right] = \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\mathrm{E}\left[\boldsymbol{X}'\boldsymbol{\epsilon}\right] \tag{4}$$

we see that the properties of $\boldsymbol{b}$ depend on the properties of $\boldsymbol{X}$, $\boldsymbol{\epsilon}$, and their relation. In the so-called **classical regression model**, $\boldsymbol{X}$ is assumed to be non-stochastic. This means that $\boldsymbol{X}$ can be chosen (like in an experimental situation), or is fixed in repeated samples. Neither case holds in typical financial empirical studies. We will treat $\boldsymbol{X}$ as random, and the finite sample properties derived below are considered to be *conditional* on the sample $\boldsymbol{X}$ (although we will not always indicate this explicitly). This does not preclude the possibility that $\boldsymbol{X}$ contains constants (e.g. dummy variables). The important requirement (assumption) is that $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are generated by mechanisms that are completely unrelated.

Assumption AX (strict exogeneity): The conditional expectation of *each* ϵ_i conditional on *all* observations and variables in $\boldsymbol{X}$ is zero:

$$\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0} \qquad \mathrm{E}[\epsilon_i|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n] = 0 \qquad (i=1,\ldots,n).$$

According to this assumption, $\boldsymbol{X}$ cannot be used to obtain information about $\boldsymbol{\epsilon}$. If **AX** is satisfied, the following properties hold:

1. (unconditional mean): $\mathrm{E}[\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]]=\mathrm{E}[\boldsymbol{\epsilon}]=\boldsymbol{0}$.
2. (conditional expectation): $\mathrm{E}[\boldsymbol{y}|\boldsymbol{X}]=\hat{\boldsymbol{y}}=\boldsymbol{X}\boldsymbol{\beta}$.
3. Regressors and disturbances are **orthogonal**

$$\mathrm{E}[x_{il}\epsilon_j] = 0 \quad (i,j=1,\ldots,n;\ l=1,\ldots,K),$$

since $\mathrm{E}[x_{il}\epsilon_j]=\mathrm{E}[\mathrm{E}[x_{il}\epsilon_j|x_{il}]]=\mathrm{E}[x_{il}\mathrm{E}[\epsilon_j|x_{il}]]=0$. This implies that regressors are orthogonal to the disturbances from the same *and* all other observations. Orthogonality with respect to the *same* observations is expressed by

$$\mathrm{E}[\boldsymbol{X}'\boldsymbol{\epsilon}] = \boldsymbol{0}.$$

Orthogonality is equivalent to zero correlation between $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$:

$$\mathrm{cov}[\boldsymbol{X},\boldsymbol{\epsilon}] = \mathrm{E}[\boldsymbol{X}'\boldsymbol{\epsilon}] - \mathrm{E}[\boldsymbol{X}]\mathrm{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}.$$

Note that this orthogonality must not be confused with orthogonality between $\boldsymbol{X}$ and the residuals $\boldsymbol{e}$ from LS estimation (see section 1.1.2). There it is a consequence of choosing $\boldsymbol{b}$ such the sum of squared errors is minimized. Here it is an assumption that refers to the unknown $\boldsymbol{\epsilon}$.

4. $y_i\epsilon_i = \boldsymbol{x}_i'\epsilon_i\boldsymbol{\beta} + \epsilon_i^2 \Rightarrow \mathrm{E}[y_i\epsilon_i] = \mathrm{E}[\epsilon_i^2] = \mathrm{V}[\epsilon_i]$.

If **AX** holds, the explanatory variables are (strictly) **exogenous**. The term **endogeneity** (i.e. one or all explanatory variables are endogenous) is used if **AX** does not hold (broadly speaking, if $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are correlated). Note that sometimes, instead of assuming **AX** to hold, the assumptions $\mathrm{E}[\boldsymbol{\epsilon}]=\mathbf{0}$ or $\mathrm{E}[\boldsymbol{X}'\boldsymbol{\epsilon}]=\mathbf{0}$ are made instead.

For example, **AX** is violated when a regressor, *in fact*, is determined on the basis of the dependent variable $\mathbf{y}$. This is the case in any situation where $\mathbf{y}$ and $\mathbf{X}$ (at least one of its columns) are determined simultaneously. A classic example are regressions attempting to analyze the effect of the number of policemen on the crime rate. These are bound to fail whenever the police force is driven by the number of crimes committed. Solutions to this kind of problem are discussed in section 1.9.1. Another example are regressions relating the performance of funds to their size. It is conceivable that an unobserved variable like the skill of fund managers affects size and performance. If that is the case, **AX** is violated.

Another important case where **AX** does not hold is a model where the **lagged dependent variable** is used as a regressor:

$$y_t = \phi y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + \epsilon_t \qquad y_{t+1} = \phi y_t + \boldsymbol{x}_{t+1}'\boldsymbol{\beta} + \epsilon_{t+1} \qquad y_{t+2} = \ldots.$$

AX requires the disturbance ϵ_t to be uncorrelated with regressors from any other observation, e.g. with y_t from the equation for $t+1$. **AX** is violated because $\mathrm{E}[y_t\epsilon_t]\neq 0$.

There are two main reasons for adding y_{t-1} to a regression: (a) to account for autocorrelated residuals (see section 1.7.3), and (b) to account for potentially missing regressors (see section 1.6.7 for a detailed treatment of the omitted variable bias). The effect of omitted regressors is captured by ϵ_t which affects y_t. In a time series context one can assume (or hope) that y_{t-1} partly reflects that missing information, in particular with rather frequently observed data. Hence, we are faced with a situation where the bias from adding the lagged dependent variable may be accepted to avoid the bias from omitted regressors.[9]

Predictive regressions are obtained when a predictor x_t enters only with a lag:

$$y_t = \beta_0 + \beta_1 x_{t-1} + \epsilon_t.$$

For dependent variables like asset returns (i.e. $y_t = \ln p_t/p_{t-1}$) a typically used predictor is the dividend-price ratio (i.e. $x_t = \ln d_t/p_{t-1}$. Stambaugh (1999) argues that, despite $\mathrm{E}[\epsilon_t|x_{t-1}]=0$, in a predictive regression $\mathrm{E}[\epsilon_t|x_t] \neq 0$, and thus **AX** is violated. To understand this reasoning, we consider

$$\underbrace{\ln p_t - \ln p_{t-1}}_{y_t} = \beta_1 \underbrace{(\ln d_{t-1} - \ln p_{t-1})}_{x_{t-1}} + \epsilon_t,$$

[9] As shown below, (a) the effects of adding a lagged dependent variable depend on the resulting residual autocorrelation, and (b) omitted regressors lead to biased and *inconsistent* coefficients.

$$\underbrace{\ln p_{t+1} - \ln p_t}_{y_{t+1}} = \beta_1 (\underbrace{\ln d_t - \ln p_t}_{x_t}) + \epsilon_{t+1} \ldots,$$

where β_0=0 for simplicity. Disturbances ϵ_t affect the price in t, (and, for given p_{t-1}, the return during the period $t-1$ to t). Thus, they are correlated with p_t, and hence with the regressor in the equation for $t+1$. Although the mechanism appears similar to the case of a lagged dependent variable, here the correlation between the disturbances and very specifically defined predictors x_t is the source of violation of **AX**. Stambaugh (1999) shows that this leads to a finite-sample bias (see below) in the estimated parameter b_1, *irrespective* of β_1 (e.g. even if β_1=0).

Assumption AH (homoscedasticity; uncorrelatedness): This assumptions covers two aspects. It states that the (conditional) variance of the disturbances is constant across observations (assuming that **AX** holds):

$$\mathrm{V}[\epsilon_i|\boldsymbol{X}] = \mathrm{E}[\epsilon_i^2|\boldsymbol{X}] - (\mathrm{E}[\epsilon_i|\boldsymbol{X}])^2 = \mathrm{E}[\epsilon_i^2|\boldsymbol{X}] = \sigma^2 \quad \forall i.$$

The errors are said to be **heteroscedastic** if their variance is not constant.

The second aspect of **AH** relates to the (conditional) covariance of $\boldsymbol{\epsilon}$ which is assumed to be zero:

$$\mathrm{cov}[\epsilon_i, \epsilon_j|\boldsymbol{X}] = 0 \quad \forall i \neq j \qquad \mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\boldsymbol{X}] = \mathrm{V}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma^2 \boldsymbol{I}.$$

This aspect of **AH** implies that the errors from different observations are not correlated. In a time series context this correlation is called **serial** or **autocorrelation**.

Assumption AN (normality): Assumptions **AX** and **AH** imply that the mean and variance of $\boldsymbol{\epsilon}|\boldsymbol{X}$ are $\mathbf{0}$ and $\sigma^2\boldsymbol{I}$. Adding the assumption of normality we have

$$\boldsymbol{\epsilon}|\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I}).$$

Since $\boldsymbol{X}$ plays no role in the distribution of $\boldsymbol{\epsilon}$, we have $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$. This assumption is useful to construct test statistics (see section 1.2.3), although many of the subsequent results do not require normality.

1.2.2 Properties

Expected value of $\boldsymbol{b}$ (**AL**,**AR**,**AX**): We first take the conditional expectation of (3)

$$\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}] = \boldsymbol{\beta} + \mathrm{E}[\boldsymbol{H}\boldsymbol{\epsilon}|\boldsymbol{X}] \qquad \boldsymbol{H} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Since $\boldsymbol{H}$ is a function of the conditioning variable $\boldsymbol{X}$ (see Review 2), it follows that

$$\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}] = \boldsymbol{\beta} + \boldsymbol{H}\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}],$$

and by assumption **AX** ($\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]=0$) we find that $\boldsymbol{b}$ is unbiased:

$$\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}] = \boldsymbol{\beta}.$$

By using the law of iterated expectations we can also derive the following unconditional result[10] (again using **AX**):

$$\mathrm{E}[\boldsymbol{b}] = \mathrm{E}_{\mathrm{x}}[\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}]] = \boldsymbol{\beta} + \mathrm{E}_{\mathrm{x}}[\boldsymbol{H}\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]] = \boldsymbol{\beta}.$$

We note that assumptions **AH** and **AN** are not required for unbiasedness, whereas **AX** is critical. Since a model with a lagged dependent variable violates **AX**, all coefficients in such a regression will be biased.

Covariance of $\boldsymbol{b}$ (**AL**,**AR**,**AX**,**AH**): The covariance of $\boldsymbol{b}$ conditional on $\boldsymbol{X}$ is given by

$$\begin{aligned}
\mathrm{V}[\boldsymbol{b}|\boldsymbol{X}] &= \mathrm{E}[(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})'|\boldsymbol{X}] \\
&= \mathrm{E}[\boldsymbol{H}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{H}'|\boldsymbol{X}] \\
&= \boldsymbol{H}\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\boldsymbol{X}]\boldsymbol{H}' \\
&= \boldsymbol{H}(\sigma^2\boldsymbol{I})\boldsymbol{H}' = \sigma^2\boldsymbol{H}\boldsymbol{H}' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad \text{since } \boldsymbol{H}\boldsymbol{H}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned} \tag{5}$$

For the special case of a single regressor the variance of b_1 is given by

$$\mathrm{V}[b_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sigma^2}{(n-1)\sigma_x^2}, \tag{6}$$

which shows that the precision of the estimate increases with the sample size and the variance of the regressor σ_x^2, and decreases with the variance of the disturbances.

To derive the unconditional covariance of $\boldsymbol{b}$ we use the variance decomposition

$$\mathrm{E}[\mathrm{V}[\boldsymbol{b}|\boldsymbol{X}]] = \mathrm{V}[\boldsymbol{b}] - \mathrm{V}[\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}]].$$

[10] To verify that $\boldsymbol{b}$ is unbiased conditionally *and* unconditionally by simulation one could generate samples of $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ for *fixed* $\boldsymbol{X}$ using many realizations of $\boldsymbol{\epsilon}$. The average over the OLS estimates $\boldsymbol{b}|\boldsymbol{X}$ – corresponding to $\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}]$ – should be equal to $\boldsymbol{\beta}$. However, if $\boldsymbol{X}$ is also allowed to vary across samples the average over $\boldsymbol{b}$ – corresponding to the unconditional mean $\mathrm{E}[\boldsymbol{b}]=\mathrm{E}[\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}]]$ – should also equal $\boldsymbol{\beta}$.

Since $\mathrm{E}[\boldsymbol{b}|\boldsymbol{X}]=\boldsymbol{\beta}$ the second term is zero and

$$\mathrm{V}[\boldsymbol{b}] = \mathrm{E}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}] = \sigma^2\mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}],$$

which implies that the unconditional covariance of $\boldsymbol{b}$ depends on the population covariance of the regressors.

Variance of $\boldsymbol{e}$ (**AL**,**AR**,**AX**,**AH**): The variance of $\boldsymbol{b}$ is expressed in terms of σ^2 (the population variance of ϵ). To estimate the covariance of $\boldsymbol{b}$ from a sample we replace σ^2 by the unbiased estimator

$$s_e^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K} \qquad \mathrm{E}[s_e^2] = \sigma^2.$$

Its square root s_e is the **standard error of regression**. s_e is measured in the same units as y. It may be a more informative measure for the goodness of fit than R^2, which is expressed in terms of variances (measured in *squared* units of y).

The estimated **standard error** of $\boldsymbol{b}$ denoted by $\mathrm{se}[\boldsymbol{b}]$ is the square root of the diagonal of

$$\hat{\mathrm{V}}[\boldsymbol{b}|\boldsymbol{X}] = s_e^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Efficiency (**AL**,**AR**,**AX**,**AH**): The **Gauss-Markov Theorem** states that the OLS estimator $\boldsymbol{b}$ is not only unbiased but has the minimum variance of all *linear* unbiased estimators (BLUE) and is thus efficient. This result holds whether $\boldsymbol{X}$ is stochastic or not. If **AN** holds (the disturbances are normal) $\boldsymbol{b}$ has the minimum variance of *all* unbiased (linear or not) estimators (see Greene (2003), p.47,48).

Sampling distribution of $\boldsymbol{b}$ (**AL**,**AR**,**AX**,**AH**,**AN**): Given (3) and **AN** the distribution of $\boldsymbol{b}$ is normal for given $\boldsymbol{X}$:

$$\boldsymbol{b}|\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

The sample covariance of $\boldsymbol{b}$ is obtained by replacing σ^2 with s_e^2, and is given by $\hat{\mathrm{V}}[\boldsymbol{b}]$ defined above.

Example 2: The standard error of regression from example 1 is 18.2 billion US$. This can be compared to the standard deviation of real investment which amounts to 34 billion US$. s_e is used to compute the (estimated) standard errors for the estimated coefficients which are given by

$$\mathrm{se}[\boldsymbol{b}]=(0.0503 \quad 0.0515 \quad 0.00328 \quad 0.00365)'.$$

Further details can be found in the files `investment.R` or `investment.xls`.

1.2.3 Testing hypothesis

Review 4: A **null hypothesis** H_0 formulates a restriction with respect to an unknown parameter of the population $\theta=\theta_0$. In a **two-sided test** the alternative hypothesis H_a is $\theta\neq\theta_0$. The test procedure is a rule that rejects H_0 if the sample estimate $\hat{\theta}$ is 'too far away' from θ_0. This rule can be based on the $1-\alpha$ **confidence interval** $\hat{\theta}\pm Q(\alpha/2)\text{se}[\hat{\theta}]$, where $Q(\alpha)$ denotes the α-quantile of the sampling distribution of $\hat{\theta}$. H_0 is rejected if θ_0 is outside the confidence interval.

If $Y\sim N(\mu,\sigma^2)$ and $Z=(y-\mu)/\sigma$ then $Z\sim N(0,1)$. $\Phi(Z)=P[Y\leq y^*]=\Phi((y^*-\mu)/\sigma)$ is the standard normal distribution function (e.g. $\Phi(-1.96)=0.025$). z_α is the α-quantile of the standard normal distribution, such that $P[Z\leq z_\alpha]=\alpha$ (e.g. $z_{0.025}=-1.96$).

> **Example 3:** Consider a sample of n observations from a normal population with mean μ and standard deviation σ. The sampling distribution of the sample mean $\bar{y}$ is also normal. The standard error of the mean is $\sigma/\sqrt{n}$. The $1-\alpha$ confidence interval for the unknown mean μ is $\bar{y}\pm z_{\alpha/2}\sigma/\sqrt{n}$. The *estimated* standard error of the mean $\text{se}[\bar{y}]=s/\sqrt{n}$ is obtained by replacing σ with the sample estimate s. In this case the $1-\alpha$ confidence interval is given by $\bar{y}\pm T(\alpha/2,n-1)s/\sqrt{n}$ where $T(\alpha,n-1)$ denotes the α-quantile of the t-distribution (e.g. $T(0.025,20)=-2.086$). If n is large the standard normal and t-quantiles are practically equal. In that case the interval is given by $\bar{y}\pm z_{\alpha/2}s/\sqrt{n}$.

A **type I error** is committed if H_0 is rejected although it is true. The probability of a type I error is the **significance level** (or **size**) α. If H_0 is rejected, $\hat{\theta}$ is said to be *significantly different* from θ_0 at a level of α. A type II error is committed if H_0 is not rejected although it is false. The **power** of a test is the probability of correctly rejecting a false null hypothesis. The power depends on the true parameter (which is usually unknown).

A **test statistic** is based on a sample estimate $\hat{\theta}$ and θ_0. It is a random variable. The distribution of the test statistic (usually under H_0) can be used to specify a rule for rejecting H_0. H_0 is rejected if the test statistic exceeds **critical values** which depend on α (and other parameters). In a two-sided test the critical values are the $\alpha/2$-quantiles and $1-\alpha/2$-quantiles of the distribution. In a one-sided test of the form $H_0\geq\theta_0$ (and $H_a<\theta_0$) the critical value is the α-quantile (this implies that H_0 is rejected if $\hat{\theta}$ is 'far below' θ_0). If $H_0\leq\theta_0$ the critical value is the $1-\alpha$ quantile. The **p-value** is that level of α for which there is indifference between accepting or rejecting H_0.

> **Example 4:** We consider a hypothesis about the mean of a population. $\mu=\mu_0$ can be tested against $\mu\neq\mu_0$ using the t-statistic (or t-ratio) $t=(\bar{y}-\mu_0)/\text{se}[\bar{y}]$. t has a standard normal or t-distribution depending on whether σ or s is used to compute $\text{se}[\bar{y}]$. If s is used, the t-statistic is compared to $\pm T(\alpha/2,n-1)$ in a two-sided test. One-sided tests use $\pm T(\alpha,n-1)$. In a two-sided test, H_0 is rejected if $|t|>|T(\alpha/2,n-1)|$.

If $\boldsymbol{\epsilon}$ is normally distributed the t-statistic

$$t_i=\frac{b_i-\beta_i}{\text{se}[b_i]}$$

has a t-distribution with $n-K$ degrees of freedom (df). $\text{se}[b_i]$ (the standard error of b_i) is the square root of the i-th diagonal element of $\hat{V}[\boldsymbol{b}]$). t_i can be used to test hypotheses about single elements of $\boldsymbol{\beta}$.

A joint test of β_j=0 (j=1,...,k) can be based on the statistic

$$F = \frac{(n-K)R^2}{k(1-R^2)},$$

which has an F-distribution with df=(k,$n-K$) if the disturbances are normal.

> **Example 5:** The t-statistics for the estimated coefficients from example 1 are given by $(-1.44 \quad 4.59 \quad -1.08 \quad -0.0755)'$. As it turns out only the coefficient of real GNP is significantly different from zero at a level of α=5%. The F-statistic is 12.8 with a p-value<0.001. Thus, we reject the hypothesis that the coefficients are jointly equal to zero. Further details can be found in the file `investment.xls`.

> **Exercise 2:** Use the results from exercise 1 and test the estimated coefficients for individual and joint significance.

In general, hypothesis tests about $\boldsymbol{\beta}$ can be based on imposing a linear restriction $\boldsymbol{r}$ (a $K\times1$ vector consisting of zeros and $\pm$ ones) on $\boldsymbol{\beta}$ and $\boldsymbol{b}$, and compare $\delta=\boldsymbol{r}'\boldsymbol{\beta}$ to $d=\boldsymbol{r}'\boldsymbol{b}$. If d differs significantly from δ we conclude that the sample is inconsistent with (or, does not support) the hypothesis expressed by the restriction. Since $\boldsymbol{b}$ is normal, $\boldsymbol{r}'\boldsymbol{b}$ is also normal, and the test statistic

$$t = \frac{d-\delta}{\text{se}[d]} \qquad \text{se}[d] = \sqrt{\boldsymbol{r}'\left[s_e^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]\boldsymbol{r}}$$

has a t-distribution with df=$n-K$.

We can consider several restrictions at once by using the $m\times K$ matrix $\boldsymbol{R}$ to define $\boldsymbol{\delta}=\boldsymbol{R}\boldsymbol{\beta}$ and $\boldsymbol{d}=\boldsymbol{R}\boldsymbol{b}$. Under the null that all restrictions hold we can define the **Wald statistic**

$$W = (\boldsymbol{d}-\boldsymbol{\delta})'\left[s_e^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1}(\boldsymbol{d}-\boldsymbol{\delta}). \tag{7}$$

W has a χ^2_m-distribution if the sample is large enough (see section 1.5) (or s_e^2 in (7) is replaced by the usually unknown σ^2). Instead, one can use the test statistic W/m which has an F-distribution with df=(m,$n-K$). In small samples, a test based on W/m will be more conservative (i.e. will have larger p-values).

So far, restrictions have been tested using the estimates from the unrestricted model. Alternatively, restrictions may directly be imposed when the parameters are estimated. This will lead to a loss of fit (i.e. R^2 will decrease). If R_r^2 is based on the parameter vector $\boldsymbol{b}_r$ (where some of the parameters are fixed rather than estimated) and R_u^2 is based on the unrestricted estimate, the test statistic

$$F = \frac{(n-K)(R_u^2-R_r^2)}{m(1-R_u^2)}$$

has an F-distribution with df=(m,$n-K$). It can be shown that $F=W/m$ (see Greene (2003), section 6.3). If F is significantly different from zero, H_0 is rejected and the restrictions are considered to be jointly invalid.

The distribution of the test statistics t, F and W depends on assumption **AN** (normality of disturbances). In section 1.3 we will comment on the case that **AN** does not hold.

1.2.4 Example 6: CAPM, beta-factors and multi-factor models

The **Capital Asset Pricing Model** (**CAPM**) considers the equilibrium relation between the expected return of an asset or portfolio (μ_i=E[y^i]), the risk-free return r_f, and the expected return of the market portfolio (μ_m=E[y^m]). Based on various assumptions (e.g. quadratic utility or normality of returns) the CAPM states that

$$\mu_i - r_f = \beta_i(\mu_m - r_f). \tag{8}$$

This relation is also known as the **security market line** (SML). In the CAPM the so-called **beta-factor** β_i defined as

$$\beta_i = \frac{\text{cov}[y^i, y^m]}{\text{V}[y^m]}$$

is the appropriate measure of an asset's risk. The (total) variance of the asset's returns is an inappropriate measure of risk since a part of this variance can be diversified away by holding the asset in a portfolio. The risk of the market portfolio cannot be diversified any further. The beta-factor β_i shows how the asset responds to market-wide movements and measures the market risk or systematic risk of the asset. The risk premium an investor can expect to obtain (or requires) is proportional to β_i. Assets with β_i>1 imply more risk than the market and should thus earn a proportionately higher risk premium.

Observed returns of the asset (y_t^i; t=1,...,n) and the market portfolio (y_t^m) can be used to estimate β_i or to test the CAPM. Under the assumption that observed returns deviate from expected returns we obtain

$$y_t^i - \mu_i = u_t^i \qquad y_t^m - \mu_m = u_t^m.$$

When we substitute these definitions for the expected values in the CAPM we obtain the so-called **market model**

$$y_t^i = \alpha_i + \beta_i y_t^m + \epsilon_t^i,$$

where α_i=(1−β_i)r_f and ϵ_t^i=u_t^i−$\beta_i u_t^m$. The coefficients α_i and β_i in this equation can be estimated by OLS. If we write the regression equation in terms of (observed) *excess* returns x_t^i=y_t^i−r_f and x_t^m=y_t^m−r_f we obtain

$$x_t^i = \beta_i x_t^m + \epsilon_t^i.$$

Thus the *testable implication* of the CAPM is that the constant term in a simple linear regression using excess returns should be equal to zero. In addition, the CAPM implies that there must not be any other risk factors than the market portfolio (i.e. the coefficients of such factors should not be significantly different from zero).

We use monthly data on the excess return of two industry portfolios (consumer goods and hi-tech) compiled by French[11]. We regress the excess returns of the two industries

[11] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. The files `capm.wf1` and `capm.txt` are based on previous versions of data posted there. These files have been compiled using the datasets which are now labelled as "5 Industry Portfolios" and "Fama/French 3 Factors" (which includes the risk-free return r_f).

on the excess market return based on a value-weighted average of all NYSE, AMEX, and NASDAQ firms (all returns are measured in percentage terms). Using data from January 2000 to December 2004 (n=60) we obtain the following estimates for the consumer goods portfolio (p-values in parenthesis; details can be found in the file `capm.wf1`)

$$\underset{}{x_t^i} = \underset{(0.36)}{0.343} + \underset{(0.0)}{0.624}\, x_t^m + e_t^i \qquad R^2 = 0.54 \qquad s_e = 2.9,$$

and for the hi-tech portfolio

$$\underset{}{x_t^i} = \underset{(0.11)}{-0.717} + \underset{(0.0)}{1.74}\, x_t^m + e_t^i \qquad R^2 = 0.87 \qquad s_e = 3.43.$$

The coefficients 0.624 and 1.74 indicate that a change in the (excess) market return by one percentage point implies a change in the expected excess return by 0.624 percentage points and 1.74 percentage points, respectively. In other words, the hi-tech portfolio has much higher market risk than the consumer goods portfolio.

The market model can be used to decompose the total variance of an asset into market- and firm-specific variance as follows (assuming that $\text{cov}[y^m, \epsilon^i]$=0):

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{\epsilon^i}^2.$$

$\beta_i^2\sigma_m^2$ can be interpreted as the risk that is market-specific or systematic (cannot be diversified since it is due to market-wide movements) and $\sigma_{\epsilon^i}^2$ is firm-specific (or idiosyncratic) risk. Since R^2 can also be written as $(\beta_i^2\sigma_m^2)/\sigma_i^2$ it measures the proportion of the market-specific variance in total variance. The R^2 from the two equations imply that 53% and 86% of the variance in the portfolio's returns are systematic. The higher R^2 from the hi-tech regression indicates that this industry is better diversified than the consumer goods industry. The p-values of the constant terms indicate that the CAPM implication cannot be rejected. This conclusion changes, however, when the sample size is increased.

The CAPM makes an (equilibrium) statement about *all* assets as expressed by the security market line (8). In order to test the CAPM, beta-factors $\hat{\beta}_i$ for many assets are estimated from the market model using time-series regressions. Then mean returns $\bar{y}_i$ for each asset (as an average across time) are computed, and the cross-sectional regression

$$\bar{y}_i = \lambda_f + \lambda_m \hat{\beta}_i + \eta_i$$

is run. The estimates for λ_f and λ_m (the market risk premium) are estimates of r_f and $(\mu_m - r_f)$ in equation (8). If the CAPM is valid, the mean returns of all assets should be located on the SML – i.e. on the line implied by this regression. However, there are some problems associated with this regression. The usual OLS standard errors of the estimated coefficients are incorrect because of heteroscedasticity in the residuals. In addition, the regressors $\hat{\beta}_i$ are subject to an **errors-in-variables** problem since they are not observed and will not correspond to the 'true' beta-factors.

Fama and MacBeth (1973) have suggested a procedure to improve the precision of the estimates. They first estimate beta-factors $\hat{\beta}_{it}$ for a large number of assets by running

the market model regression using monthly[12] time series of excess returns. The estimated beta-factors are subsequently used as regressors in the cross-sectional regression

$$y_t^i = \lambda_{ft} + \lambda_{mt}\hat{\beta}_{it} + \eta_{it}.$$

Note that $\hat{\beta}_{it}$ is based on an excess return series which ends one month before the cross-sectional regression is estimated (i.e. using x_s^i and x_s^m for $s{=}t{-}n,\ldots,t{-}1$). The cross-sectional regression is run in each month of the sample period and a times series of estimates $\hat{\lambda}_{ft}$ and $\hat{\lambda}_{mt}$ is obtained. The sample means and the standard errors of $\hat{\lambda}_{ft}$ and $\hat{\lambda}_{mt}$ are used as the final estimates for statistical inference[13]. Although the Fama-MacBeth approach yields improved estimates, Shanken (1992) has pointed out further deficiencies and has suggested a correction.

The CAPM has been frequently challenged by empirical evidence indicating significant risk premia associated with other factors than the market portfolio. A crucial aspect of the CAPM (in addition to assumptions about utility or return distributions) is that the market portfolio must include *all available* assets (which is hard to achieve in empirical studies). According to the **Arbitrage Pricing Theory** (**APT**) by Ross (1976) there exist *several* risk factors F_j that are *common* to a set of assets. The factors are assumed to be uncorrelated, but no further assumptions about utility or return distributions are made. These risk factors (and not only the market risk) capture the systematic risk component. Although the APT does not explicitly specify the nature of these factors, empirical research has typically considered two types of factors. One factor type corresponds to macroeconomic conditions such as inflation or industrial production (see Chen et al., 1986), and a second type corresponds to portfolios (see Fama and French, 1992). Considering only two common factors (for notational simplicity) the asset returns are governed by the factor model

$$y_t^i = \alpha_i + \beta_{i1}F_{t1} + \beta_{i2}F_{t2} + \epsilon_t^i,$$

where β_{ji} are the **factor sensitivities** (or **factor loadings**). The expected return of a single asset in this two-factor model is given by

$$\mathrm{E}[y^i] = \mu_i = \lambda_0 + \lambda_1\beta_{i1} + \lambda_2\beta_{i2},$$

where λ_j is the factor risk premium of F_j and $\lambda_0{=}r_f$. Using $\mathrm{V}[F_j]{=}\sigma_j^2$ and $\mathrm{cov}[F_1, F_2]{=}0$ the total variance of an asset can be decomposed as follows:

$$\sigma_i^2 = \beta_{i1}^2\sigma_1^2 + \beta_{i2}^2\sigma_2^2 + \sigma_{\epsilon^i}^2.$$

Estimation of the beta-factors is done by **factor analysis**, which is not treated in this text. For further details of the APT and associated empirical investigations see Roll and Ross (1980).

We briefly investigate one version of multi-factor models using the so-called Fama-French benchmark factors SMB (small minus big) and HML (high minus low) to test whether

[12] Using monthly data is not a prerequisite of the procedure. It could be performed using other data frequencies as well.

[13] See `Fama-MacBeth.xlsx` for an illustration of the procedure using only 30 assets and the S&P500 index.

excess returns depend on other factors than the market return. The factor SMB measures the difference in returns of portfolios of small and large stocks, and is intended to measure the so-called **size effect**. HML measures the difference between value stocks (having a high book value relative to their market value) and growth stocks (with a low book-market ratio).[14] The estimated regression equations are (details can be found in the file `capm.wf1`)

$$\underset{}{x_t^i} = \underset{(0.8)}{0.085} + \underset{(0.0)}{0.68}\, x_t^m - \underset{(0.30)}{0.089}\,\mathrm{SMB}_t + \underset{(0.0)}{0.29}\,\mathrm{HML}_t + e_t \qquad R^2 = 0.7$$

for the consumer goods portfolio and

$$\underset{}{x_t^i} = \underset{(0.07)}{-0.83} + \underset{(0.0)}{1.66}\, x_t^m + \underset{(0.04)}{0.244}\,\mathrm{SMB}_t - \underset{(0.21)}{0.112}\,\mathrm{HML}_t + e_t \qquad R^2 = 0.89$$

for the hi-tech portfolio. Consistent with the CAPM the constant terms in the first case is not significant. The beta-factor remains significant in both industries and changes only slightly compared to the market model estimates. However, the results indicate a significant return premium for holding value stocks in the consumer goods industry. For the hi-tech portfolio we find support for a size-effect. Overall, the results can be viewed as supporting multi-factor models.

> **Exercise 3:** Retrieve excess returns for industry portfolios of your choice from French's website. Estimate beta-factors in the context of multi-factor models. Interpret the results and test implications of the CAPM.

[14] Further details on the variable definitions and the underlying considerations can be found on French's website http://mba.tuck.dartmouth.edu/pages/faculty/ken.french.

1.2.5 Example 7: Interest rate parity

We consider a European investor who invests in a riskless US deposit with rate r_f. He buys US dollars at the spot exchange rate S_t (S_t is the amount in Euro paid/received for one dollar), invests at r_f, and after one period converts back to Euro at the rate S_{t+1}. The one-period return on this investment is given by

$$\ln S_{t+1} - \ln S_t + r_f.$$

Forward exchange rates F_t can be used to hedge against the currency risk (introduced by the unknown S_{t+1}) involved in this investment. If F_t denotes the rate fixed at t to buy/sell US dollars in t+1 the (certain) return is given by

$$\ln F_t - \ln S_t + r_f.$$

Since this return is riskless it must equal the return r_f^d from a domestic riskless investment to avoid arbitrage. This leads to the **covered interest rate parity** (CIRP)

$$r_f^d - r_f = \ln F_t - \ln S_t.$$

The left hand side is the interest rate differential and the right hand side is the **forward premium**.

The **uncovered interest rate parity** (UIRP) is defined in terms of the *expected* spot rate

$$r_f^d - r_f = \mathrm{E}_t[\ln S_{t+1}] - \ln S_t.$$

$\mathrm{E}_t[\ln S_{t+1}]$ can differ from $\ln F_t$ if the market pays a risk premium for taking the risk of an unhedged investment. A narrowly defined version of the UIRP assumes risk neutrality and states that the risk premium is zero (see Engel, 1996, for a survey)

$$\mathrm{E}_t[\ln S_{t+1} - \ln S_t] = \ln F_t - \ln S_t.$$

Observed exchange rates S_{t+1} can deviate from F_t, but the expected difference must be zero. The UIRP can be tested using the **Fama regression**

$$s_t - s_{t-1} = \beta_0 + \beta_1(f_{t-1} - s_{t-1}) + \epsilon_t,$$

where s_t=ln S_t and f_t=ln F_t. The UIRP imposes the testable restrictions β_0=0 and β_1=1.[15] We use a data set[16] from Verbeek (2004) and obtain the following results (t-statistics in parenthesis)

$$s_t - s_{t-1} = \underset{(0.72)}{0.0023} + \underset{(0.67)}{0.515}\,(f_{t-1} - s_{t-1}) + e_t \qquad R^2 = 0.00165.$$

[15] Hayashi (2000, p.424) discusses the question, why UIRP cannot be tested on the basis of s_t=β_0+$\beta_1 f_{t-1}$+ϵ_t.

[16] This data is available from `http://eu.wiley.com/legacy/wileychi/verbeek2ed/datasets.html`. We use the corrected data set `forward2c` from chapter 4 (foreign exchange markets). Note that the exchange and forward rates in this dataset are expressed in terms of US dollars paid/received for one Euro. To make the data consistent with the description in this section we have defined the logs of spot and forward rates accordingly (although this does not change the substantive conclusions). Details can be found in the files `uirp.R` or `uirp.xls`.

Testing the coefficients individually shows that b_0 is not significantly different from 0 and b_1 is not significantly different from 1.

To test both restrictions at once we define

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\delta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The Wald statistic for testing both restrictions equals 3.903 with a p-value of 0.142. The p-value of the F-statistic $W/2$=1.952 is 0.144. Alternatively, we can use the R^2 from the restricted model with β_0=0 and β_1=1. This requires to define restricted residuals according to $(s_t - s_{t-1}) - (f_{t-1} - s_{t-1})$. The associated R^2 is negative and the F-statistic is again 1.952. Thus, the joint test confirms the conclusion derived from testing individual coefficients, and we cannot reject UIRP (which does not mean that UIRP holds!).

> **Exercise 4:** Repeat the analysis and tests from example 7 but use the US dollar/British pound exchange and forward rates in the files `forward2c.dat`, `uirp.xls`, or `uirp.wf1` to test the UIRP.

1.2.6 Prediction

Regression models can be also used for out-of-sample **prediction**. Suppose the estimated model from n observations is $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+\boldsymbol{e}$ and we want to predict y_0 given a new observation of the regressors $\boldsymbol{x}_0$ which has not been included in the estimation (hence: out-of-sample). From the Gauss-Markov theorem it follows that the prediction

$$\hat{y}_0 = \boldsymbol{x}_0'\boldsymbol{b}$$

is the BLUE of $\mathrm{E}[y_0]$. Its variance is given by

$$\mathrm{V}[\hat{y}_0] = \mathrm{V}[\boldsymbol{x}_0'\boldsymbol{b}] = \boldsymbol{x}_0'\mathrm{V}[\boldsymbol{b}]\boldsymbol{x}_0 = \boldsymbol{x}_0'\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0,$$

and reflects the sampling error of $\boldsymbol{b}$. The prediction error is

$$e_0 = y_0 - \hat{y}_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon_0 - \boldsymbol{x}_0'\boldsymbol{b} = \epsilon_0 + \boldsymbol{x}_0'(\boldsymbol{\beta} - \boldsymbol{b}),$$

and its variance is given by

$$\mathrm{V}[e_0] = \sigma^2 + \mathrm{V}[\boldsymbol{x}_0'(\boldsymbol{\beta} - \boldsymbol{b})] = \sigma^2 + \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0.$$

The variance can be estimated by using s_e^2 in place of σ^2. For the special case of a single regressor the variance of e_0 is given by (see (6) and Kmenta (1971), p.240)

$$\sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].$$

This shows that the variance of the prediction (error) increases with the distance of $\boldsymbol{x}_0$ from the mean of the regressors and decreases with the sample size. The (estimated) variance of the disturbances can be viewed as a lower bound for the variance of the out-of-sample prediction error.
If σ^2 is replaced by s_e^2 we can compute a $1-\alpha$ prediction interval for y_0 from

$$\hat{\boldsymbol{y}}_0 \pm z_{\alpha/2}\mathrm{se}[\boldsymbol{e}_0],$$

where $\mathrm{se}[\boldsymbol{e}_0]$ is the square root of the estimated variance $\hat{\mathrm{V}}[\boldsymbol{e}_0]$. These calculations, using example 1, can be found in the file `investment.xls` on the sheet `prediction`.

1.3 Large sample properties of least squares estimates[17]

Review 5:[18] We consider the asymptotic properties of an estimator $\hat{\theta}_n$ which hold as the sample size n grows without bound.

Convergence: The random variable $\hat{\theta}_n$ **converges in probability** to the (non-random) constant c if, for any $\epsilon>0$,

$$\lim_{n\to\infty} \mathrm{P}[|\hat{\theta}_n - c| > \epsilon] = 0.$$

c is the **probability limit** of $\hat{\theta}_n$ and is denoted by $\operatorname{plim} \hat{\theta}_n=c$.
Rules for scalars x_n and y_n:

$$\operatorname{plim}(x_n + y_n) = \operatorname{plim} x_n + \operatorname{plim} y_n \qquad \operatorname{plim}(x_n \cdot y_n) = \operatorname{plim} x_n \cdot \operatorname{plim} y_n.$$

Rules for vectors and matrices:

$$\operatorname{plim} \boldsymbol{X}\boldsymbol{y} = \operatorname{plim} \boldsymbol{X} \cdot \operatorname{plim} \boldsymbol{y}.$$

Rule for a nonsingular matrix $\boldsymbol{X}$:

$$\operatorname{plim} \boldsymbol{X}^{-1} = (\operatorname{plim} \boldsymbol{X})^{-1}.$$

Consistency: $\hat{\theta}_n$ is **consistent** for θ if $\operatorname{plim} \hat{\theta}_n=\theta$. $\hat{\theta}_n$ is consistent if the **asymptotic bias** is zero and the **asymptotic variance** is zero:

$$\lim_{n\to\infty} \mathrm{E}[\hat{\theta}_n] - \theta = 0 \qquad \lim_{n\to\infty} \mathrm{aV}[\hat{\theta}_n] = 0.$$

Example: The sample mean $\bar{y}$ from a population with μ and σ^2 is consistent for μ since $\mathrm{E}[\bar{y}]=\mu$ and $\mathrm{aV}[\bar{y}]=\sigma^2/n$. Thus $\operatorname{plim} \bar{y}=\mu$.

Consistency of a mean of functions: Consider a random sample $(y_1,\ldots,y_n)$ from a random variable Y and any function $f(y)$. If $\mathrm{E}[f(Y)]$ and $\mathrm{V}[f(Y)]$ are finite constants then

$$\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} f(y_i) = \mathrm{E}[f(Y)].$$

Limiting distribution: $\hat{\theta}_n$ with cdf F_n **converges in distribution** to a random variable θ with cdf F (this is denoted by $\hat{\theta}_n \xrightarrow{d} \theta$) if

$$\lim_{n\to\infty} |F_n - F| = 0$$

for every continuity point of F. F is the **limiting** or **asymptotic distribution** of θ_n.
A consistent estimator $\hat{\theta}_n$ is **asymptotically normal** if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}(0, v) \quad \text{or} \quad \hat{\theta}_n \overset{a}{\sim} \mathrm{N}(\theta, v/n),$$

where $\mathrm{aV}[\hat{\theta}_n]=v/n$ is the asymptotic variance of $\hat{\theta}_n$.

[17] Most of this subsection is based on Greene (2003), sections 5.2 and 5.3.
[18] Greene (2003); section D.

Central Limit Theorem: If $\bar{y}$ is the sample mean of a random sample $(y_1,\ldots,y_n)$ from a distribution with mean μ and variance σ^2 (which need not be normal)

$$\sqrt{n}(\bar{y}-\mu) \xrightarrow{d} \mathrm{N}(0,\sigma^2) \quad \text{or} \quad \bar{y} \overset{a}{\sim} \mathrm{N}(\mu,\sigma^2/n).$$

Expressed differently,

$$z_n = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$$

is asymptotically standard normal: $z_n \overset{a}{\sim} \mathrm{N}(0,1)$.

The finite sample properties of OLS estimates only hold if assumptions **AL**, **AR**, **AX**, and **AH** are satisfied. **AN** is required to obtain the exact distribution of $\boldsymbol{b}$ and to derive (the distribution of) test statistics. Large-sample theory drops **AN** and adds other assumptions about the data generating mechanism. The sample is assumed to be large enough so that certain asymptotic properties hold, and an approximation of the distribution of OLS estimates can be derived.

1.3.1 Consistency

Consistency relates to the properties of $\boldsymbol{b}$ as $n\to\infty$. Therefore we use the formulation

$$\boldsymbol{b}_n = \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon}\right). \tag{9}$$

This shows that the large-sample properties of $\boldsymbol{b}_n$ depend on the behavior of the sample averages of $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{\epsilon}$. In addition to the assumptions from the previous subsection we assume that $(\boldsymbol{x}_i,\epsilon_i)$ are an i.i.d. sequence of random variables:

Aiid: $(\boldsymbol{x}_i,\epsilon_i) \sim$ i.i.d.

To prove consistency we consider the probability limit of $\boldsymbol{b}_n$:

$$\begin{aligned}\operatorname{plim}\boldsymbol{b}_n &= \boldsymbol{\beta} + \operatorname{plim}\left[\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\cdot\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon}\right)\right] \\ &= \boldsymbol{\beta} + \operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\cdot\operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon}\right).\end{aligned} \tag{10}$$

We have to make sure that the covariance matrix of regressors $\boldsymbol{X}$ is 'well behaved'. This requires that all elements of $\boldsymbol{X}'\boldsymbol{X}/n$ converge to finite constants (i.e. the corresponding population moments). This is expressed by the assumption

$$\overline{\mathbf{AR}}\text{:} \quad \operatorname{plim}\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}, \tag{11}$$

where $\boldsymbol{Q}$ is a positive definite matrix.

Regarding the second probability limit in (10), Greene (2003, p.66) defines

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\epsilon_i = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}_i = \bar{\boldsymbol{w}}_n$$

and uses $\mathbf{AX}$ to show that

$$\mathrm{E}[\bar{\boldsymbol{w}}_n] = \mathbf{0} \qquad \mathrm{V}[\bar{\boldsymbol{w}}_n] = \mathrm{E}[\bar{\boldsymbol{w}}_n \bar{\boldsymbol{w}}_n'] = \frac{\sigma^2}{n} \frac{\mathrm{E}[\boldsymbol{X}'\boldsymbol{X}]}{n}.$$

The variance of $\bar{\boldsymbol{w}}_n$ will converge to zero, which implies that $\operatorname{plim} \bar{\boldsymbol{w}}_n = \mathbf{0}$, or

$$\operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\epsilon}\right) = \mathbf{0}.$$

Thus the probability limit of $\boldsymbol{b}_n$ is given by

$$\operatorname{plim} \boldsymbol{b}_n = \boldsymbol{\beta} + \boldsymbol{Q}^{-1} \cdot \mathbf{0},$$

and we conclude that $\boldsymbol{b}_n$ is consistent:

$$\operatorname{plim} \boldsymbol{b}_n = \boldsymbol{\beta}.$$

1.3.2 Asymptotic normality

Large-sample theory is *not* based on the normality assumption **AN**, but derives an approximation of the distribution of OLS estimates. We rewrite (9) as

$$\sqrt{n}(\boldsymbol{b}_n - \boldsymbol{\beta}) = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\epsilon}\right) \tag{12}$$

to derive the asymptotic distribution of $\sqrt{n}(\boldsymbol{b}_n - \boldsymbol{\beta})$ using the central limit theorem. By **AR** the probability limit of the first term on the right hand side of (12) is $\boldsymbol{Q}^{-1}$. Next we consider the limiting distribution of

$$\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\epsilon} = \sqrt{n}\left(\bar{\boldsymbol{w}}_n - \mathrm{E}[\bar{\boldsymbol{w}}_n]\right).$$

$\bar{\boldsymbol{w}}_n$ is the average of n i.i.d. random vectors $\boldsymbol{w}_i = \boldsymbol{x}_i\epsilon_i$. From the previous subsection we know that $\mathrm{E}[\bar{\boldsymbol{w}}_n] = \boldsymbol{0}$. Greene (2003, p.68) shows that the variance of $\sqrt{n}\bar{\boldsymbol{w}}_n$ converges to $\sigma^2\boldsymbol{Q}$. Thus, in analogy to the univariate case, we can apply the central limit theorem. The means of the i.i.d. random vectors $\boldsymbol{w}_i$ converge to a normal distribution:

$$\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\epsilon} = \sqrt{n}\bar{\boldsymbol{w}}_n \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{Q}).$$

We can now complete the derivation of the limiting distribution of (12) by including $\boldsymbol{Q}^{-1}$ to obtain

$$\boldsymbol{Q}^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\epsilon} \xrightarrow{d} \mathrm{N}(\boldsymbol{Q}^{-1}\boldsymbol{0}, \boldsymbol{Q}^{-1}(\sigma^2\boldsymbol{Q})\boldsymbol{Q}^{-1})$$

or

$$\sqrt{n}(\boldsymbol{b}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{Q}^{-1}) \qquad \boldsymbol{b}_n \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \frac{\sigma^2}{n}\boldsymbol{Q}^{-1}).$$

Note that the asymptotic normality of $\boldsymbol{b}$ is not based on **AN** but on the central limit theorem. The asymptotic covariance of $\boldsymbol{b}_n$ is estimated by using $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ to estimate $(1/n)\boldsymbol{Q}^{-1}$ and $s_e^2 = \mathrm{SSE}/(n-K)$ to estimate σ^2:

$$\widehat{\mathrm{aV}}[\boldsymbol{b}_n] = s_e^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

This implies that t- and F-statistics are asymptotically valid even if the residuals are not normal. If F has an F-distribution with df$=(m,n-k)$ then $W = mF \overset{a}{\sim} \chi^2_m$.
In small samples the t-distribution may be a reasonable approximation[19] even when **AN** does not hold. Since it is more conservative than the standard normal, it may be preferable to use the t-distribution. By a similar argument, using the F-distribution (rather than $W = mF$ and the χ^2 distribution) can be justified in small samples when **AN** does not hold.

[19] If **AN** does not hold the finite sample distribution of the t-statistic is unknown.

1.3.3 Time series data[20]

With time series data the strict exogeneity assumption **AX** is usually hard to maintain. For example, a company's returns may depend on the current, exogenous macroeconomic conditions and the firm's past production (or investment, finance, etc.) decisions. To the extent that the company decides upon the level of production based on past realized returns (which include past disturbances), the current disturbances may be correlated with regressors in future equations. More generally, strict exogeneity might not hold if regressors are policy variables which are set depending on past outcomes.

If **AX** does not hold (e.g. in a model with a lagged dependent variable), $\boldsymbol{b}_n$ is biased. In the previous subsections consistency and asymptotic normality have been established on the basis of **Aiid** and $\overline{\mathbf{AR}}$. However, with time series data the i.i.d. assumption need not hold and the applicability of limit theorems is not straightforward. Nevertheless, consistent estimates in a time series context can still be obtained. The additional assumptions needed are based on the following concepts.

A **stochastic process** Y_t is a sequence[21] of random variables $Y_{-\infty},\ldots,Y_0,Y_1,\ldots,Y_{+\infty}$. An observed sequence y_t $(t{=}1,\ldots,n)$ is a sample or **realization** (one possible outcome) of the stochastic process. Any statistical inference about Y_t must be based on the *single draw* y_t from the so-called **ensemble** of realizations of the process. Two properties are crucial in this context: the process has to be **stationary** (i.e. the underlying distribution of Y_t does not change with t) and **ergodic** (i.e. each individual observation provides unique information about the process; adjacent observations must not be too similar). More formally, a stationary process is **ergodic** if any two random variables Y_t and $Y_{t-\ell}$ are asymptotically (i.e. $\ell{\to}\infty$) independent.

A stochastic process is characterized by the **autocovariance** γ_ℓ

$$\gamma_\ell = \mathrm{E}[(Y_t-\mu)(Y_{t-\ell}-\mu)] \qquad \mu = \mathrm{E}[Y_t], \tag{13}$$

or the **autocorrelation** ρ_ℓ

$$\rho_\ell = \frac{\gamma_\ell}{\gamma_0} = \frac{\gamma_\ell}{\sigma^2}. \tag{14}$$

A stochastic process is **weakly** or **covariance stationary** if $\mathrm{E}[Y_t^2]{<}\infty$ and if $\mathrm{E}[Y_t]$, $\mathrm{V}[Y_t]$ and γ_ℓ do not depend on t (i.e. γ_ℓ and ρ_ℓ only depend on ℓ). If Y_t is **strictly stationary** the joint distribution of Y_t and $Y_{t-\ell}$ does not depend on the time shift ℓ. If Y_t is weakly stationary and normally distributed then Y_t is also strictly stationary.

According to the **ergodic theorem**, averages from a single observed sequence will converge to the corresponding parameters of the population, if the process is stationary and ergodic. If Y_t is stationary and ergodic with $\mathrm{E}[Y_t]{=}\mu$, the sample mean obtained from a single realization y_t converges to μ asymptotically:

$$\lim_{n\to\infty} \frac{1}{n}\bar{y}_n = \sum_{t=1}^{n} y_t = \mu.$$

[20] Most of this subsection is based on Greene (2003), section 12.4.

[21] We use the index t since stochastic processes are frequently viewed in terms of chronologically ordered sequences across *time*. However, the index set is arbitrary and everything we say holds as well if the index refers to other entities (e.g. firms).

If Y_t is covariance stationary it is sufficient that $\sum_{\ell=0}^{\infty}|\gamma_\ell|<\infty$ (absolute summability) for the process to be ergodic for the mean. The theorem extends to any (finite) moment of stationary and ergodic processes. In the special case where Y_t is a normal and stationary process, then absolute summability is enough to insure ergodicity for all moments. Whereas many tests for stationarity are available (see section 2.3.3), ergodicity is difficult to test and is usually *assumed* to hold. Quickly decaying *estimated* autocorrelations can be taken as *empirical* evidence of stationarity and ergodicity.

In other words, the ergodic theorem implies that consistency does not require independent observations. Greene (2003, p.73) shows that consistency and asymptotic normality of the OLS estimator can be preserved in a time-series context by replacing **AX** with[22]

$$\overline{\mathbf{AX}}\text{:}\quad \mathrm{E}[\epsilon_t|\boldsymbol{x}_{t-\ell}]=0\quad(\forall\ell\geq 0),$$

replacing $\overline{\mathbf{AR}}$ by

$$\mathbf{AR}_t\text{:}\quad \operatorname{plim}\frac{1}{n-\ell}\sum_{t=\ell+1}^{n}\boldsymbol{x}_t\boldsymbol{x}_{t-\ell}'=\boldsymbol{Q}(\ell),$$

where $\boldsymbol{Q}(\ell)$ is a finite matrix, and by requiring that $\boldsymbol{Q}(\ell)$ has to converge to a matrix of zeros as $\ell\to\infty$. These properties of $\boldsymbol{Q}(\ell)$ can be summarized by the assumption that $\boldsymbol{x}_t$ is stationary and ergodic. In addition, the autocorrelation ρ_ℓ of the disturbances ϵ_t has to be zero (for all ℓ), although not always explicitly stated.

This has the following implications for models with a lagged dependent variables:

$$y_t=\phi_1 y_{t-1}+\cdots+\phi_p y_{t-p}+\boldsymbol{z}_t'\boldsymbol{\beta}+\epsilon_t.$$

Although estimates of ϕ_i and $\boldsymbol{\beta}$ are biased (since **AX** is violated), they are consistent provided $\overline{\mathbf{AX}}$ holds, $\boldsymbol{x}_t=[y_{t-1},\ldots,y_{t-p},\boldsymbol{z_t}]$ is stationary and ergodic, and ϵ_t is not autocorrelated. In section 1.7.3 we take a closer look at the case when ϵ_t is autocorrelated.

[22] Other authors (e.g. Hayashi, 2000, p.109) assume that ϵ_t and $\boldsymbol{x}_t$ are contemporaneously uncorrelated ($\mathrm{E}[\boldsymbol{x}_t\epsilon_t]=\mathbf{0}$), as implied by $\overline{\mathbf{AX}}$.

1.4 Maximum likelihood estimation

Review 6:[23] We consider a random sample y_i (i=1, ..., n) to estimate the parameters μ and σ^2 of a random variable $Y \sim \mathrm{N}(\mu, \sigma^2)$. The maximum likelihood (ML) estimates are those values for the parameters of the underlying distribution which make the observed sample most likely (i.e. would generate it most frequently).

The **likelihood (function)** $L(\boldsymbol{\theta})$ is the joint density evaluated at the observations y_i (i=1, ..., n) as a function of the parameter (vector) $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta}) f(y_2|\boldsymbol{\theta}) \cdots f(y_n|\boldsymbol{\theta}).$$

$f(y_i|\boldsymbol{\theta})$ is the value of the density function at y_i given the parameters $\boldsymbol{\theta}$. To simplify the involved calculations the logarithm of the likelihood function (the **log-likelihood**) is maximized:

$$\ln L(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}) \longrightarrow \max .$$

The ML method requires an assumption about the distribution of the population. Using the density function of the normal distribution and $\boldsymbol{\theta}$=(μ, σ^2) we have

$$\ln f(y_i|\mu, \sigma^2) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_i - \mu)^2}{2\sigma^2},$$

and the log-likelihood as a function of μ and σ^2 is given by

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

From the first derivatives with respect to μ and σ^2

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \qquad \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2$$

we obtain the ML estimates

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \tilde{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

To estimate more general models the constants μ and σ^2 can be replaced by conditional mean μ_i and variance σ_i^2, provided the standardized residuals ϵ_i=$(y_i - \mu_i)/\sigma_i$ are i.i.d. Then the likelihood depends on the coefficients in the equations which determine μ_i and σ_i^2.

The ML estimate of a regression model requires the specification of a distribution for the disturbances. If ϵ_i=$y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ is assumed to be i.i.d.[24] and normal $\epsilon \sim \mathrm{N}(0, \sigma^2)$, the log-likelihood is given by

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{15}$$

[23] For details see Kmenta (1971), p.174 or Wooldridge (2003), p.746.

[24] Note that the i.i.d. assumption is not necessary for the observations but only for the residuals.

The necessary conditions for a maximum are

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}}: \quad \frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{\epsilon} = \mathbf{0}$$

$$\frac{\partial \ell}{\partial \sigma^2}: \quad -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = 0.$$

The solution of these equations gives the estimates

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \tilde{s}_e^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n}.$$

ML estimators are attractive because of their large sample properties: provided that the model is correctly specified they are consistent, asymptotically efficient and asymptotically normal[25]:

$$\boldsymbol{b} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{I}(\boldsymbol{\beta})^{-1}).$$

$\boldsymbol{I}(\boldsymbol{\beta})$ is the **information matrix** evaluated at the true parameters. Its inverse can be used to estimate the covariance of $\boldsymbol{b}$. Theoretically, $\boldsymbol{I}(\boldsymbol{\beta})$ is minus the expected value of the second derivatives of the log-likelihood. In practice this expectation can be computed in either of the two following ways (see Greene, 2003, p.480). One way is to evaluate the Hessian matrix (the second derivatives of ℓ) at $\boldsymbol{b}$

$$\hat{\boldsymbol{I}}(\boldsymbol{b}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{b}\, \partial \boldsymbol{b}'},$$

where the second derivatives are usually calculated numerically. Another way is the BHHH estimator[26] which is based on the outer-product of the gradient (or **score** vector):

$$\hat{\boldsymbol{I}}(\boldsymbol{b}) = \sum_{i=1}^{n} \boldsymbol{g}_i \boldsymbol{g}_i' \qquad \boldsymbol{g}_i = \frac{\partial \ell_i(\boldsymbol{b})}{\partial \boldsymbol{b}} \qquad \text{or} \quad \hat{\boldsymbol{I}}(\boldsymbol{b}) = \boldsymbol{G}'\boldsymbol{G},$$

where $\boldsymbol{g}_i$ is the $K\times 1$ gradient for observation i, and $\boldsymbol{G}$ is a $n\times K$ matrix with rows equal to the transpose of the gradients for each observation.

In general, the Hessian and the BHHH approach do not yield the same results, even when the derivatives are available in closed form. The two estimates of $\boldsymbol{I}$ can also differ when the model is misspecified. **Quasi-ML** (**QML**) estimates are based on maximizing the likelihood using a distribution which is known to be incorrect (i.e. using the wrong density when formulating (15)). For instance, the normal distribution is frequently used as an approximation when the true distribution is unknown or cumbersome to use.

Significance tests of regression coefficients are based on the asymptotic normality of the ML estimates. z-statistics (rather than t-statistics) are frequently used to refer to the standard normal distribution of the test statistic $z_i=(b_i-\beta_i)/\mathrm{se}[b_i]$, where the standard error $\mathrm{se}[b_i]$ is the square root of the i-th diagonal element of the inverse of $\hat{\boldsymbol{I}}(\boldsymbol{b})$, and $z_i \overset{a}{\sim} \mathrm{N}(0,1)$.

The major weakness of ML estimates is their potential bias in small samples (e.g. the variance estimate is scaled by n rather than $n-K$).

[25] Greene (2003), p.473.

[26] BHHH refers to the initials of Berndt, Hall, Hall, and Hausman who have first proposed this approach.

Example 8: We use the quarterly investment data from 1950-2000 from Table 5.1 in Greene (2003) (see exercise 1). The explanatory variables are the log of real output, the nominal interest rate and the rate of inflation. to estimate the same regression as in example 1 by *numerically maximizing* the log-likelihood. The dependent variable is the log of real investment. Details can be found in the file `investment-ml.xls`.

The estimated ML coefficients are almost equal to the OLS estimates, and depend on the settings which trigger the convergence of the numerical optimization algorithm. The standard errors are based on the outer-product of the gradients, and are slightly different from those based on the inverse of $\boldsymbol{X}'\boldsymbol{X}$. Accordingly, the z-statistics differ from the t-statistics. The interest rate turns out to be the only regressor which is not statistically significant at the 5% level.

Exercise 5: Use the *annual* data and the regression equation from example 1 (see file `investment.xls`) and estimate the model by maximum likelihood.

1.5 LM, LR and Wald tests[27]

Suppose the ML estimate $\hat{\boldsymbol{\theta}}$ (a $K\times 1$ parameter vector) shall be used to test m linear restrictions H_0: $\boldsymbol{\delta}=\boldsymbol{R\theta}$. Three test principles can be used for that purpose.

The **Wald test** is based on unrestricted estimates. If the restrictions are valid, $\boldsymbol{d}=\boldsymbol{R}\hat{\boldsymbol{\theta}}$ will not deviate significantly from $\boldsymbol{\delta}$. The Wald test statistic for m restrictions defined as

$$W = (\boldsymbol{d}-\boldsymbol{\delta})'(\mathrm{V}[\boldsymbol{d}])^{-1}(\boldsymbol{d}-\boldsymbol{\delta}).$$

The covariance of $\boldsymbol{d}$ can be estimated by $\boldsymbol{R}\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]\boldsymbol{R}'$, where $\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]$ can be based on the inverse of the information matrix. Using $\hat{\mathrm{V}}[\hat{\boldsymbol{\theta}}]=\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})^{-1}$ we obtain

$$(\boldsymbol{d}-\boldsymbol{\delta})'\left[\boldsymbol{R}\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{R}'\right]^{-1}(\boldsymbol{d}-\boldsymbol{\delta}) \overset{a}{\sim} \chi^2_m. \tag{16}$$

The **likelihood-ratio (LR) test** requires estimating the model with and without restrictions. The LR test statistic is

$$2[\ell_u-\ell_r] \overset{a}{\sim} \chi^2_m, \tag{17}$$

where ℓ_u is the unrestricted log-likelihood, and ℓ_r the log-likelihood obtained by imposing m restrictions. If the restrictions are valid, the difference between ℓ_r and ℓ_u will be close to zero.

If parameters are estimated by OLS, the LR test statistic can be computed using the residuals $\boldsymbol{e}_u$ and $\boldsymbol{e}_r$ from unrestricted and restricted OLS regressions, respectively. For m restrictions the LR test statistic is given by

$$\mathrm{LR} = n[\ln(\boldsymbol{e}_r'\boldsymbol{e}_r)-\ln(\boldsymbol{e}_u'\boldsymbol{e}_u)] \qquad \mathrm{LR}\sim\chi^2_m.$$

The **Lagrange multiplier (LM) test** (or **score test**) is based on maximizing the log-likelihood under the restrictions using the Lagrangian function

$$\mathcal{L}(\boldsymbol{\theta}_r) = \ell(\boldsymbol{\theta}_r) + \boldsymbol{\lambda}'(\boldsymbol{R\theta}_r-\boldsymbol{\delta}).$$

The estimates $\hat{\boldsymbol{\theta}}_r$ and $\hat{\boldsymbol{\lambda}}$ can be obtained from the first order conditions

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_r}: \quad \frac{\partial\ell}{\partial\boldsymbol{\theta}_r}+\boldsymbol{\lambda}'\boldsymbol{R}=\boldsymbol{0}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda}}: \quad \boldsymbol{R\theta}_r-\boldsymbol{\delta}=\boldsymbol{0}.$$

Lagrange multipliers measure the improvement in the likelihood which can be obtained by relaxing constraints. If the restrictions are valid (i.e. hold in the data), imposing them is not necessary, and $\hat{\boldsymbol{\lambda}}$ will not differ significantly from zero. This consideration leads to

[27] This section is based on Greene (2000), p.150 and Greene (2003), section 17.5.

H_0: $\boldsymbol{\lambda}=\mathbf{0}$ (hence LM test). This is equivalent to testing the derivatives evaluated at the restricted estimates $\hat{\boldsymbol{\theta}}_r$:

$$\boldsymbol{g}_r = \frac{\partial \ell(\hat{\boldsymbol{\theta}}_r)}{\partial \hat{\boldsymbol{\theta}}_r} = -\hat{\boldsymbol{\lambda}}' \boldsymbol{R}.$$

Under H_0: $\boldsymbol{g}_r=\mathbf{0}$ the LM test statistic is given by

$$\boldsymbol{g}_r' \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_r)^{-1} \boldsymbol{g}_r \overset{a}{\sim} \chi^2_m. \tag{18}$$

$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_r)=\boldsymbol{G}_r'\boldsymbol{G}_r$, where $\boldsymbol{G}_r$ is a $n\times K$ matrix with rows equal to the transpose of the gradients for each observation evaluated at the *restricted* parameters.

Alternatively, the Lagrange multiplier (LM) test statistic can be derived from a regression of the restricted residuals $\boldsymbol{e}_r$ on all regressors including the constant (see Greene (2003), p.496). This version of LM is defined as

$$\text{LM} = nR_e^2 \qquad \text{LM} \sim \chi^2_m,$$

where R_e^2 is the coefficient of determination from that auxiliary regression.

Wald, LR and LM tests of linear restrictions in multiple regressions are asymptotically equivalent. Depending on how the information matrix is estimated, the test statistics and the associated conclusions will differ. In small samples the χ^2 distribution may lead to too many rejections of the true H_0. Alternatively, the t-statistic (for a single restriction) or the F-statistic can be used.

Example 9: We use the data and results from example 7 to test the restrictions β_0=0 and β_1=1 using OLS based LR and LM tests. Using the residuals from unrestricted and restricted residuals we obtain LR=3.904 with a p-value of 0.142. Regressing the residuals from the restricted model on $\boldsymbol{X}$ we obtain LM=0.402 with a p-value of 0.818. Details can be found in the file `uirp.xls`.

Example 10: We use the same data to estimate the Fama regression numerically by ML. The coefficients are virtually identical to the OLS estimates, while the standard errors (derived from the inverse of the information matrix) differ slightly from the OLS standard errors. To test the restriction β_0=0 and β_1=1 we use ML based Wald, LR and LM tests. All three tests agree that these restriction cannot be rejected with p-values ranging from 0.11 to 0.17. Details can be found in the file `uirp-ml.xls`.

Exercise 6: Use the data from exercise 4 (i.e. US dollar/British pound exchange and forward rates) to test the restriction β_0=0 and β_1=1 using OLS *and* ML based Wald, LR and LM tests.

1.6 Specifications

The specification of the regression equation has key importance for a successful application of regression analysis (in addition to a careful definition and selection of variables). The linearity assumption **AL** may appear to be a very strong restriction. However, $\boldsymbol{y}$ and $\boldsymbol{X}$ can be arbitrary functions of the underlying variables of interest. Thus, as we will show in this section, there exist several linear formulations to model a variety of practically relevant and interesting cases.

1.6.1 Log and other transformations

The **log-linear model**[28]

$$\ln y = \ln b_0 + b_1 \ln x_1 + \cdots + b_k \ln x_k + e = \hat{y}^{\ln} + e$$

corresponds to the multiplicative expression

$$y = b_0 x_1^{b_1} \cdots x_k^{b_k} \exp\{e\}.$$

In this model b_i is the (estimated) elasticity of y with respect to x_i:

$$\frac{\partial y}{\partial x_i} = b_i b_0 x_1^{b_1} \cdots x_i^{b_i - 1} \cdots x_k^{b_k} \exp\{e\} = b_i \frac{y}{x_i} \quad \Longrightarrow \quad b_i = \frac{\partial y}{y} \Big/ \frac{\partial x_i}{x_i}.$$

In other words, a change in x_i by p percent leads to a c.p. change in $\hat{y}$ by $p \cdot b_i$ percent. This implies that the change in $\hat{y}$ in response to a change in x_i depends on the levels of y and x_i, whereas these levels are irrelevant in the linear model.

To compute $\hat{y}$ using the fitted values $\hat{y}^{\ln}$ from the log-linear model we have to account for the properties of the lognormal distribution (see review 9). If the residuals from the log-linear model are (approximately) normal the expected value of y is given by

$$\hat{y}_i = \exp\{\hat{y}_i^{\ln} + 0.5 s_e^2\},$$

where s_e is the standard error of the residuals from the log-linear model. Note that these errors are given by

$$e_i = \ln y_i - \hat{y}_i^{\ln},$$

where $\hat{y}_i^{\ln}$ is the fitted value of $\ln y_i$. e_i is not equal to $\ln y_i - \ln \hat{y}_i$ because of Jensen's inequality $\ln \mathrm{E}[y] > \mathrm{E}[\ln y]$ (see review 1). The standard error of residuals s_e is an *approximation* for the magnitude of the *percentage* error $(y_i - \hat{y}_i)/\hat{y}_i$.

In the **semi-log model**

$$\ln y = b_0 + b_1 x_1 + \cdots + b_k x_k + e$$

[28] In section 1.6 we will formulate regression models in terms of estimated parameters since these are usually used for interpretations.

the expected c.p. *percentage* change in $\hat{y}$ is given by $b_i \cdot 100$, if x_i changes by one unit. More accurately, $\hat{y}$ changes by $(\exp\{b_i\}-1)\times 100$ percent.[29] This model is appropriate when the growth rate of y is assumed to be a linear function of the regressors. The chosen specification will mainly be driven by assumptions about the nature of the underlying relationships. However, taking logs is frequently also used to reduce or eliminate heteroscedasticity.

Another version of a semi-log model is

$$y = b_0 + b_1 \ln x_1 + \cdots + b_k \ln x_k + e.$$

Here, a one percent change in x_i yields a c.p. change in $\hat{y}$ of $0.01 \cdot b_i$ units.

The **logistic model**

$$\ln \frac{y}{1-y} = b_0 + b_1 x_1 + \cdots + b_k x_k + e \qquad 0 < y < 1$$

implies that $\hat{y}$ is s-shaped according to:

$$\hat{y} = \frac{\exp\{b_0 + b_1 x_1 + \cdots + b_k x_k\}}{1 + \exp\{b_0 + b_1 x_1 + \cdots + b_k x_k\}} = \frac{1}{1 + \exp\{-(b_0 + b_1 x_1 + \cdots + b_k x_k)\}}.$$

1.6.2 Dummy variables

Explanatory variables which are measured on a nominal scale (i.e. the variables are qualitative in nature) can be used in regressions after they have been recoded. A binary valued (0-1) dummy variable is defined for each category except one which constitutes the **reference category**. Suppose there are $m+1$ categories (e.g. industries or regions). We define m dummy variables d_i (d_i=1 if an observation belongs to category i and 0 otherwise). Note that defining a dummy for each category leads to an exact linear relationship among the regressors. If the model contains an intercept the sum of all dummies is equal to the first column of $\boldsymbol{X}$, and $\boldsymbol{X}$ will not have full rank. The coefficients δ_i in the regression model

$$\hat{y} = b_0 + b_1 x_1 + \cdots + \delta_1 d_1 + \cdots + \delta_m d_m$$

correspond to parallel shifts of the regression line (hyperplane). δ_i represents the change in $\hat{y}$ for a c.p. shift from the reference category to category i.

If categories have a natural ordering, an alternative definition of dummy variables may be appropriate. In this case all dummy variables $d_1,\ldots,d_j$ are set equal to 1 if an observation belongs to category j. Now δ_j represents the expected change in $\hat{y}$ for a c.p. shift from category $j-1$ to category j.

[29] The fitted value of y implied by the semi-log model $\ln y = b_0 + b_1 x + e$ is given by $\hat{y}_0 = \exp\{b_0 + b_1 x_0\}$ (ignoring the term $\exp\{0.5 s_e^2\}$) for an initial level x_0. If x_0 is increased by Δ we get $\hat{y}_1 = \exp\{b_0 + b_1(x_0+\Delta)\} = \exp\{b_0 + b_1 x_0\}\exp\{b_1\Delta\} = \hat{y}_0 \exp\{b_1\Delta\}$. Hence, $\hat{y}_1 - \hat{y}_0 = \hat{y}_0 \exp\{b_1\Delta\} - \hat{y}_0 = \hat{y}_0(\exp\{b_1\Delta\}-1)$, and $(\hat{y}_1 - \hat{y}_0)/\hat{y}_0 = \exp\{b_1\Delta\}-1$. The fitted value in log-terms is given by $\hat{y}_0^{\ln} = b_0 + b_1 x_0$. Increasing x_0 by Δ we get $\hat{y}_1^{\ln} = b_0 + b_1(x_0+\Delta)$, and $\hat{y}_1^{\ln} - \hat{y}_0^{\ln} = b_1\Delta$.

1.6.3 Interactions

Dummy variables cannot be used to model changes in the slope (e.g. differences in the propensity to consume between men and women). If the slope is assumed to be different among categories the following specification can be used:

$$\hat{y} = b_0 + b_1 x_1 + b_2 d + b_3 d x_1.$$

The product dx_1 is an **interaction term**. If d=0 this specification implies $\hat{y}$=b_0+$b_1 x_1$, and if d=1 it implies $\hat{y}$=(b_0+b_2)+(b_1+b_3)x_1. Thus, the coefficient b_3 measures the expected c.p. change in the slope of x_1 when switching categories.

It is important to note that the presence of an interaction term changes the 'usual' interpretation of the coefficients associated with the components of the interaction. First, the coefficient b_1 of x_1 must be interpreted as the slope of the reference category (for which d=0). Second, the coefficient b_2 of the dummy variable is not the expected c.p. difference between the two categories anymore (except for x_1=0). Now the difference depends on the level of x_1. Even if x_1 is held constant the difference in $\hat{y}$ when changing from d=1 to d=0 is given by b_2+$b_3 x_1$.

Interactions are not confined to using dummy variables but can be based on two 'regular' regressors. The equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

implies a change in the slope of x_1 that depends on the level of x_2 and vice versa. To simplify the interpretation of the coefficients it is useful to evaluate $\hat{y}$ for typical values of one of the two variables (e.g. using $\bar{x}_2$ and $\bar{x}_2 \pm s_{x_2}$).

The presence of an interaction has an important effect on the stand-alone coefficients of the components of the interaction term. In the equation above, neither b_1 nor b_2 can be interpreted meaningfully without fixing specific levels of x_1 and x_2. In particular, it is not possible to compare the coefficients from the equation above to those from the regression

$$\hat{y} = b'_0 + b'_1 x_1 + b'_2 x_2.$$

For example, b'_1 is only comparable to b_1 if x_2=0. However, if x_2 is a strictly positive variable with a typical range far above zero, b_1 and b'_1 will have very different orders of magnitude.

This aspect is also relevant in the context of testing whether the slope with respect to x_1 is zero. In a regression which includes an interaction term the slope with respect to x_1 is given by (b_1+$b_3 x_2$)x_1. Hence, the 'effective' slope coefficient to be tested depends on the level of x_2. In other words, every level of x_2 implies a different t-statistic and p-value. While a regression including interaction terms allows for testing whether slopes depend on the levels of the regressors in the interaction term (i.e. whether b_3 differs from zero), it does not allow for directly testing the slopes.

If interactions are defined using logs of variables such as in the following so-called **translog** model

$$\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + b_3 \ln x_1 \ln x_2 + e$$

the conditional expectation of y is given by

$$\hat{y} = b_0 x_1^{b_1 + b_3 \ln x_2} x_2^{b_2} \exp\{0.5 s_e^2\}.$$

This implies that a c.p. change of x_2 by p percent leads to an expected change of the elasticity b_1 by pb_3. However, if $\hat{y}$ is defined as

$$\hat{y} = b_0 x_1^{b_1 + b_3 x_2} x_2^{b_2} \exp\{0.5 s_e^2\},$$

it is necessary to estimate the model

$$\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + b_3 (\ln x_1) x_2 + e.$$

In this case a c.p. change of x_2 by one unit leads to an expected change of the elasticity b_1 by b_3 units.

1.6.4 Difference-in-differences

As a special case, an interaction can be defined as the product of a time-dummy and another dummy, identifying group membership. (Quasi) natural experiments are typical situations where this is an appropriate specification. The purpose of the analysis is to find out about the effect of a certain stimulus or a special event (e.g. a policy or legal change, a crisis, an announcement, etc.). Such experiments are characterized by a treatment and a control group: the treatment group consists of those objects under study which are subject to the special event, whereas the remaining, unaffected subjects constitute the control group.

The effect size can be estimated by comparing the means of the two groups before and after the event. This difference (among groups) of the difference (over time) – hence, difference-in-differences (or, diff-in-diff) – can be simply estimated from the coefficient b_3 in the regression

$$\hat{y} = b_0 + b_1 T + b_2 d + b_3 T d,$$

where T denotes a time-dummy (i.e. being 0 before and 1 after the event), and d is the dummy distinguishing the treatment (d=1) and the control (d=0) group. Note that b_0 is the average of y for T=0 and d=0 (i.e. the control group before the event), b_1 estimates the average change in y over the two time periods for d=0, b_2 estimates the average difference between treatment and control for T=0. b_3 is the estimate which is of primary interest in such studies.

Note that the simple formulation above is only appropriate if no other regressors need to be accounted for. If this is not the case, the model has to be extended as follows:

$$\hat{y} = b_0 + b_1 T + b_2 d + b_3 T d + \mathbf{Xb}.$$

As soon as the term $\mathbf{Xb}$ is included in the specification, the coefficient b_3 is still the main object of interest, however it is not a difference of sample averages any more, but has the corresponding ceteris paribus interpretation.

The diff-in-diff approach can be used to account for a so-called *selection bias*. For example, when assessing the effect of an MBA on salaries, people who choose to do an MBA may already have higher salaries than those who do not. Thus, the assignment to treatment and control group is not random but depends on (existing or expected) salaries. This problem of so-called *self-selection* results whenever subjects enter the treatment sample for reasons which are related to the dependent variable.

The appropriateness of the difference-in-differences approach rests on the **parallel-trends assumption**. Absent the effect under study, the dependent variable of the two groups must not have different "trends" (i.e. must not have differing slopes with respect to time). If this assumption is violated, the effect is over- or underestimated (because of diverging or converging trends), and partially but falsely attributed to the treatment. In the MBA-salary example this assumption is violated, when the salaries of people who choose to do an MBA already increase more quickly than the salaries of those, who do not.

Note that the interaction term Td already accounts for different slopes with respect to time. Therefore, it is impossible to separate the effect under study from possibly different trends of the two groups which have nothing to do with the effect under study.

1.6.5 Example 11: Hedonic price functions

Hedonic price functions are used to define the implicit price of key attributes of goods as revealed by their sales price. We use a subset of a dataset used in Wooldridge (2003, p.194)[30] consisting of the price of houses (y), the number of bedrooms (x_1), the size measured in square feet (x_2) and a dummy variable to indicate the style of the house (x_3) (see `hedonic.wf1`). We estimate the regression equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \epsilon,$$

where the interaction-term $x_1 x_2$ is used to model the importance of the number of bedrooms *depending on* the size of the house. The underlying hypothesis is that additional bedrooms in large houses have a stronger effect on the price than in small houses (i.e. it is expected that $\beta_4 > 0$). The estimated equation is

$$\underset{}{\hat{y} =} \underset{(0.034)}{199.45} - \underset{(0.072)}{45.51}\, x_1 + \underset{(0.575)}{0.025}\, x_2 + \underset{(0.191)}{20.072}\, x_3 + \underset{(0.014)}{0.026}\, x_1 x_2.$$

The interaction term is significant and has the expected sign. To facilitate the model's interpretation it is useful to evaluate the regression equation using typical values for one of the variables in the interaction term. Mean and standard deviation of size (x_2) are 2013.7 and 578.7. We can formulate equations for the expected price for small, average and large houses as a function of style (x_3) and the number of bedrooms (x_1):

$$\begin{aligned}
x_2{=}1500: &\quad \hat{y} = 236.48 + 20.072 x_3 - 6.63 x_1 \\
x_2{=}2000: &\quad \hat{y} = 248.82 + 20.072 x_3 + 6.33 x_1 \\
x_2{=}2500: &\quad \hat{y} = 261.16 + 20.072 x_3 + 19.29 x_1.
\end{aligned}$$

[30] Source: Go to the book companion site of Wooldridge (2003) at `http://www.cengage.com/` (latest edition), click on "Data Sets", download one of the zip-files and choose "`HPRICE1.*`".

This shows that a regression equation with interactions can be viewed as a model with varying intercept and slope, where this variation depends on one of the interaction variables. The first equation shows that additional bedrooms in small houses lead to a c.p. drop in the expected price (probably because those bedrooms would be rather small and thus unattractive). We find a positive effect of bedrooms in houses with (above) average size.

1.6.6 Example 12: House price changes induced by siting decisions

We consider a part of the dataset used by Kiel and McClain (1995)[31] who study the impact of building an incinerator on house prices. Prices are available for 1978, *before* any rumors about potentially building an incinerator, and 1981 when its construction began.[32] The purpose of the analysis is to quantify the effect of building a new incinerator on house prices. For simplicity we first ignore the control variables considered in the original study.

Running a regression of house prices on a dummy indicating whether a house is near[33] the incinerator (d=1) makes no sense. If the incinerator was built in an area where house prices were already (relatively) low, the coefficient would not estimate the impact of the incinerator. In other words, the sample suffers from a selection bias. To determine the effect of the incinerator we must compare the average house prices before and after rumors by distinguishing houses near and far from the incinerator.

A non-regression based approach to estimate the effect is shown in `diff-in-diff.xlsx`. More specifically, we find that prices for houses far from the incinerator (control group) have increased by 18790.3 from 1978 to 1981, whereas prices of houses nearby (treatment group) have increased by only 6926.4. The difference of these differences between treatment and control group is –11863.9; i.e. houses nearby have increased less strongly – an effect that can be attributed to the incinerator. Alternatively, one could compare the difference in prices for the two types of houses in 1978 (–18824.4; i.e. houses nearby are cheaper) and 1981 (–30688.3; i.e. houses nearby are now much cheaper). This also results in an estimated effect of –11863.9. Simply comparing averages leads to the same results as a regression with dummy variables and an interaction with time, *provided* no further regressors (i.e. control variables) are used.

Adding further regressors (see `diff-in-diff.R` or `diff-in-diff.wf1`) accounts for various (additional) features of the houses to be compared. This avoids any biases from ignoring other effects (see section 1.6.7). The revised estimate turns out to be –14177.9. This estimate cannot be simply derived from comparing averages, but measures the effect of the incinerator on house prices after controlling for other features of houses (i.e. ceteris paribus). Note that adding further regressors not only controls for additional features but (usually) also improves the goodness of fit. Thereby, standard errors of coefficients are reduced, and statistical inference is enhanced.

[31] Source: Go to the book companion site of Wooldridge (2003) at `http://www.cengage.com/` (latest edition), click on "Data Sets", download one of the zip-files, and choose "`KIELMC.*`'.

[32] Note that this is a pooled dataset, i.e. prices for *different* houses are considered in the two years.

[33] We ignore the available measure for the *distances* between houses and the incinerator to be able to illustrate the difference-in-differences approach.

1.6.7 Omitted and irrelevant regressors

Relevant variables may have been omitted from a regression equation by mistake or because of a lack of data. Omitting relevant variables may have serious consequences. To investigate the potential effects we suppose that the correctly specified model is given by

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{19}$$

but the model is estimated without $\boldsymbol{X}_2$. The OLS estimate of $\boldsymbol{\beta}_1$ is given by

$$\boldsymbol{b}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}.$$

We rewrite (19) as $\boldsymbol{y}$=$\boldsymbol{X}_1\boldsymbol{\beta}_1$+$\boldsymbol{\epsilon}_1$ where $\boldsymbol{\epsilon}_1$=$\boldsymbol{X}_2\boldsymbol{\beta}_2$+$\boldsymbol{\epsilon}$ and substitute for $\boldsymbol{y}$ in the equation for $\boldsymbol{b}_1$ to obtain

$$\begin{aligned}\boldsymbol{b}_1 &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1)\\ &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\epsilon}_1\\ &= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\epsilon}_1.\end{aligned}$$

The expectation of $\boldsymbol{b}_1$ is given by

$$\mathrm{E}[\boldsymbol{b}_1] = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\mathrm{E}[\boldsymbol{X}_1'\boldsymbol{\epsilon}_1].$$

This shows that assumption **AX** is violated since

$$\mathrm{E}[\boldsymbol{X}_1'\boldsymbol{\epsilon}_1] = \mathrm{E}[\boldsymbol{X}_1'(\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon})] = \boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2$$

is non-zero unless $\boldsymbol{X}_1'\boldsymbol{X}_2$=$\mathbf{0}$ (i.e. all elements of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are uncorrelated) or $\boldsymbol{\beta}_2$=$\mathbf{0}$. Thus $\boldsymbol{b}_1$ is *biased* and *inconsistent* if there are omitted regressors which are correlated with included regressors. The expected value of $\boldsymbol{b}_1$ is given by the so-called **omitted variable formula**

$$\mathrm{E}[\boldsymbol{b}_1] = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2. \tag{20}$$

The formula shows that the bias depends on the term that is multiplied with $\boldsymbol{\beta}_2$. This term is equal to the coefficients from a regression of omitted regressors $\boldsymbol{X}_2$ on included regressors $\boldsymbol{X}_1$.

As a further consequence, the standard error of regression s_e and the standard errors of $\boldsymbol{b}_1$ will also be biased. Thus, statistical tests about $\boldsymbol{\beta}_1$ are not meaningful. Usually, s_e will be too high, and se[$\boldsymbol{b}_1$] can be lower than if $\boldsymbol{X}_2$ is included in the regression (see Greene, 2000, p.337).

In the simple case of only two regressors, where the correct equation is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{21}$$

and x_2 is omitted from the estimated regression, the bias is given by

$$\mathrm{E}[b_1] - \beta_1 = \beta_2 \frac{\mathrm{cov}[x_1 x_2]}{\mathrm{V}[x_1]}.$$

If x_1 and x_2 are uncorrelated, the estimate of β_0 is still biased[34] and inconsistent. In this case the estimate b_1 is unbiased, but the standard error of b_1 is too large (see Kmenta, 1971, p.392).

The estimated coefficients in the reduced regression for a specific sample can also be computed from the omitted variable formula. If $\boldsymbol{b}_1^{\text{full}}$ denotes the subset of coefficients from the full model corresponding to $\boldsymbol{X}_1$ and $\boldsymbol{b}_2^{\text{full}}$ corresponds to $\boldsymbol{X}_2$, the coefficients in the reduced model are given by

$$\boldsymbol{b}_1 = \boldsymbol{b}_1^{\text{full}} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{b}_2^{\text{full}}.$$

The omission of explanatory variables cannot be detected by a statistical test. Indirect (but possibly ambiguous) evidence may be obtained from the analysis of residuals. Note that the correlation between the residuals of the OLS regression $\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}$ and $\boldsymbol{X}$ *cannot* be used to detect this problem. It is one of the implications of the LS principle that this correlation is *always* zero (see section 1.1.2). Proxy variables may be used instead of actually required, but unavailable regressors. Proxies should be highly correlated with the unavailable variables, but one can only make assumptions about this correlation. The negative consequences of omitted variables can be mitigated or eliminated using instrumental variable estimates (see section 1.9), or panel data analysis[35].

Including irrelevant variables in the model leads to inefficient but unbiased and consistent estimates. The inefficiency can be shown by considering an alternative definition of the variance of the OLS estimate b_j (see Greene, 2003, p.57)

$$\mathrm{V}[b_j] = \frac{\sigma^2}{(1-R_j^2)\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}, \tag{22}$$

where R_j^2 is the R^2 from a regression of regressor j on all other regressors (including a constant term). This definition shows that c.p. the variance of b_j will increase with the correlation between variable x_j and other regressors. This fact is also known as the **multicollinearity** problem which becomes relevant if R_j^2 is very close to one.

Suppose that the correct model is given by $y=\beta_0+\beta_1 x_1+\epsilon$ but the irrelevant variable x_2 (i.e. $\beta_2=0$) is added to the estimated regression equation. Denote the estimate of β_1 from the overfitted model by $\tilde{b}_1$. The variance of b_1 (from the correct regression) is given by (6), p.11 whereas $\mathrm{V}[\tilde{b}_1]$ is given by (22). Thus, unless x_1 and x_2 are uncorrelated *in the sample*, $\mathrm{V}[\tilde{b}_1]$ will be larger than necessary (i.e. larger than $\mathrm{V}[b_1]$).

Exact multicollinearity holds when there are exact linear relationships among some regressors (i.e. $\boldsymbol{X}$ does not have full rank). This can easily be corrected by eliminating redundant regressors (e.g. superfluous dummies). Typical signs of strong (but not exact) multicollinearity are wrong signs or implausible magnitudes of coefficients, as well as a strong sensitivity to changes in the sample (dropping or adding observations). The inflating effect on standard errors of coefficients may lead to cases where several coefficients are individually insignificant, but eliminating them (jointly) from the model leads to a significant drop in R^2 (based on an F-test).

[34] Exception: the mean of x_2 is zero.

[35] For an introduction to the principles of panel data analysis, see Wooldridge (2003), chapters 13 and 14).

The consequences of including irrelevant regressors (inefficiency) have to be compared to the consequences of omitting relevant regressors (bias and inconsistency). We hesitate to formulate a general recommendation, but it is worth while asking "What is the point of estimating a parameter more precisely if it is potentially biased?"

1.6.8 Selection of regressors

The search for a correct specification of a regression model is usually difficult. The selection procedure can either start from a model with one or only a few explanatory variables, and subsequently add variables to the equation (the *specific to general* approach). Alternatively, one can start with a large model and subsequently eliminate insignificant variables. The second approach (*general to specific*) is preferable, since the omission of relevant variables has more drawbacks than the inclusion of irrelevant variables. In any case, it is strongly recommended to select regressors on the basis of a sound theory or a thorough investigation of the subject matter. A good deal of common sense is always useful.

The following guidelines can be used in the model selection process:

1. The selection of variables must not be based on simple correlations between the dependent variable and preselected regressors. Because of the potential bias associated with omitted variables any selection should be done in the context of estimating *multiple* regressions.

2. If the p-value of a coefficient is above the significance level this indicates that the associated variable can be eliminated. If several coefficients are insignificant one can start by eliminating the variable with the largest p-value and re-estimate the model.

3. If the p-value indicates elimination but the associated variable is considered to be of key importance theoretically, the variable should be kept in the model (in particular if the p-value is not far above the significance level). A failure to find significant coefficients may be due to insufficient data or a random sample effect (bad luck).

4. Statistical significance alone is not sufficient. There should be a very good reason for a variable to be included in a model and its coefficient should have the expected sign.

5. Adding a regressor will always lead to an increase of R^2. Thus, R^2 is not a useful selection criterion. A number of model selection criteria have been defined to facilitate the model choice in terms of a compromise between goodness of fit and the **principle of parsimony**. The adjusted R^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{s_e^2}{s_y^2} \qquad s_e^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}$$

is a criterion that can be used for model selection. Note that removing a variable whose t-statistic is less than 1 leads to an increase of $\bar{R}^2$ (R^2 always drops if a regressors is removed!), and a decrease of the standard error of regression (s_e). It has been found, however, that $\bar{R}^2$ puts too little penalty on the loss in degrees of freedom. Alternative criteria are **Akaike's information criterion**

$$\text{AIC} = -\frac{2\ell}{n} + \frac{2K}{n}$$

and the **Schwarz criterion**[36]

$$\text{SC} = -\frac{2\ell}{n} + \frac{K \ln n}{n},$$

where $\ell = -0.5n[1+\ln(2\pi)+\ln(\boldsymbol{e}'\boldsymbol{e}/n)]$. We finally note that model selection criteria must never be used to compare models with different dependent variables (e.g. to compare linear and log-linear models).

Exercise 7: Consider the data on the salary of 208 employees in the file `salary.wf1`[37]. Estimate and choose a regression model for salary using available information such as gender, education level, experience and others. Note that EDUC is a *categorical* variable measuring the education level in terms of degrees obtained (1=finished high school, 2=finished some college courses, 3=obtained a bachelor's degree, 4=took some graduate courses, 5=obtained a graduate degree). Use model formulations which allow you to test for gender-specific payment behavior.

[36] These are the definitions of AIC and SC used in EViews. Alternatively, the first term in the definition of AIC and SC can be replaced by $\ln(\boldsymbol{e}'\boldsymbol{e}/n) = \ln(\tilde{s}_e^2)$.

[37] Source: Albright et al. (2002, p.686), Example 13.3.

1.7 Regression diagnostics

The purpose of diagnostic checking is to test whether important assumptions of regression analysis hold. In this subsection we will present some frequently applied tests, discuss some implications associated with violated assumptions, and provide simple remedies to correct for negative consequences. Note that a model that passes diagnostic tests need not necessarily be correctly specified.

1.7.1 Non-normality

The **Jarque-Bera test** is based on the null hypothesis of a normal distribution and takes skewness S and kurtosis U into account:

$$\text{JB} = \frac{n-K}{6}\left[S^2 + \frac{1}{4}(U-3)^2\right] \qquad \text{JB} \sim \chi^2_2,$$

where

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^3}{\tilde{s}^3} \qquad U = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^4}{\tilde{s}^4}.$$

If OLS residuals are not normally distributed, OLS estimates are unbiased and consistent, but *not efficient* (see Kmenta (1971), p.248). There exist other estimators with greater accuracy (of course, only if the correct (or a more suitable) distribution is used in those estimators). In addition, the t-statistics for significance testing are not appropriate. However, this is only true in small samples, and when the deviation from the normal distribution is 'strong'. A failure to obtain normal residuals in a regression may indicate missing regressors and/or other specification problems (although the specific kind of problem cannot be easily inferred). At any rate, normality of the dependent variables is not a requirement of OLS (as can be derived from sections 1.2.1 and 1.2.2).

> **Example 13:** We use the data from example 8 and estimate the regression equation by OLS. Details can be found in the file `investment quarterly.wf1`. The distribution of residuals is positively skewed (0.25). This indicates an asymmetric distribution whose right tail is slightly longer than the left one. The kurtosis is far greater than three (5.08) which indicates more concentration around the mean than a normal distribution. JB is 38.9 with a p-value of zero. This clearly indicates that we can reject H_0 and we conclude that the residuals are not normally distributed.

1.7.2 Heteroscedasticity

Heteroscedasticity means that the variance of disturbances is not constant across observations

$$\mathrm{V}[\epsilon_i] = \sigma_i^2 = \omega_i \sigma^2 \quad \forall i,$$

and thus violates assumption **AH**. To analyze the implications of heteroscedasticity we assume that the covariance matrix is diagonal

$$\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \sigma^2 \boldsymbol{\Omega} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{23}$$

If the variance of ϵ is not given by $\sigma^2 \boldsymbol{I}$ but $\sigma^2 \boldsymbol{\Omega}$, the model is a so-called **generalized least squares** (**GLS**) model.

It can be shown that the finite sample properties of the OLS estimator are not affected if only **AH** is violated (see Greene (2003), section 10.2.1). However, the covariance of $\boldsymbol{b}$ is not given by (5), p.11 but is given by

$$\mathrm{V}[\boldsymbol{b}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{\Omega})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{24}$$

Provided that $\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}/n$ converges to a positive definite matrix it can be shown that in the presence of heteroscedasticity the OLS estimator $\boldsymbol{b}$ is unbiased, consistent and asymptotically normal (see Greene (2003), section 10.2.2):

$$\boldsymbol{b} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\boldsymbol{Q}^{-1}\bar{\boldsymbol{Q}}_n\boldsymbol{Q}^{-1}\right), \tag{25}$$

where $\boldsymbol{Q}$ is defined in (11), p.23 and

$$\bar{\boldsymbol{Q}}_n = \operatorname{plim} \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}.$$

However, the OLS estimator $\boldsymbol{b}$ is *inefficient* since it does not use all the information available in the sample. The estimated standard errors of $\boldsymbol{b}$ are biased and the associated t- and F-statistics are incorrect. For instance, if σ_i^2 and a regressor x_j are positively correlated, the bias in the standard error of b_j is negative (see Kmenta (1971), p.256). Depending on the correlation between the heteroscedasticity and the regressors (and their cross-products) the consequences may be substantial (see Greene (2000), p.502-505).

The **Breusch-Pagan test** for heteroscedasticity is based on an auxiliary regression of e_i^2 on a constant and the regressors. Under the null of homoscedasticity we can use the R^2 from this regression to compute the test statistic $nR^2 \sim \chi_k^2$ (k is the number of regressors *excluding* the constant). The **White-test** for heteroscedasticity is based on regressing e_i^2 against a constant, the regressors and their squares. In a more general version of the test the cross products of regressors may be added, too. Under the null of homoscedasticity the

test statistic is also $nR^2 \sim \chi^2_k$, where k is the number of regressors in the auxiliary regression excluding the constant. The advantage of the White-test is that no assumptions about the type of heteroscedasticity are required. On the other hand, rejecting H_0 need not be due to heteroscedasticity but may indicate other specification errors (e.g. omitted variables).

In section 1.8 we will present estimators that make use of some knowledge about $\boldsymbol{\Omega}$. If no such information is available the OLS estimator may still be retained. However, to improve statistical inference about coefficients the estimated standard errors can be corrected using the **White heteroscedasticity consistent** (WHC) estimator

$$\widehat{\mathrm{aV}}[\boldsymbol{b}] = \frac{n}{n-K}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\,. \tag{26}$$

Example 14: We use a dataset[38] on hourly earnings (y), employment duration (x_1) and years of schooling (x_2) (n=49) (see `earnings.wf1`). A plot of the residuals from the estimated regression (t-statistics in parenthesis)

$$\underset{}{\ln y} = \underset{(6.1)}{1.22} + \underset{(4.4)}{0.027}\,x_1 + \underset{(3.6)}{0.126}\,x_2 + e$$

against x_1 shows a strong increase in the variance of e. The White-test statistic 23.7 based on the regression (p-values in parenthesis)

$$e^2 = \underset{(0.5)}{0.1} - \underset{(0.15)}{0.022}\,x_1 + \underset{(0.008)}{0.0009}\,x_1^2 + \underset{(0.09)}{0.12}\,x_2 - \underset{(0.08)}{0.019}\,x_2^2 + u \qquad R^2 = 0.484$$

is highly significant (the p-value is very close to zero) and we can firmly reject the homoscedasticity assumption. The t-statistics based on the WHC standard errors are 9.7, 4.0 and 4.9, respectively. Thus, in this example, the conclusions regarding the significance of coefficients are not affected by the heteroscedasticity of residuals.

Exercise 8: Test the residuals from the models estimated in exercise 7 for non-normality and heteroscedasticity.

[38] Source: Thomas (1997), p.293.

1.7.3 Autocorrelation

Autocorrelation (or **serial correlation**) is only relevant in case of time series data. It means that consecutive disturbances are correlated – which violates assumption **AH**. For instance, if the dependent variable is subject to seasonality (e.g. a monthly time series which has local peaks during the months of summer and troughs during winter) which is not accounted for by the regressors, the residuals e_t and e_{t-12} will be correlated.

To analyze the implications of autocorrelation we assume that the covariance matrix of disturbances is given by

$$\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}, \tag{27}$$

where ρ_ℓ is the (auto)correlation between ϵ_t and $\epsilon_{t-\ell}$. It can be shown (see Greene (2003), section 10.2.2) that under this assumption, autocorrelated disturbances have the same consequences as heteroscedasticity. The OLS estimator $\boldsymbol{b}$ is unbiased, consistent, asymptotically normal as in (25), but inefficient. This implies that the standard errors of the coefficients are *biased*. For instance, if the majority of autocorrelations is positive the standard errors are too small (see Kmenta (1971), p.273).

Autocorrelations can be estimated from the sample by

$$r_\ell = \frac{1}{\boldsymbol{e}'\boldsymbol{e}} \sum_{t=\ell+1}^{n} e_t e_{t-\ell},$$

and tests for the significance of individual autocorrelations can be based on

$$r_\ell \overset{a}{\sim} \mathrm{N}(-1/n, 1/n).$$

The asymptotic properties of r_ℓ hold if the disturbances $\boldsymbol{\epsilon}$ are uncorrelated (see Chatfield (1989), p.51). The **Ljung-Box Q-statistic**

$$Q_p = n(n+2) \sum_{\ell=1}^{p} \frac{r_\ell^2}{n-\ell} \qquad Q_p \sim \chi_p^2$$

can be used as a joint test for all autocorrelations up to lag p. The **Durbin-Watson test** $\mathrm{DW} \approx 2(1-r_1)$ has a long tradition in econometrics. However, it only takes the autocorrelation at lag 1 into account and has other conceptual problems; e.g. it is not appropriate if the lagged dependent variable is used as a regressor (see Greene (2003), p.270).

The **Breusch-Godfrey test** is based on an auxiliary regression of e_t on p lagged residuals and the original regressors. Under the null of no autocorrelation we can use the R^2 from this regression to compute the test statistic $nR^2 \sim \chi_p^2$.

Similar to the WHC estimator the **Newey-West** (HAC) estimator can be used to account for residual autocorrelation without changing the model specification. It is a covariance

estimator that is consistent in the presence of both heteroscedasticity and autocorrelation (hence HAC) of unknown form. It is given by

$$\widehat{\mathrm{aV}}[\boldsymbol{b}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\hat{\boldsymbol{\Omega}}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

where

$$\hat{\boldsymbol{\Omega}} = \frac{n}{n-K}\left(\sum_{t=1}^{n} e_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' + \sum_{j=1}^{q} w_j \sum_{t=j+1}^{n} \boldsymbol{x}_t e_t e_{t-j} \boldsymbol{x}_{t-j}' + \boldsymbol{x}_{t-j} e_{t-j} e_t \boldsymbol{x}_t'\right).$$

w_j=1−j/(q+1), and the truncation lag q determines how many autocorrelations are taken into account. Newey and West (1987) suggest to set q=4(n/100)$^{2/9}$.

We now take a closer look at implications of autocorrelated residuals and consider the model[39]

$$y_t = \beta x_t + u_t \tag{28}$$

$$u_t = \rho u_{t-1} + \epsilon_t \qquad |\rho| < 1 \quad \epsilon_t \sim \text{i.i.d.}$$

The first equation may be viewed as being incorrectly specified, as we are going to show now. The second equation for the autocorrelated residuals u_t is a so-called **first order autoregression AR(1)**. Substituting u_t into the first equation (y_t=βx_t+ρu_{t-1}+ϵ_t) and using u_{t-1}=y_{t-1}−βx_{t-1}, we obtain

$$y_t = \rho y_{t-1} + \beta x_t - \rho\beta x_{t-1} + \epsilon_t.$$

This shows that the autocorrelation in u_t can be viewed as a result of missing lags in the original equation. If we run a regression *without* using y_{t-1} and x_{t-1}, we have an omitted variables problem. The coefficient δ of the incomplete regression y_t=δx_t+ν_t is given by

$$\delta = \frac{\sum y_t x_t}{\sum x_t^2} = M \sum y_t x_t \qquad M = \frac{1}{\sum x_t^2}.$$

Substituting for y_t from the complete regression we obtain

$$\begin{aligned}\mathrm{E}[\delta] &= \mathrm{E}[M \textstyle\sum x_t(\rho y_{t-1} + \beta x_t - \rho\beta x_{t-1} + \epsilon_t)] \\ &= \rho\mathrm{E}[M \textstyle\sum x_t y_{t-1}] + \beta\mathrm{E}\big[M \textstyle\sum x_t^2\big] - \rho\beta\mathrm{E}[M \textstyle\sum x_t x_{t-1}] + \mathrm{E}[M \textstyle\sum x_t \epsilon_t]\,.\end{aligned}$$

To simplify this relation it is useful to write the AR(1) equation as

$$u_t = \rho^t u_0 + \sum_{i=0}^{t-1} \rho^i \epsilon_{t-i},$$

[39] For the sake of simplicity we consider only a single regressor x_t, and assume that y_t and x_t have mean zero.

which implies that equation (28) can be written as ($\rho^t u_0$ vanishes for large t since $|\rho|<1$)

$$y_t = \beta x_t + \sum_{i=0}^{t-1} \rho^i \epsilon_{t-i}.$$

We also make use of $\mathrm{E}[x_t \epsilon_{t-i}]=0$ ($\forall i \geq 0$), and note that the autocorrelation of x_t is given by $\rho_x = M \sum x_t x_{t-1}$. We find that δ is unbiased since its expectation is given by

$$\begin{aligned} \mathrm{E}[\delta] &= \rho \mathrm{E}[M \textstyle\sum x_t (\beta x_{t-1} + \sum \rho^i \epsilon_{t-i})] + \beta - \rho\beta\rho_x \\ &= \rho\beta\rho_x + \beta - \rho\beta\rho_x = \beta. \end{aligned}$$

Thus, despite the incomplete regression and the presence of autocorrelated residuals, we obtain unbiased estimates.

We now add a lagged dependent variable to equation (28). From section 1.2 we know that a lagged dependent variable leads to biased estimates. However, the estimates are consistent provided that assumptions $\overline{\mathbf{AX}}$ and $\mathbf{AR}_t$ hold, and the disturbances ϵ_t are autocorrelated (see section 1.3.3). We now investigate what happens if the disturbances are autocorrelated. We consider the model

$$\begin{aligned} y_t &= \phi y_{t-1} + \beta x_t + u_t \\ u_t &= \rho u_{t-1} + \epsilon_t \qquad |\rho| < 1 \quad \epsilon_t \sim \text{i.i.d.}, \end{aligned}$$

which can be written as

$$y_t = (\phi + \rho) y_{t-1} - \phi\rho y_{t-2} + \beta x_t - \rho\beta x_{t-1} + \epsilon_t.$$

Suppose we run a regression *without* using y_{t-2} and x_{t-1}. From the omitted variable formula (20) we know that the resulting bias depends on the coefficients of omitted regressors ($\boldsymbol{\beta}_2 = [-\phi\rho \;\; -\rho\beta]'$ in the present case), and the matrix of coefficients from regressing y_{t-2} and x_{t-1} on included regressors. This matrix will be proportional to the following matrix (i.e. we ignore the inverse of the matrix associated with included regressors):

$$\begin{pmatrix} \sum y_{t-1} y_{t-2} & \sum y_{t-1} x_{t-1} \\ \sum x_t y_{t-2} & \sum x_t x_{t-1} \end{pmatrix}.$$

The elements in the first row will be non-zero (if $\phi \neq 0$ and $\beta \neq 0$), and thus the estimated coefficient of y_{t-1} is biased. It is more difficult to say something general about the first element in the second row, but autocorrelation in x_t leads to a bias in β. Greene (2003, p.266) considers the simplified case $\beta=0$, and states that the probability limit of the estimated coefficient of y_t on y_{t-1} alone is given by $(\phi+\rho)/(1+\phi\rho)$. Although the consequences of serial correlation in ϵ_t in a regression with a lagged dependent variable have to be determined on a case-by-case basis, coefficients of such regressions are generally biased an inconsistent.

Example 15: We consider the data set analyzed by Coen et al. (1969) who formulate a regression for the Financial Times index (y_t) using the UK car production index (p_t) lagged by six quarters and the Financial Times commodity index (c_t) lagged by seven quarters as regressors. Details can be found in the files `coen.R` or `coen.wf1`. These lags where found by "*graphing the series on transparencies and then superimposing them* (p.136)". The estimated equation is (all p-values are very close to zero; we report t-statistics below coefficients for later comparisons)

$$y_t = \underset{(11.6)}{653} + \underset{(14.1)}{0.47}\, p_{t-6} - \underset{(9.9)}{6.13}\, c_{t-7} + e_t. \tag{29}$$

The Coen et al. study has raised considerable debate (see the discussion in their paper and in Granger and Newbold, 1971, 1974) because the properties of the residuals had not been thoroughly tested. As it turns out DW=0.98, and the Breusch-Godfrey test statistic using p=1 is 12.4 with a p-value below 0.001. This is evidence of considerable autocorrelation. In fact, using Newey-West HAC standard errors, the t-statistics are reduced to 11.4 and 7.6, respectively.

Stock prices or indices are frequently claimed to follow a random walk (see section 2.3) y_t=y_{t-1}+ϵ_t (ϵ_t∼i.i.d.). Thus we add the lagged dependent variable y_{t-1} to Coen et al.'s equation and find

$$y_t = \underset{(4.1)}{276} + \underset{(6.9)}{0.661}\, y_{t-1} + \underset{(2.3)}{0.127}\, p_{t-6} - \underset{(3.8)}{2.59}\, c_{t-7} + e_t. \tag{30}$$

The residuals in this equation are not autocorrelated which indicates that the estimates are consistent (to the extent that no regressors have been omitted). The coefficients and the t-statistics of p_{t-6} and c_{t-7} are considerably lower than before. It is not straightforward to test whether the coefficient of y_{t-1} is equal to one for reasons explained in section 2.3.3. In sum, our results raise some doubt about the highly significant lagged relationships found by Coen et al.

Example 16: We briefly return to example 7 where we have considered tests of the UIRP based on one-month forward rates and monthly data. Since forward rates are also available for other maturities, this provides further opportunities to test the UIRP. We use the three-month forward rate F_t^3 for which the UIRP implies

$$\mathrm{E}_t[\ln S_{t+3}] = \ln F_t^3.$$

This can be tested by running the regression

$$s_t - s_{t-3} = \beta_0 + \beta_1(f_{t-3}^3 - s_{t-3}) + \epsilon_t.$$

The estimated regression is[40]

$$s_t - s_{t-3} = \underset{(1.76)}{0.01} + \underset{(1.86)}{0.994}\,(f_{t-3}^3 - s_{t-3}) + e_t \qquad R^2 = 0.0123.$$

Before we draw any conclusions from this regression it is important to note that the observation frequency need not (and in the present case does not) conform to the maturity. s_t−s_{t-3} is a three-month return (i.e. the sum of three consecutive monthly returns). This introduces autocorrelation in three-month returns even though the monthly returns are not serially correlated (similar to section 1.8.3). This is known as the **overlapping samples problem**. In fact, the residual autocorrelations at lags 1 and 2 are highly significant (and positive), and the p-value of the Breusch-Godfrey

[40] See file `uirp.wf1` for details.

test is zero. Thus, the standard errors cannot be used since they will most likely be biased (the bias will be negative since the autocorrelations are positive).

One way to overcome this problem is to use quarterly data (i.e. to use only every third monthly observation). However, this leads to a substantial loss of information, and reduces the power of the tests. Alternatively, we can use Newey-West standard errors to find t-statistics of b_0 and b_1 equal to 1.29 and 1.21, which are much lower, as expected. Whereas a Wald test based on the usual standard errors has a p-value of about 0.017 (which implies a rejection of the UIP), the p-value of a Wald test based on Newey-West standard errors is 0.19.

Exercise 9: Use the data in the file `coen.txt`, `coen.xlsx` or `coen.wf1` to estimate and test alternative models for the Financial Times index. Make use of additional regressors available in that file.

Exercise 10: Use the US dollar/British pound exchange rate and the three-month forward rate in the files `forward2c.dat`, `uirp.xls`, or `uirp.wf1` to test the UIRP.

1.8 Generalized least squares

We now consider alternative estimators to overcome the inefficiency of OLS estimates associated with features of disturbances (i.e. violations of assumption **AH**) introduced in sections 1.7.2 and 1.7.3. In general the matrix $\mathbf{\Omega}$ in (23) or (27) is unknown. If its structure is known, or assumptions are made about its structure, it is possible to derive alternative estimators.

1.8.1 Heteroscedasticity

We first consider the problem of heteroscedasticity and suppose that the variance of disturbances is given by

$$\mathrm{V}[\epsilon_i] = \sigma_i^2 = \omega_i \sigma^2 \quad \forall i.$$

In the method of **weighted least squares (WLS)** the regression equation is multiplied by a suitable variable λ_i[41]

$$\lambda_i y_i = \beta_0 \lambda_i + \sum_{j=1}^{k} \beta_j \lambda_i x_{ij} + \lambda_i \epsilon_i \qquad i = 1, \ldots, n.$$

The variance of the disturbance term $\epsilon_i^*{=}\lambda_i \epsilon_i$ is

$$\mathrm{V}[\epsilon_i^*] = \mathrm{V}[\lambda_i \epsilon_i] = \lambda_i^2 \mathrm{E}[\epsilon_i^2] = \lambda_i^2 \sigma_i^2.$$

Obviously, if λ_i is chosen such that it is equal to $1/\sqrt{\omega_i}$ the variance of the disturbances in the modified equation is constant

$$\mathrm{V}[\epsilon_i^*] = \lambda_i^2 \sigma_i^2 = \sigma_i^2/\omega_i = \sigma^2 \qquad \text{if } \lambda_i{=}1/\sqrt{\omega_i},$$

and OLS estimation of the modified equation should give efficient estimates for β. This can be achieved if λ is chosen such that it is proportional to the reciprocal of the standard deviation of the disturbances. Since ω_i cannot be observed, one can try to define λ in terms of a regressor of the model. The Breusch-Pagan- or White-test may serve as starting points for this choice. From a regression of e^2 on all regressors and their squares one may find, for example, a significant coefficient for x_j^2. In this case $\lambda{=}1/x_j$ may be a good choice and the modified regression equation is given by:

$$y/x_j = b_0/x_j + b_1 x_1/x_j + b_j + b_k x_k/x_j + e/x_j,$$

or, using transformed variables $y^*{=}y/x_j$ and $x_l^*{=}x_l/x_j$ ($l{=}0, \ldots, k$; $x_0{=}1$):

$$y^* = b_0 x_0^* + b_1 x_1^* + b_j + b_k x_k^* + e^*. \tag{31}$$

Note that the coefficient b_j is the constant term in the modified regression but is still the estimator of the coefficient for regressor x_j. The variance of the modified residuals is

$$\mathrm{V}[e^*] = \mathrm{E}[e^2/x_j^2].$$

[41] The index i is meant to emphasize that λ_i differs across observations.

To the extent that e^2 and x_j^2 are related (as indicated by the White-test regressions) the variance of e^* should be approximately constant. If the White-test shows a significant coefficient for x_j, a good choice may be $\lambda{=}1/\sqrt{x_j}$.

Standard errors for coefficients estimated by WLS are based on the covariance derived from $\boldsymbol{X}^*$ (the matrix of weighted regressors) and $e^*{=}\boldsymbol{X}^*\boldsymbol{b}_{\text{WLS}}$:

$$\hat{\text{V}}[\boldsymbol{b}_{\text{WLS}}] = s_{e^*}^2(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}.$$

The R^2 from the modified regression (31) must not be compared to the R^2 from the original equation since the dependent variables are not the same. It does not describe the relation of interest and thus the R^2 from the transformed equation is rather useless. Equation (31) is mainly a device to obtain efficient estimates and correct statistical inference. Therefore, parameter estimates should also be interpreted and used in the context of the original (untransformed) model.

Another approach is based on using weights that are estimated from the data. It may be assumed that heteroscedasticity depends on regressors $\boldsymbol{z}$ (which may include original regressors $\boldsymbol{x}$). Plausible candidates are $\sigma_i^2{=}\boldsymbol{z}_i'\boldsymbol{\beta}_z$ or $\sigma_i^2{=}\sigma^2\exp\{\boldsymbol{z}_i'\boldsymbol{\beta}_z\}$. In the method of **feasible generalized least squares (FGLS)** estimates of σ_i^2 are used to obtain an estimate of the matrix $\boldsymbol{\Omega}$ defined in (23). The corresponding FGLS estimator is given by

$$\boldsymbol{\beta}_{\text{FGLS}} = (\boldsymbol{X}'\hat{\boldsymbol{\Omega}}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}\boldsymbol{y}.$$

The FGLS estimator is asymptotically efficient if the estimate to construct $\hat{\boldsymbol{\Omega}}$ is consistent. Estimates $\hat{\sigma}_i^2$ can be obtained by using $\boldsymbol{z}_i'\boldsymbol{b}_z$ from the auxiliary regressions

$$e_i^2 = \boldsymbol{z}_i'\boldsymbol{b}_z + u_i \qquad \text{or} \qquad \ln e_i^2 = \boldsymbol{z}_i'\boldsymbol{b}_z + v_i.$$

Example 17: We use data and results from example 14. The White test regression has shown a significant relation between e^2 and x_1^2. We use $1/x_1$ as the weight in a WLS regression and obtain (t-statistics in parenthesis)

$$\underset{}{\ln y^*} = \underset{(27.9)}{1.23} + \underset{(8.7)}{0.042}\,x_1^* + \underset{(1.8)}{0.025}\,x_2^* + e^*.$$

Note the small changes in the estimated coefficients (which are expected) and the substantial changes in the t-statistics compared to example 14.

Exercise 11: Consider the data from example 11 (see file `hedonic.wf1`). Estimate a model excluding the interaction term. Test the residuals from this model for heteroscedasticity, and obtain WLS or FGLS estimates if required.

1.8.2 Autocorrelation

Autocorrelation is another case where assumption **AH** is violated. To overcome the associated inefficiency of OLS we assume, for simplicity, that autocorrelations in the covariance matrix (27) can be expressed in terms of the first order (lag one) autocorrelation ρ_1 only:

$$\rho_\tau = \rho_1^\tau \qquad \tau = 1, \ldots, n-1.$$

This is equivalent[42] to the model

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

$$u_t = \rho_1 u_{t-1} + \epsilon_t \qquad \epsilon_t \sim \text{i.i.d.}$$

Upon substitution of u_t from the second equation we find

$$y_t = \beta_0(1-\rho_1) + \beta_1 x_{t1} - \beta_1\rho_1 x_{t-1,1} + \cdots + \beta_k x_{tk} - \beta_k\rho_1 x_{t-1,k} + \rho_1 y_{t-1} + \epsilon_t.$$

Alternatively, we can use transformed variables $y_t^*{=}y_t{-}\rho_1 y_{t-1}$ and $x_{tj}^*{=}x_{tj}{-}\rho_1 x_{t-1,j}$ (the so-called **partial differences**):

$$y_t^* = \beta_0^* + \beta_1 x_{t1}^* + \cdots + \beta_k x_{tk}^* + \epsilon_t.$$

ϵ_t is uncorrelated, and estimating the transformed equation by OLS gives efficient estimates. ρ_1 is unknown, but we can use FGLS if ρ_1 is replaced by a consistent estimate. Several options are available to estimate ρ_1 consistently (e.g. Cochrane-Orcutt or Prais-Winsten; see Greene (2003), section 12.9). The simplest is to use the first order autocorrelation of the residuals from the original (consistent) regression.

We also note that autocorrelation of disturbances may be viewed as the result of a misspecified equation (see section 1.7.3). In other words, the (original) equation has to be modified to account for the dynamics of the variables and responses involved. According to this view, GLS is not appropriate to resolve the inefficiency of OLS. A starting point for the reformulation may be obtained from the equation using partial differences.

WLS and FGLS may be summarized in terms of the Cholesky decomposition of the inverse of $\boldsymbol{\Omega}$:

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{C}'\boldsymbol{C} \qquad \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}' = \boldsymbol{I}.$$

The matrix $\boldsymbol{C}$ is used to transform the model such the transformed disturbances are homoscedastic and non-autocorrelated:

$$\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\epsilon} \quad \Longrightarrow \quad \boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$$

$$\mathrm{E}[\boldsymbol{C}\boldsymbol{\epsilon}] = \boldsymbol{C}\mathrm{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$$

$$\mathrm{V}[\boldsymbol{C}\boldsymbol{\epsilon}] = \boldsymbol{C}\mathrm{V}[\boldsymbol{\epsilon}]\boldsymbol{C}' = \boldsymbol{C}\sigma^2\boldsymbol{\Omega}\boldsymbol{C}' = \sigma^2\boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}' = \sigma^2\boldsymbol{I}.$$

If the transformed equation is estimated by OLS the **GLS** estimator is obtained:

$$\boldsymbol{\beta}_{\text{GLS}} = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y}^* = (\boldsymbol{X}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{y}.$$

[42] The autocorrelations of u_t from the model $u_t{=}\rho_1 u_{t-1}{+}\epsilon_t$ can be shown to be given by $\rho_\tau{=}\rho_1^\tau$ (see section 2.2).

Example 18: We consider the model (29) estimated in example 15. The estimated residual autocorrelation at lag 1 is $\hat{\rho}_1$=0.5 and can be used to form partial differences (e.g. y_t^*=y_t−0.5y_{t-1}). The FGLS estimates are given by (t-statistics below coefficients)

$$\underset{}{y_t^*} = \underset{(7.2)}{299.7} + \underset{(8.7)}{0.446}\, p_{t-6}^* - \underset{(5.8)}{5.41}\, c_{t-7}^* + e_t^*.$$

The increased efficiency associated with the FGLS estimator cannot be derived from comparing these estimates to those in example 15. It is worth noting, however, that the t-statistics are much lower and closer to those obtained in equation (30).

Exercise 12: Consider a model you have estimated in exercise 9. Test the residuals for autocorrelation, and obtain FGLS estimates if required.

1.8.3 Example 19: Long-horizon return regressions

Suppose a time series of (overlapping) h-period returns is computed from single-period returns as follows (see section 2.1):

$$y(h)_t = y_t + y_{t-1} + \cdots + y_{t-h} \qquad y_t = \ln p_t - \ln p_{t-1}.$$

In the context of long-horizon return regressions the conditional expected value of $y(h)_t$ is formed on the basis of the information set available at the date when forecasts are made; i.e. at date $t-h-1$. For example, in case of a single regressor a regression equation is formulated as

$$y(h)_t = \beta_0 + \beta_1 x_{t-h-1} + \epsilon_t.$$

Alternatively, the model can be reformulated by shifting the time axis as follows:

$$y(h)_{t+h} = \beta_0 + \beta_1 x_{t-1} + \epsilon_{t+h}.$$

This model specification corresponds to the problem of predicting the sum of h single-period returns during the period t until $t+h$ on the basis of information available at date $t-1$.

We consider weekly observations of the DAX p_t which are used to form a time series of annual returns

$$y(52)_t = \ln p_t - \ln p_{t-52}.$$

The following regressors are available: the dividend yield d_t, the spread between ten-year and one-month interest rates s_t, the one-month real interest rate r_t, and the growth rate of industrial production g_t[43]. Real interest rates are computed by subtracting the inflation rate from one-month interest rates. The inflation rate i_t is calculated from the consumer price index c_t, using $i_t = \ln c_t - \ln c_{t-52}$. The growth rate of industrial production is computed in the same way from the industrial production index. Details can be found in the file `long.wf1`.

Each explanatory variable is included with a lag of 53 weeks and the estimated equation is[44]

$$y(52)_t = -0.41 + 39.8 d_{t-53} - 1.76 s_{t-53} - 8.79 r_{t-53} + 0.638 g_{t-53} + e_t.$$

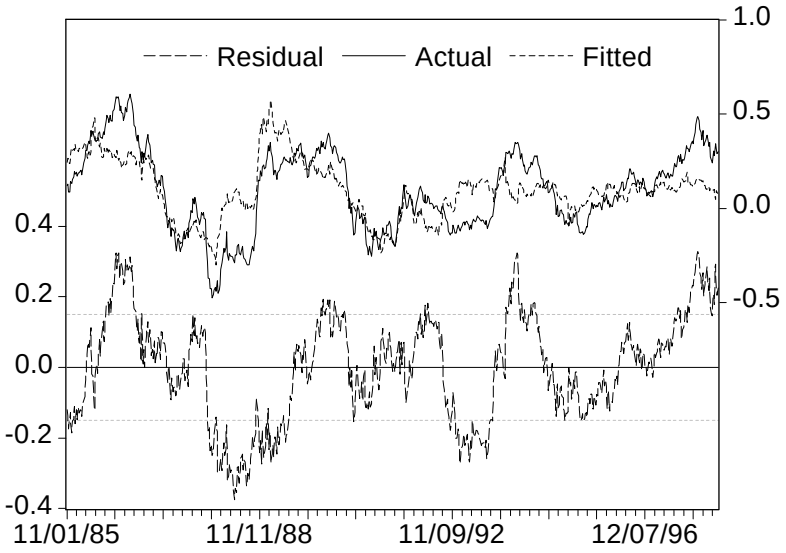
The present result is typical for some similar cases known from literature which support the 'predictability' of long-horizon returns. All parameters are highly significant which leads to the (possibly premature) conclusion that the corresponding explanatory variables are relevant when forming expectations.

The most obvious deficiency of this model is the substantial autocorrelation of the residuals (r_1=0.959). The literature on **return predictability** typically reports that R^2 increases

[43] Source: Datastream; January 1983 to December 1997; 782 observations.

[44] All p-values are less than 0.001 except for s_{t-53} with a p-value equal to 0.002. Because of missing values the sample used in the regression starts on January 11, 1985.

Figure 1: Fit and residuals of the long-horizon return regression of annual DAX returns.



with h (see Kirby, 1997). This property of the residuals is mainly caused by the way multi-period returns are constructed. The (positive) autocorrelation of residuals causes a (negative) bias of the standard errors of the parameters. In extreme cases this may lead to the so-called **spurious regression** problem[45].

No matter whether this is, in fact, a spurious regression case, Figure 1 shows that data and in-sample fit may deviate strongly from each other for very long periods. Therefore, when this model is used for out-of-sample forecasts, large errors in the estimation of expected returns can be expected over long time intervals.

Note that the residual autocorrelation cannot be simply corrected by using partial differences (e.g. $y_t - \hat{\rho}_1 y_{t-1}$). Such a formulation would imply single-period expectations, contrary to the intension of long-horizon regressions. On the other hand, partial differences based on longer periods (e.g. $y_t - \hat{\rho}_{53} y_{t-53}$) would not account for the short-term autocorrelation.

A viable alternative is to use Newey-West HAC standard errors which shows that only the coefficients of d_{t-53} and r_{t-53} remain significant. The p-values of s_{t-53} and g_{t-53} increase to 0.28 and 0.11, respectively. Valkanov (2003) provides theoretical results why t-statistics in long-horizon regressions do not converge to well-defined distributions, and proposes a rescaled t-statistic.

[45] For details see Granger and Newbold (1974).

1.9 Endogeneity and instrumental variable estimation[46]

1.9.1 Endogeneity

In sections 1.2 and 1.3 the exogeneity assumptions $\mathbf{AX}$ and $\overline{\mathbf{AX}}$ were found to be crucial for the properties of OLS estimates. The term **endogeneity** (i.e. regressors are not exogenous, but are correlated with the disturbances) refers to violations of these assumptions. There are several circumstances which may lead to endogeneity. As shown in section 1.6.7 omitted variables lead to biased and inconsistent OLS estimates of regression coefficients, because regressors and disturbances are correlated in this case. Two further reasons for endogeneity are measurement errors and simultaneity (see below). Roberts and Whited (2012) provide a comprehensive treatment of endogeneity, its sources, and econometric techniques aimed at addressing that problem in the context of corporate finance.

$\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are correlated in case of **measurement errors** (or **errors-in-variables**) (see Greene, 2003, section 5.6). For instance, the Fama-MacBeth two-step procedure mentioned in example 6 leads to an errors-in-variables problem, since the second step uses generated regressors. To provide some insight into the associated consequences we consider the regression $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is measured with error: $\tilde{\boldsymbol{X}}=\boldsymbol{X}+\boldsymbol{u}$. $\boldsymbol{u}$ is a mean-zero error, uncorrelated with $\boldsymbol{y}$ and $\boldsymbol{X}$. Upon substituting $\boldsymbol{X}$ with $\tilde{\boldsymbol{X}}-\boldsymbol{u}$ we obtain

$$\boldsymbol{y} = (\tilde{\boldsymbol{X}} - \boldsymbol{u})\boldsymbol{\beta} + \boldsymbol{\epsilon} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{v} \qquad \boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{u}\boldsymbol{\beta}.$$

Regressors and disturbances in this regression are correlated, since

$$\begin{aligned} \mathrm{E}[\tilde{\boldsymbol{X}}'\boldsymbol{v}] &= \mathrm{E}[(\boldsymbol{X} + \boldsymbol{u})'(\boldsymbol{\epsilon} - \boldsymbol{u}\boldsymbol{\beta})] \\ &= \mathrm{E}[\boldsymbol{X}'\boldsymbol{\epsilon} + \boldsymbol{u}'\boldsymbol{\epsilon} - \boldsymbol{X}'\boldsymbol{u}\boldsymbol{\beta} - \boldsymbol{u}'\boldsymbol{u}\boldsymbol{\beta}] \\ &= -\mathrm{E}[\boldsymbol{u}'\boldsymbol{u}]\boldsymbol{\beta} = -\sigma_u^2\boldsymbol{\beta}. \end{aligned}$$

Note that *all* coefficients are biased and inconsistent, even if only one of the regressors in $\boldsymbol{X}$ is measured with error (see Greene, 2003, p.85). The bias can be derived on the basis of relation (4):

$$\mathrm{E}[\boldsymbol{b}] = \boldsymbol{\beta} + \mathrm{E}\left[(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\right]\mathrm{E}[\tilde{\boldsymbol{X}}'\boldsymbol{v}] = \left(1 - \sigma_u^2\mathrm{E}\left[(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\right]\right)\boldsymbol{\beta}.$$

In case of the slope in a simple regression we have

$$\mathrm{E}[b] = \beta\left(1 - \frac{\sigma_u^2}{\sigma_{\tilde{x}}^2}\right).$$

Since $\mathrm{E}[\boldsymbol{X}'\boldsymbol{u}]=0$ and $\sigma_{\tilde{x}}^2=\sigma_x^2+\sigma_u^2$ we obtain

$$\beta\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right).$$

This shows that b will be biased towards zero, since the term in parenthesis is less than one.

[46] Most of this section is based on Greene (2003), section 5.4, and Wooldridge (2002), sections 5.1 and 6.2.

A very typical case of endogeneity is **simultaneity** (see Greene, 2003, p.378). Simultaneity arises if at least one explanatory variable x_k is not exogenous but is determined (partly) as a function of y, and thus x_k and ϵ are correlated. A frequently used example is the case of demand and supply functions in the following simple model:

$$d = \beta_0^d + \beta_1^d p + \epsilon^d \qquad s = \beta_0^s + \beta_1^s p + \epsilon^s \qquad d = s \text{ (in equilibrium)}.$$

Supply and demand differ *conceptually* but usually they cannot be separately measured. Thus, we can only observe quantities sold q (representing the market equilibrium values of d and s). Using q=d=s we can solve the equations

$$q = \beta_0^d + \beta_1^d p + \epsilon^d \qquad q = \beta_0^s + \beta_1^s p + \epsilon^s$$

for p and obtain

$$p = \frac{\beta_0^s - \beta_0^d}{\beta_1^d - \beta_1^s} + \frac{\epsilon^s - \epsilon^d}{\beta_1^d - \beta_1^s}.$$

Thus, p is a function of the disturbances from *both* equations, and is thus endogenous in a regression of q on p. Its covariance with ϵ^d and ϵ^s is given by[47]

$$\text{cov}[p, \epsilon^d] = -\frac{\text{V}[\epsilon^d]}{\beta_1^d - \beta_1^s} \qquad \text{cov}[p, \epsilon^s] = \frac{\text{V}[\epsilon^s]}{\beta_1^d - \beta_1^s}.$$

In this case, endogeneity is a consequence of market equilibrium. If we estimate a regression of q on p it is not clear whether the result is the slope of a demand or supply function. It can be shown, however, that the probability limit of the coefficient of p in a regression of q on p is given by (see Hayashi, 2000, p.189)

$$\frac{\beta_1^d \text{V}[\epsilon^s] + \beta_1^s \text{V}[\epsilon^d]}{\text{V}[\epsilon^s] + \text{V}[\epsilon^d]}.$$

Thus, the estimated coefficient is neither the slope of the demand nor the supply function, but a weighted average of both. If supply shocks dominate (i.e. $\text{V}[\epsilon^s]$>$\text{V}[\epsilon^d]$) the estimate will be closer to the slope β_1^d of the demand function. This may hold in case of agricultural products which are more exposed to supply shocks (e.g. weather conditions). Positive(!) slopes of a "demand" function may be found in case of manufactured goods which are more subject to demand shifts over the business cycle. In an extreme case, where there are no demand shocks, the observed quantities correspond to the intersections of a (constant) demand curve and many supply curves. This would allow us to *identify* the estimated slope as the slope of a demand function. This observation will be the basis for a solution to the endogeneity bias described below.

[47] For simplicity we assume that ϵ^d and ϵ^s are uncorrelated.

1.9.2 Instrumental variable estimation

Instrumental variable (**IV**) or **two-stage least squares** (**2SLS**) estimation is a method that can be used when exogeneity is violated. We start by considering the case of only one *endogenous* regressor x_k being correlated with ϵ in the regression

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon. \tag{32}$$

IV-estimation is based on using an observable variable z – the so-called instrumental variable or instrument – which satisfies the following conditions:

1. the instrument z must be uncorrelated with ϵ (orthogonality condition):

   $$\text{corr}[z, \epsilon] = 0.$$

   This condition implies that the instrument and a potentially omitted variable (absorbed in ϵ) must be uncorrelated. If this condition is violated, the instrument is considered to be *invalid*.

2. the coefficient b_z of z in the so-called **first-stage regression**

   $$x_k = b_0 + \sum_{j=1}^{k-1} b_j x_j + b_z z + v = \hat{x}_k + v$$

   must be non-zero: $b_z \neq 0$. If this condition (which is also called *relevance condition*) is violated, the instrument is called *weak*.

3. z must not appear in the original regression (32). Roberts and Whited (2012) call $\text{corr}[z, \epsilon]=0$ the *exclusion condition*, expressing that the only way z must affect y is through the endogenous regressor x_k (but not directly via equation 32).

It is frequently stated that z must be correlated with the endogenous regressor. Note that the second requirement is stronger, since it refers to the *partial* correlation between z and x_k. Whereas the second condition can be tested, the first requirement of zero correlation between z and ϵ cannot be tested directly. Since ϵ is unobservable, this assumption must be maintained or justified on the basis of economic arguments. However, in section 1.9.3 we will show that it is possible to test for exogeneity of regressors and the appropriateness of instruments.

Consistent estimates can be obtained by using $\hat{x}_k$ from the first-stage regression to replace x_k in the original regression (32):

$$y = \beta_0^{\text{IV}} + \sum_{j=1}^{k-1} \beta_j^{\text{IV}} x_j + \beta_k^{\text{IV}} \hat{x}_k + \epsilon^{\text{IV}}.$$

This (second-stage) equation can be estimated with OLS (important details regarding standard errors of coefficients are treated below). $\hat{x}_k$ is exogenous by construction, since it

only depends on exogenous regressors and an instrument uncorrelated with ϵ. Thus, the endogeneity is removed from equation (32), and the resulting IV-estimates of its parameters are consistent.

We briefly return to the example of supply and demand functions described in section 1.9.1. A suitable instrument for a demand equation is an observable variable which leads to supply shifts and hence price changes (e.g. temperature variations affect the *supply* of coffee and its price). We are able to identify the demand function and estimate its slope, if the instrument is uncorrelated with demand shocks (i.e. temperature has little or no impact on the *demand* for coffee). IV-estimates can be obtained by first regressing price on temperature (and other regressors), and then use the *fitted* prices as a regressor (among others) to explain quantities sold. Consequently, the second regression can be considered to be a demand equation.

In general, IV-estimation replaces those elements of $\boldsymbol{X}$ which are correlated with $\boldsymbol{\epsilon}$ by a set of instruments which are uncorrelated with $\boldsymbol{\epsilon}$, but related to the endogenous elements of $\boldsymbol{X}$ by first-stage regressions. The number of instruments can be larger than the number of endogenous regressors. However, the number of instruments must not be less than the number of (potentially) endogenous regressors (this is the so-called *order condition*). To derive the IV-estimator in more general terms we define a matrix $\boldsymbol{Z}$ of exogenous regressors which includes the exogenous elements of $\boldsymbol{X}$ (including the constant) and the instruments. Regressors suspected to be endogenous are not included in $\boldsymbol{Z}$. We assume that $\boldsymbol{Z}$ is uncorrelated with $\boldsymbol{\epsilon}$

$$\mathrm{E}[\boldsymbol{Z}'\boldsymbol{\epsilon}] = \boldsymbol{0}. \tag{33}$$

IV-estimation can be viewed as a two-stage LS procedure. The first stage involves regressing each original regressor x_i on all instruments $\boldsymbol{Z}$:

$$\boldsymbol{x}^i = \boldsymbol{Z}\boldsymbol{b}_z^i + \boldsymbol{v}^i = \hat{\boldsymbol{x}}^i + \boldsymbol{v}^i \qquad i = 1, \ldots, K.$$

Regressors in $\boldsymbol{Z}$ that are also present in the original matrix $\boldsymbol{X}$ are exactly reproduced by this regression. The resulting fitted values are used to construct the matrix $\hat{\boldsymbol{X}}$ which is equal to $\boldsymbol{X}$, except for the columns which correspond to the (suspected) endogenous regressors. $\hat{\boldsymbol{X}}$ is used in the second stage in the regression

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\boldsymbol{b}_{\mathrm{IV}} + \boldsymbol{e}. \tag{34}$$

Since $\hat{\boldsymbol{X}}$ only represents exogenous information the IV-estimator given by

$$\boldsymbol{b}_{\mathrm{IV}} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$$

is consistent. $\hat{\boldsymbol{X}}$ can be written as (see Greene, 2003, p.78)

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{b}_z, \tag{35}$$

where $\boldsymbol{b}_z$ is a matrix (or vector) of coefficients from first-stage regressions. If $\boldsymbol{Z}$ has the same number of columns as $\boldsymbol{X}$ (i.e. the number of instruments is equal to the number of endogenous regressors) the IV-estimator is given by

$$\boldsymbol{b}_{\mathrm{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

If the following conditions are satisfied

$$\text{plim}\,\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{Q}_z \qquad |\boldsymbol{Q}_z| > 0$$

$$\text{plim}\,\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{Q}_{zx} \qquad |\boldsymbol{Q}_{zx}| > 0$$

$$\text{plim}\,\frac{1}{n}\boldsymbol{Z}'\boldsymbol{\epsilon} = \boldsymbol{0}$$

the IV-estimator can be shown to be consistent

$$\text{plim}\,\boldsymbol{b}_{\text{IV}} = \boldsymbol{\beta} + \text{plim}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \cdot \text{plim}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{\epsilon}\right) = \boldsymbol{\beta}, \tag{36}$$

and asymptotically normal (see Greene, 2003, p.77):

$$\boldsymbol{b}_{\text{IV}} \overset{a}{\sim} \text{N}(\boldsymbol{\beta}, \boldsymbol{Q}_{zx}^{-1}\boldsymbol{Q}_z\boldsymbol{Q}_{zx}^{-1}).$$

The asymptotic variance of $\boldsymbol{b}_{\text{IV}}$ is estimated by

$$\tilde{s}_e^2(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{X}'\boldsymbol{Z})^{-1}, \tag{37}$$

where

$$\tilde{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{b}_{\text{IV}})^2.$$

Note that the standard errors and $\tilde{s}_e^2$ are derived from residuals based on $\boldsymbol{X}$ rather than $\hat{\boldsymbol{X}}$:

$$\boldsymbol{e}_{\text{IV}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{\text{IV}}. \tag{38}$$

This further implies that the R^2 from IV-estimation based on these residuals cannot be interpreted. The variance of $\boldsymbol{e}_{\text{IV}}$ can be even larger than the variance of $\boldsymbol{y}$, and thus R^2 can become negative.

If the IV-estimation is done in two OLS-based steps, the standard errors from running the regression (34) will differ from those based on (37), and thus will be incorrect (see Wooldridge, 2002, p.91). We also note that the standard errors of $\boldsymbol{b}_{\text{IV}}$ are always larger than the OLS standard errors, since $\boldsymbol{b}_{\text{IV}}$ only uses that part of the (co)variance of $\boldsymbol{X}$ which appears in the fitted values $\hat{\boldsymbol{X}}$. Thus, the potential reduction of the bias is associated with a loss in efficiency.

So far, the treatment of this subject has been based on the assumption that instruments satisfy the conditions stated above. However, violations of these conditions may have serious consequences (see Murray, 2006, p.124). First, invalid instruments – correlated with the disturbance term – yield biased and inconsistent estimates, which can be even more biased than the OLS estimates. Second, if instruments are too weak it may not be possible to eliminate the bias associated with OLS, and standard errors may be misleading even in very large samples. Thus, the selection of instruments is crucial for the properties of IV-estimates. In the next section we will investigate those issues more closely.

1.9.3 Selection of instruments and tests

Suitable instruments are usually hard to find. In the previous section we have mentioned weather conditions. They may serve as instruments to estimate a demand equation for coffee, since they may be responsible for supply shifts, and they are correlated with prices but not with demand shocks. In a study on crime rates Levitt (1997) uses data on electoral cycles as instruments to estimate the effects associated with hiring policemen on crime rates. Such a regression is subject to a simultaneity bias since more policemen should lower crime rates, however, cities with a higher crime rate tend to hire more policemen. Electoral cycles may be suitable instruments: they are exogenously given (predetermined), and expenditures on security may be higher during election years (i.e. the instrument is correlated with the endogenous regressor). In a time series context, lagged variables of the endogenous regressors may be reasonable candidates. The instruments $\boldsymbol{X}_{t-1}$ may be highly correlated with $\boldsymbol{X}_t$ but uncorrelated with $\boldsymbol{\epsilon}_t$. According to Roberts and Whited (2012) suitable instruments can be derived from biological or physical events or features. They stress the importance of understanding the economics of the question at hand, i.e. that the instrument must only affect y via the endogenous regressor. For example, institutional changes may be suitable as long as the economic question under study was not one of the original reasons for the institutional change.

Note that instruments are distinctly different from proxy variables (which may serve as a substitute for an otherwise omitted regressor). Whereas a proxy variable should be highly correlated with an omitted regressor, an instrument must be highly correlated with a (potentially) endogenous regressor x_k. However, the higher the correlation of an instrument z with x_k, the less justified may be the assumption that z is uncorrelated with ϵ – given that the correlation between x_k and ϵ is causing the problem in the first place!

IV-estimation will only lead to consistent estimates if suitable instruments are found (i.e. $\mathrm{E}[\boldsymbol{Z}'\boldsymbol{\epsilon}]=\boldsymbol{0}$) and $\boldsymbol{b}_z\neq\boldsymbol{0}$). If instruments are not exogenous because (33) is violated (i.e. $\mathrm{E}[\boldsymbol{Z}'\boldsymbol{\epsilon}]\neq\boldsymbol{0}$) the IV-estimator is inconsistent (see (36)). Note also that the consistency of $\boldsymbol{b}_{\mathrm{IV}}$ critically depends on $\mathrm{cov}[\boldsymbol{Z},\boldsymbol{X}]$ even if $\mathrm{E}[\boldsymbol{Z}'\boldsymbol{\epsilon}]\approx\boldsymbol{0}$. (36) shows that poor instruments (being only weakly correlated with $\boldsymbol{X}$) may lead to strongly biased estimates even in large samples (i.e. inconsistency prevails). As noted above, estimated IV standard errors are always larger than OLS standard errors, but they can be strongly biased downward when instruments are weak (see Murray, 2006, p.125).

Given these problems associated with IV-estimation, it is important to first test for the endogeneity of a regressor (i.e. does the endogeneity problem even exist?). This can be done with the **Hausman test** described below. However, that test requires valid and powerful instruments. Thus, it is necessary to first investigate the properties of potential instruments.

Testing $\boldsymbol{b}_z$: Evidence for weak instruments can be obtained from first-stage regressions. The joint significance of the instruments' coefficients can be tested with an F-test (weak instruments will lead to low F-statistics). If weak instruments are found, the worst should be dropped and replaced by more suitable instruments (if possible at all). To emphasize the impact of (very) weak instruments consider an extreme case, where the instruments' coefficients in the first-stage regression are all zero. In that case, $\hat{\boldsymbol{x}}^k$ could not be used in the second stage, since it would be a linear combination of the other exogenous regressors. As a rule of thumb, n times R^2 from the first-stage regression should be larger than the number of instruments so that the bias

of IV will tend to be less than the OLS bias (see Murray, 2006, p.124). Staiger and Stock (1997) consider instruments to be weak, if the first-stage F-statistic, testing the coefficients $\boldsymbol{b}_z$ of instruments entering the first-stage regression for being jointly zero, is less than ten. Dealing with more than one endogeneous regressor is even more demanding (see Stock et al., 2002).

Testing corr$[z, \epsilon]$: If there are more instruments than necessary,[48] $\boldsymbol{\beta}$ is **overidentified**. Rather than eliminating some instruments they are usually kept in the model, since they may increase the efficiency of IV-estimates. However, a gain in efficiency requires valid instruments (i.e. they must be truly exogenous). Since invalid instruments lead to inconsistent IV-estimates, it is necessary to test the overidentifying restrictions. Suppose two instruments z_1 and z_2 are available, but only one (z_1) is used in 2SLS (this is the *just identified* case). Whereas the condition corr$[z_1, \epsilon]$=0 cannot be tested (both $\boldsymbol{e}$ from (34) and $\boldsymbol{e}_{\text{IV}}$ from (38) are uncorrelated with z by the LS principle), we can test whether z_2 is uncorrelated with e, and may thus be a suitable instrument. The same applies vice versa, if z_2 is used in 2SLS and z_1 is tested.

In general, overidentifying restrictions can be tested by regressing the residuals from the 2SLS regression $\boldsymbol{e}_{\text{IV}}$ as defined in (38) on all exogenous regressors and all instruments (see Wooldridge, 2002, p.123). Valid (i.e. exogenous) instruments should not be related to 2SLS residuals. Under H_0: $\text{E}[\boldsymbol{Z}'\boldsymbol{\epsilon}]=\boldsymbol{0}$, the test statistic is $nR^2 \sim \chi^2_m$, where R^2 is taken from this regression, and m is the difference between the number of endogenous regressors and the number of instruments. A failure to reject the overidentifying restrictions is an indicator of valid instruments.

If acceptable instruments have been found, we can proceed to test for the presence of endogeneity. The Hausman test is based on comparing $\boldsymbol{b}_{\text{IV}}$ and $\boldsymbol{b}_{\text{LS}}$, the IV and OLS estimates of $\boldsymbol{\beta}$. A significant difference is an indicator of endogeneity. The Hausman test is based on the null hypothesis that $\text{plim}\,(1/n)\boldsymbol{X}'\boldsymbol{\epsilon}=\boldsymbol{0}$ (i.e. H_0: $\boldsymbol{X}$ is not endogenous). In this case OLS and IV are both consistent. If H_0 does not hold, only IV is consistent. However, a failure to reject H_0 may be due to invalid or weak instruments. Murray (2006, p.126) reviews alternative procedures, which are less affected by weak instruments, and provides further useful guidelines.

The Hausman test is based on $\boldsymbol{d}=\boldsymbol{b}_{\text{IV}}-\boldsymbol{b}_{\text{LS}}$, H_0: $\text{plim}\,\boldsymbol{d}=\boldsymbol{0}$. The Hausman test statistic is given by

$$H = \boldsymbol{d}'\widehat{\text{aV}}[\boldsymbol{d}]^{-1}\boldsymbol{d} = \frac{1}{s_e^2}\boldsymbol{d}'[(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1} - (\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\boldsymbol{d} \qquad H \sim \chi^2_m,$$

where $\hat{\boldsymbol{X}}$ is defined in (35), $m=K-K_0$ and K_0 is the number of regressors for which H_0 must not be tested (because they are known – actually, *assumed* – to be exogenous).

A simplified but asymptotically equivalent version of the test is based on the residuals from the first-stage regression associated with the endogenous regressor (v^k), and an auxiliary OLS regression (see Wooldridge, 2002, p.119), where v^k is added to the original regression[49]

[48] By the order condition the number of instruments must be greater or equal to the number of endogenous regressors.

[49] Note that adding v^k to equation (32) does not change any of the coefficients in (32) estimated by 2SLS.

(32):

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + b_v v^k + \epsilon.$$

Note that v^k represents the endogenous part of x_k (if there is any). If there is an endogeneity problem, the coefficient b_v will pick up this effect (i.e. the endogenous part of ϵ will move to $b_v v^k$). Thus, endogeneity can be tested by a standard t-test of b_v (based on heteroscedasticity-consistent standard errors). If the coefficient is significant, we reject H_0 (no endogeneity) and conclude that the suspected regressor is endogenous. However, failing to reject H_0 need not indicate the absence of endogeneity, but may be due to weak instruments. If more than one regressor is suspected to be endogenous, a first-stage regression is run for each one of them. All residuals thus obtained are added to the original equation, and a F-test for the residuals' coefficients being jointly equal to zero can be used.

Example 20: We use a subset of wage data from Wooldridge (2002, p.89) (example 5.2)[50] to illustrate IV-estimation. Wages are assumed to be (partially) determined by the unobservable variable ability. Another regressor in the wage regression is education (measured by years of schooling), which can be assumed to be correlated with ability. Since ability is an omitted variable, which is correlated with (at least) another regressor, this will lead to inconsistent OLS estimates. The dummy variable 'near' indicates whether someone grew up near a four-year college. This variable can be used as an instrument: it is exogenously given (i.e. uncorrelated with the error term which contains 'ability'), and most likely correlated with education. The first-stage regression shows that the coefficient of the instrument is highly significant (i.e. there seems to be no danger of using a weak instrument). The condition $\text{corr}[z, \epsilon]=0$ cannot be tested since only one instrument is available. The coefficient of v (the residual from the first-stage regression) is highly significant (p-value 0.0165) which indicates an endogeneity problem (as expected). Comparing OLS and IV estimates shows that the IV coefficient of education is three times as high as the OLS coefficient. However, the IV standard error is more than ten times as high as the OLS standard error. Similar results are obtained for other coefficients (see `wage.R`, `wage.wf1`, or `wage.xls`).

[50] Source: Go to the book companion site of Wooldridge (2003) at `http://www.cengage.com/` (latest edition), click on "Data Sets", download one of the zip-files and choose "`CARD.*`"; 874 observations have been extracted from the original sample.

1.9.4 Example 21: Consumption based asset pricing

We consider an investor who maximizes the expected utility of present and future consumption by solving

$$\max \mathrm{E}_t\left[\sum_{\tau=0}^{T} \delta^\tau U(C_{t+\tau})\right],$$

subject to the budget constraint

$$C_t + W_t = L_t + (1 + R_t)W_{t-1},$$

where δ is the time discount factor, C_t is the investor's consumption, W_t is financial wealth, L_t is labor income and R_t is the return from investments. A first-order condition for the investor's intertemporal consumption and investment problem is given by the intertemporal **Euler equation**

$$\delta \mathrm{E}_t[U'(C_{t+1})(1 + R_{t+1})] = U'(C_t).$$

We derive this equation in a simplified, two-period setting, ignoring labor income. In this case the objective is given by

$$\max \mathrm{E}_t[U(C_t) + \delta U(C_{t+1})],$$

and the (reformulated) budget constraint is

$$C_{t+1} = (1 + R_{t+1})W_t - W_{t+1}.$$

The agent has to decide how much to consume and invest now (in t) in order to maximize expected utility. This implies maximizing current and expected utility with respect to either current consumption or wealth (because of the budget constraint, only one of the two needs to be determined):

$$\max_{W_t} U(C_t) + \delta \mathrm{E}_t[U((1 + R_{t+1})W_t - W_{t+1}))].$$

Taking first derivatives and applying the chain results gives

$$-U'(C_t) + \delta \mathrm{E}_t\big[U'((1 + R_{t+1})W_t - W_{t+1})(1 + R_{1+t})\big] = 0$$

which simplifies to

$$U'(C_t) = \delta \mathrm{E}_t[U'(C_{t+1})(1 + R_{t+1})].$$

Thus, the optimal solution is obtained by equating the expected marginal utility from investment to the marginal utility of consumption. The Euler equation can be rewritten in terms of the so-called **stochastic discount factor** M_t

$$\mathrm{E}_{t-1}[(1 + R_t)M_t] = 1 \qquad M_t = \delta \frac{U'(C_t)}{U'(C_{t-1})}.$$

This equation is also known as the **consumption based CAPM**.
Using the power utility function

$$U(C) = \frac{C^{1-\gamma}}{1-\gamma} \qquad U'(C) = C^{-\gamma} \quad (\gamma \ldots \text{coefficient of relative risk aversion})$$

the Euler equation is given by

$$\mathrm{E}_{t-1}\left[(1+R_t)\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma}\right] = 1 \quad \text{or} \quad \mathrm{E}_{t-1}[(1+R_t)\delta C_t^{-\gamma}] = C_{t-1}^{-\gamma}.$$

Campbell et al. (1997, p.306) assume that R_t and C_t are lognormal and homoscedastic, use the relation $\ln \mathrm{E}[X]=\mathrm{E}[\ln X]+0.5\mathrm{V}[\ln X]$ (see (41) in section 2.1.2), and take logs of the term in square brackets to obtain

$$\mathrm{E}_{t-1}[\ln(1+R_t)] + \ln\delta - \gamma \mathrm{E}_{t-1}[\Delta \ln C_t] + 0.5c = 0.$$

c is a constant (by the assumption of homoscedasticity) involving variances and covariances of R_t and C_t. This equation implies a linear relation between expected returns and expected consumption growth, and can be used to estimate γ. We replace expectations by observed data on log returns $y_t=\ln(1+R_t)$ and consumption growth $c_t=\Delta \ln C_t$

$$y_t = \mathrm{E}_{t-1}[\ln(1+R_t)] + a_t \qquad c_t = \mathrm{E}_{t-1}[\Delta \ln C_t] + b_t.$$

Replacing expectations by observed variables implies measurement errors, which further lead to inconsistent OLS estimates if obtained from the regression equation (as shown in section 1.9.1)

$$y_t = \alpha + \gamma c_t + \epsilon_t \qquad \alpha = -\ln\delta - 0.5c \quad \epsilon_t = a_t - \gamma b_t.$$

We estimate this equation to replicate parts of the analysis by Campbell et al. (1997, p.311) using a slightly different annual dataset (n=105) prepared by Shiller[51]. Details of estimation results can be found in the files `ccapm.wf1` and `ccapm.xls`. Using OLS, the estimated equation is (p-values in parentheses)

$$\underset{}{y_t} = \underset{(.72)}{0.0057} + \underset{(.00)}{2.75}\, c_t + u_t \qquad R^2 = 0.31.$$

The estimate 2.75 is in a plausible range. However, OLS estimation is not appropriate since the regressor c_t is correlated with ϵ_t via b_t (unless γ=0). An IV-estimate of γ can be obtained using instruments which are assumed to be correlated with consumption growth. Campbell et al. use lags of the real interest rate i_t and the log dividend-price ratio d_t as instruments, arguing that ϵ_t is uncorrelated with any variables in the information set from $t-1$. Using 2SLS we obtain

$$\underset{}{y_t} = \underset{(.43)}{0.29} - \underset{(.53)}{11.2}\, c_t + e_t,$$

[51] http://www.econ.yale.edu/~shiller/data/ie_data.xls.

which shows that the estimated γ is insignificant (which saves us from the need to explain the unexpected negative sign). We use these instruments to test for their suitability, and subsequently, for testing the endogeneity of c_t. The first-stage regression of c_t on the instruments yields

$$\underset{}{c_t} = \underset{(.64)}{0.013} - \underset{(.46)}{0.042}\, i_{t-1} - \underset{(.78)}{0.0025}\, d_{t-1} + v_t \qquad R^2 = 0.006.$$

Obviously, the requirement of high correlation of instruments with the possibly endogenous regressor is not met. The F-statistic is 0.33 with a p-value of 0.72, and the instruments are considered to be very weak. nR^2<2 indicates that the IV-bias may be substantially larger than the OLS-bias.

To test the validity of the instruments based on overidentifying restrictions, we run the regression (e_t are the 2SLS residuals)

$$\underset{}{e_t} = \underset{(.71)}{0.155} - \underset{(.88)}{0.125}\, i_{t-1} + \underset{(.72)}{0.048}\, d_{t-1} + a_t \qquad R^2 = 0.0014.$$

The test statistic is 115·0.0014=0.16, with a p-value of 0.69 (m=1). We cannot reject the overidentifying restrictions (i.e. instruments are uncorrelated with 2SLS residuals e_t), and the instruments can be considered to be valid.

Despite this ambiguous evidence regarding the appropriateness of the instruments, we use the residuals from the first-stage regression to test for endogeneity:

$$\underset{}{y_t} = \underset{(.005)}{0.29} - \underset{(.024)}{11.2}\, c_t + \underset{(.005)}{14.1}\, v_t + w_t \qquad R^2 = 0.35.$$

The coefficient of v_t is significant at low levels, and we can firmly reject the H_0 of exogeneity of c_t (as expected).

This leaves us with conflicting results. On the one hand, we have no clear evidence regarding the appropriateness of the instruments (by comparing the first-stage regression and the overidentification test). On the other hand, we can reject exogeneity of c_t on theoretical grounds, but obtain no meaningful estimate for γ using 2SLS. From OLS we obtain a reasonable estimate for γ, but OLS would only be appropriate if c_t was truly exogenous (which is very doubtful).

In a related study, Yogo (2004) applies 2SLS to estimate the elasticity of intertemporal substitution, which is the reciprocal of the risk aversion coefficient γ for a specific choice of parameters in the Epstein-Zin utility function. He shows that using weak instruments (i.e. nominal interest rate, inflation, consumption growth, and the log dividend-price ratio lagged twice) leads to biased estimates and standard errors. Yogo's results imply that the lower end of a 95% confidence interval for γ is around 4.5 for the US, and not less than 2 across eleven developed countries.

> **Exercise 13:** Use the data from example 21 to replicate the analysis using the same instruments lagged by one *and* two periods.

> **Exercise 14:** Use the *monthly* data in the file `ie_data.xls` which is based on data prepared by Shiller[52] and Verbeek[53] to replicate the analysis from example 21.

[52] http://www.econ.yale.edu/~shiller
[53] http://eu.wiley.com/legacy/wileychi/verbeek2ed/

Exercise 15: Use the weekly data in the file `JEC.*` downloaded from the companion website of http://www.pearsonhighered.com/stock_watson/ to estimate a demand equation for the quantity of grain shipped. Use "price", "ice" and "seas1" to "seas12" as explanatory variables. Discuss potential endogeneity in this equation. Consider using "cartel" as an instrument. Discuss and test the appropriateness of "cartel" as an instrument.

1.10 Generalized method of moments[54]

Review 7:[55] In the **method of moments** the parameters of a distribution are estimated by equating sample and population moments. Given a sample $x_1, \ldots, x_n$ of independent draws from a distribution, the **moment condition**

$$\mathrm{E}[X - \mu] = 0$$

is replaced by the sample analog

$$\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x} = 0$$

to obtain the sample mean $\bar{x}$ as an estimate of μ. In general, to obtain estimates for a $K\times 1$ parameter vector $\boldsymbol{\theta}$ we have to consider K (sample) moment conditions

$$\mathrm{E}[m_j(X;\boldsymbol{\theta})] = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} m_{ij}(x_i;\hat{\boldsymbol{\theta}}) = 0 \qquad j = 1, \ldots, K,$$

where $m_{ij}(x_i;\hat{\boldsymbol{\theta}})$ are suitable functions of the sample and the parameter vector. The K parameters can be estimated by solving the system of K equations. The moment estimators are based on averages of functions. By the consistency of a mean of functions (see review 5) they converge to their population counterparts. They are consistent, but not necessarily efficient estimates. In many cases their asymptotic distribution is normal.

Example 22: The first and second central moments of the Gamma distribution with density

$$f(x) = \frac{(x/\beta)^{\alpha-1} e^{-x/\beta}}{\beta\Gamma(\alpha)}$$

are given by

$$\mathrm{E}[X] = \mu = \alpha\beta \quad \text{and} \quad \mathrm{E}[(X-\mu)^2] = \sigma^2 = \alpha\beta^2.$$

To estimate the two parameters α and β we define two functions

$$m_{i1} = x_i - ab \qquad m_{i2} = (x_i - ab)^2 - ab^2 = x_i^2 - (ab^2 + a^2b^2).$$

The two sample moment conditions

$$\frac{1}{n}\sum_{i=1}^{n} x_i - ab = \bar{x} - ab = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (ab^2 + a^2b^2) = 0$$

can be used to estimate a and b by solving the equations[56]

$$ab - \bar{x} = 0 \quad \text{and} \quad ab^2 - \tilde{s}^2 = 0,$$

which yields $b{=}\tilde{s}^2/\bar{x}$ and $a{=}\bar{x}^2/\tilde{s}^2$.

[54] This section is based on selected parts from Greene (2003), section 18. An alternative source is Cochrane (2001), sections 10 and 11. Note that Cochrane uses a different notation with $u_t{\equiv}\boldsymbol{m_t}$, $g_T{\equiv}\bar{\boldsymbol{m}}$, $S{\equiv}\boldsymbol{\Phi}$ and $d{\equiv}\boldsymbol{G}$.

[55] For details see Greene (2003), p.527 or Hastings and Peacock (1975), p.68.

[56] $\tilde{s}^2$ is the unadjusted sample variance $\tilde{s}^2{=}(1/n)\sum(x_i{-}\bar{x})^2$.

Example 23: We consider the problem of a time series of prices observed at irregularly spaced points in time (i.e. the intervals between observations have varying length). We want to compute mean and standard deviation of returns for a comparable (uniform) time interval by applying the method of moments (see file `irregular.xls` for a numerical example).

The observed returns are assumed to be determined by the following process:

$$Y(\tau_i) = \mu\tau_i + Z_i\sigma\sqrt{\tau_i} \qquad (i = 1, \ldots, n),$$

where τ_i is the length of the time interval (e.g. measured in days) used to compute the i-th return. Z_i is a pseudo-return, with mean zero and standard deviation one. μ and σ are mean and standard deviation of returns associated with the base interval. Assuming that Z_i and τ_i are independent (i.e. $\mathrm{E}[Z_i\sqrt{\tau_i}]=0$), we can take expectations on both sides, and replace these by sample averages to obtain

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y(\tau_i) = \mu\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n} \mu\tau_i,$$

from which we can estimate $\hat{\mu} = \bar{Y}/\bar{\tau}$.

To estimate the standard deviation σ we use

$$\frac{(Y(\tau_i) - \mu\tau_i)^2}{\tau_i} = Z_i^2\sigma^2.$$

Taking expectations and using sample averages we obtain (note that $\mathrm{E}[Z^2]=\mathrm{V}[Z]=1$):

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y(\tau_i) - \hat{\mu}\tau_i)^2}{\tau_i}.$$

1.10.1 OLS, IV and GMM

Generalized method of moments (GMM) can not only be used to estimate the parameters of a distribution, but also to estimate the parameters of an econometric model by generalizing the method of moments principle. GMM has its origins and motivation in the context of asset pricing and modeling rational expectations (see Hansen and Singleton, 1996). One of the main objectives was to estimate models without making strong assumptions about the distribution of returns.

We start by showing that the OLS estimator can be regarded as a method of moments estimator. Assumption **AX** in the context of the regression model $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon}$ implies the orthogonality condition

$$\mathrm{E}[\boldsymbol{X}'\boldsymbol{\epsilon}] = \mathrm{E}[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta})] = \mathbf{0}.$$

To estimate the $K\times 1$ parameter vector $\boldsymbol{\beta}$ we define K functions and apply them to each observation in the sample[57]

$$m_{ij}(\boldsymbol{b}) = x_{ij}(y_i - \boldsymbol{x}_i'\boldsymbol{b}) = x_{ij}e_i \qquad i = 1,\ldots,n;\ j = 1,\ldots,K.$$

The moment conditions are the sample averages

$$\bar{m}_j = \frac{1}{n}\sum_{i=1}^{n} m_{ij} = 0 \qquad j = 1,\ldots,K,$$

which are identical to the normal equations (2) which have been used to derive the OLS estimator in section 1.1:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i e_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{b}) = \frac{1}{n}\boldsymbol{X}'\boldsymbol{e} = \mathbf{0}.$$

If some of the regressors are (possibly) endogenous it is not appropriate to impose the orthogonality condition. Suppose there are instruments $\boldsymbol{Z}$ available for which $\mathrm{E}[\boldsymbol{Z}'\boldsymbol{\epsilon}]=\mathbf{0}$ holds. If $\boldsymbol{Z}$ has dimension $n\times K$ (the same as $\boldsymbol{X}$) we can obtain IV-estimates from

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{b}) = \mathbf{0}.$$

If there are more instruments than parameters we can specify the conditions

$$\frac{1}{n}\hat{\boldsymbol{X}}'\boldsymbol{e} = \mathbf{0},$$

where $\hat{\boldsymbol{X}}$ is defined in (35). Using $\hat{\boldsymbol{X}}$ generates K conditions, even when there are $L>K$ instruments[58].

[57] The notation $m_{ij}=m_j(y_i, x_{ij})$ is used to indicate the dependence of the j-th moment condition on the observation i.

[58] More instruments than necessary can be used to generate overidentifying restrictions and can improve the efficiency of the estimates.

The homoscedasticity assumption implies that the variance of residuals is uncorrelated with the regressors. This can be expressed as

$$\mathrm{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2] - \mathrm{E}[\boldsymbol{x}_i]\mathrm{E}[\epsilon_i^2] = \mathbf{0}.$$

If the model specification is correct the following expression

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i e_i^2 - \bar{\boldsymbol{x}}_i \tilde{s}_e^2$$

should be close to zero.
GMM can also be based on *conditional* moment restrictions of the form

$$\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \mathbf{0}.$$

This implies that $\boldsymbol{\epsilon}$ is not only uncorrelated with $\boldsymbol{X}$ but with any function of $\boldsymbol{X}$. Thus, depending on the way the conditional expectation is formulated, such conditions can be much stronger than unconditional restrictions. In a time series context, it can be assumed that the expectation of $\boldsymbol{\epsilon}$ conditional on past regressors is zero. Other examples are nonlinear functions of $\boldsymbol{X}$, or restrictions on the conditional variance. If $\boldsymbol{z}$ are regressors assumed to determine the (conditional) variance of disturbances, this can be expressed by the moment condition (see FGLS on p.1.8.1)

$$m_i(\boldsymbol{b}) = (y_i - \boldsymbol{x}_i'\boldsymbol{b})^2 - f(\boldsymbol{z}_i')\boldsymbol{b}_z.$$

1.10.2 Asset pricing and GMM

In example 21 we have considered the Euler equation

$$\mathrm{E}_{t-1}\left[(1 + R_t)\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma}\right] = 1,$$

and have shown how to estimate γ based on *linearizing* this equation. An alternative view is to consider the Euler equation as a testable restriction. It should hold for all assets and across all periods. This implies the following sample moment condition:

$$\frac{1}{n}\sum_{t=1}^{n} m_t(\delta, \gamma) = 0 \qquad m_t(\delta, \gamma) = (1 + R_t)\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma} - 1.$$

The returns of at least two assets are required to estimate the parameters δ and γ. Note that no linearization or closed-form solution of the underlying optimization problem is required (as opposed to the approach by Campbell et al. (1997) described in example 21). GMM can accommodate more conditions than necessary (i.e. additional instruments can be used to formulate overidentifying restrictions; see section 1.10.3).

In asset pricing or rational expectation models the errors[59] in expectations should be uncorrelated with all variables in the information set I_{t-1} of agents forming those expectations. This can be used to formulate orthogonality conditions for any instrument $z_{t-1} \in I_{t-1}$ in the following general way:

$$\mathrm{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})z_{t-1}] = 0.$$

The Euler equation in the consumption based CAPM is also expressed in terms of a conditional expectation. Thus, for any element of the information set

$$\frac{1}{n}\sum_{t=2}^{n} m_t(\delta, \gamma)z_{t-1} = 0$$

should hold.

In example 6 we have briefly described the Fama-MacBeth approach to estimate the parameters of asset pricing models. GMM provides an alternative (and possibly preferable) way to pursue that objective. We consider N assets with excess returns x_t^i, and a single-factor model with factor excess return x_t^m. The factor model implies that the following equations hold:

$$x_t^i = \beta_i x_t^m + \epsilon_t^i \quad i = 1, \ldots, N,$$

$$\mathrm{E}[x_t^i] = \lambda_m \beta_i \quad i = 1, \ldots, N.$$

Fama and MacBeth estimate β_i from the first equation for each asset. Given these estimates, λ_m is estimated from a single regression across the second set of equations (using sample means as observations of the dependent variable and estimated beta-factors as observations of the regressor). The CAPM or the APT imply a set of restrictions that should have zero expectation (at the true parameter values). The moment conditions corresponding to the first set of equations for the present example are

$$\frac{1}{n}\sum_{t=1}^{n}(x_{it} - \beta_i x_{mt})x_{mt} = 0 \quad i = 1, \ldots, N.$$

The second set of equations implies

$$\frac{1}{n}\sum_{t=1}^{n} x_{it} - \lambda_m \beta_i = 0 \quad i = 1, \ldots, N.$$

The generalization to several factors is straightforward.

[59] These errors are supposed to be evaluated at the true parameters.

1.10.3 Estimation and inference

Generalizing the method of moments we consider the case of $L>K$ moment conditions to estimate K parameters $\boldsymbol{\theta}$. Since there is no unique solution to the overdetermined system of equations we can minimize the sum of squares

$$\sum_{j=1}^{L} \bar{m}_j^2(\boldsymbol{\theta}) = \bar{\boldsymbol{m}}'\bar{\boldsymbol{m}} \qquad \bar{\boldsymbol{m}} = (\bar{m}_1, \ldots, \bar{m}_L)',$$

where

$$\bar{m}_j(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_{ij}(\boldsymbol{\theta}) \qquad j = 1, \ldots, L.$$

Minimizing this criterion gives consistent but not necessarily efficient estimates of $\boldsymbol{\theta}$. Hansen (1982) has considered estimates based on minimizing the *weighted* sum of squares

$$J = \bar{\boldsymbol{m}}'\boldsymbol{W}\bar{\boldsymbol{m}}.$$

The weight matrix $\boldsymbol{W}$ has to be positive definite. The choice of $\boldsymbol{W}$ relies on the idea of GLS estimators, with the intention to obtain efficient estimates. Elements of $\bar{\boldsymbol{m}}$ which are more precisely estimated should have a higher weight and have more impact on the value of the criterion function. If $\boldsymbol{W}$ is inversely proportional to the asymptotic covariance of $\bar{\boldsymbol{m}}$, i.e.

$$\boldsymbol{W} = \boldsymbol{\Phi}^{-1} \qquad \boldsymbol{\Phi} = \mathrm{aV}[\sqrt{n}\bar{\boldsymbol{m}}],$$

and $\operatorname{plim} \bar{\boldsymbol{m}}=\boldsymbol{0}$, the GMM estimates are consistent and efficient.

Before we proceed, we briefly refer to the asymptotic variance of the sample mean $\bar{y}$ (see review 5, p.22). It can be derived from observations y_i and is given by s^2/n where $s^2=(1/(n-1))\sum_i (y_i-\bar{y})^2$. Now, we note that $\bar{\boldsymbol{m}}$ can be viewed as a (multivariate) sample mean. It can be derived from

$$\bar{\boldsymbol{m}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}_i \qquad \boldsymbol{m}_i = m(y_i, \boldsymbol{x}_i),$$

where $\boldsymbol{m}_i$ is a $L\times 1$ vector of conditions evaluated at observation i. Similar to the asymptotic variance of the sample mean, the estimated covariance of $\bar{\boldsymbol{m}}$ can be based on the (estimated) covariance of $\boldsymbol{m}_i$:

$$\hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n-1}\sum_{i=1}^{n} \Big[\boldsymbol{m}_i(\hat{\boldsymbol{\theta}}) - \bar{\boldsymbol{m}}\Big]\Big[\boldsymbol{m}_i(\hat{\boldsymbol{\theta}}) - \bar{\boldsymbol{m}}\Big]'.$$

In general, the asymptotic covariance matrix of GMM parameter estimates can be estimated by

$$\hat{\boldsymbol{V}} = \frac{1}{n}\left(\hat{\boldsymbol{G}}'\hat{\boldsymbol{\Phi}}^{-1}\hat{\boldsymbol{G}}\right)^{-1}. \tag{39}$$

$\hat{\boldsymbol{G}}$ is the Jacobian of the moment functions (i.e. the matrix of derivatives of the moment functions with respect to the estimated parameters):

$$\hat{\boldsymbol{G}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \boldsymbol{m}_i(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}.$$

The columns of the $L \times K$ matrix $\hat{\boldsymbol{G}}$ correspond to the K parameters and the rows to the L moment conditions.

As shown in section 1.10.1, OLS and GMM lead to the same parameter estimates, if GMM is only based on the orthogonality condition $\mathrm{E}[\boldsymbol{X}'\boldsymbol{\epsilon}]$. However, if the covariance of parameters is estimated according to (39), $\hat{\boldsymbol{\Phi}}$ is given by

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{m}_i\boldsymbol{m}_i' = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{x}_i e_i \boldsymbol{x}_i' e_i = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i',$$

and $\hat{\boldsymbol{G}}$ is given by

$$\hat{\boldsymbol{G}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \boldsymbol{m}_i}{\partial \boldsymbol{b}'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{b})]}{\partial \boldsymbol{b}'} = -\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}.$$

Combining terms we find that the estimated covariance matrix for the GMM parameters of a regression model is given by

$$\frac{n}{n-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

which corresponds to White's heteroscedasticity consistent estimate (26). In this sense, estimating a regression model by GMM 'automatically' accounts for heteroscedasticity.

In practice, we need to take into account that $\hat{\boldsymbol{\Phi}}$ depends on the – yet to be determined – estimates $\hat{\boldsymbol{\theta}}$. The usually suggested two-step approach starts with an unrestricted estimate $\hat{\boldsymbol{\theta}}_u$ derived from using $\boldsymbol{W}{=}\boldsymbol{I}$, and minimizing $\bar{\boldsymbol{m}}'\bar{\boldsymbol{m}}$. The resulting estimates $\hat{\boldsymbol{\theta}}_u$ are used to construct $\hat{\boldsymbol{\Phi}}_u$, which is then used in the second step to minimize

$$J = \bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Phi}}_u^{-1}\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}).$$

The asymptotic properties of GMM estimates can be derived on the basis of a set of assumptions (see Greene, 2003, p.540). Among others, the empirical moments are assumed to obey a central limit theorem. They are assumed to have a finite covariance matrix $\boldsymbol{\Phi}/n$, so that

$$\sqrt{n}\bar{\boldsymbol{m}} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Phi}].$$

Under this and further assumptions (see Greene, 2003, p.540) it can be shown that the asymptotic distribution of GMM estimates is normal, i.e.

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathrm{N}[\boldsymbol{\theta}, \boldsymbol{V}].$$

The diagonal elements of the estimated covariance matrix $\hat{\boldsymbol{V}}$ can be used to compute t-statistics for the parameter estimates:

$$\frac{\hat{\theta}_j}{\sqrt{\hat{\boldsymbol{V}}_{jj}}} \overset{a}{\sim} \mathrm{N}(0,1).$$

Alternative estimators like the White or the Newey-West estimator can be used if required (see Cochrane, 2001, p.220).

Overidentifying restrictions can be tested on the basis of $nJ \sim \chi^2_{L-K}$. Under the null hypothesis, the restrictions are valid, and the model is correctly specified. Invalid restrictions lead to high values of J and to a rejection of the model. In the just identified case $L=K$ and $J=0$.

Despite this relatively brief description of GMM, its main advantages should have become clear. GMM does not rely on **Aiid**, requires no distributional assumptions, it may also be based on conditional moments, and allows for more conditions than parameters to be estimated (i.e. it can be used to formulate and test overidentifying restrictions). The requirements for consistency and asymptotic normality are that $\boldsymbol{m}_i$ must be well behaved (i.e. stationary and ergodic), and the empirical moments must have a finite covariance matrix.

These advantages are not without cost, however. Some of the problems (which have received insufficient space in this short treatment) associated with GMM are: In some cases the first derivative of J may not be known analytically and the optimization of the criterion function J has to be carried out numerically. Moreover, J is not necessarily a convex function which implies that there is no unique minimum, and good starting values are very important for the numerical search algorithm.

1.10.4 Example 24: Models for the short-term interest rate

Chan et al. (1992) use GMM to estimate several models for the short-term interest rate. They consider a general case where the short rate follows the diffusion

$$dr = (\alpha + \beta r)dt + \sigma r^{\gamma} dZ.$$

By imposing restrictions on the parameters special cases are obtained (e.g. the Vasicek model if γ=0, or the Brennan-Schwartz model if γ=1). The discrete-time specification of the model is given by

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \epsilon_t \qquad \mathrm{E}[\epsilon_t] = 0 \qquad \mathrm{E}[\epsilon_t^2] = \sigma^2 r_{t-1}^{2\gamma}.$$

Using $\boldsymbol{\theta}$=$(\alpha\ \beta\ \sigma\ \gamma)'$, Chan et al. impose the following moment conditions

$$\boldsymbol{m}_t(\boldsymbol{\theta}) = \left[\epsilon_t \quad \epsilon_t r_{t-1} \quad \epsilon_t^2 - \sigma^2 r_{t-1}^{2\gamma} \quad (\epsilon_t^2 - \sigma^2 r_{t-1}^{2\gamma}) r_{t-1}\right]'.$$

Conditions one and three correspond to the mean and variance of ϵ_t. Conditions two and four impose orthogonality between the regressor r_{t-1} and the error from describing the variance of the disturbances ϵ_t. The estimated covariance of the parameter estimates is based on the following components of the Jacobian (rows correspond to conditions, columns to parameters):

$$G_{t,1} = \left[\frac{\partial m_{t,1}}{\partial \alpha} = -1 \quad \frac{\partial m_{t,1}}{\partial \beta} = -r_{t-1} \quad \frac{\partial m_{t,1}}{\partial \sigma} = 0 \quad \frac{\partial m_{t,1}}{\partial \gamma} = 0\right]$$

$$G_{t,2} = \left[\frac{\partial m_{t,2}}{\partial \alpha} = -r_{t-1} \quad \frac{\partial m_{t,2}}{\partial \beta} = -r_{t-1}^2 \quad 0 \quad 0\right]$$

$$G_{t,3} = \left[-2m_{t,1} \quad -2m_{t,1} r_{t-1} \quad -2\sigma r_{t-1}^{2\gamma} \quad -2\sigma^2 r_{t-1}^{2\gamma} \ln(r_{t-1})\right]$$

$$G_{t,4} = \left[-2m_{t,1} r_{t-1} \quad -2m_{t,1} r_{t-1}^2 \quad -2\sigma r_{t-1}^{2\gamma+1} \quad -2\sigma^2 r_{t-1}^{2\gamma+1} \ln(r_{t-1})\right].$$

Chan et al. (1992) use monthly observations of the three-month rate for r_t from June 1964 to December 1989. Details of computations and some estimation results can be found in the file `ckls.xls`. Note that the estimates for α, β and σ^2 have to be scaled by Δt=1/12 to convert them into annual terms, and to make them comparable to the results presented in Chan et al. (1992).

Exercise 16: Retrieve a series of short-term interest rates from the website `http://www.federalreserve.gov/Releases/H15/data.htm` or from another source. Estimate two or three different models of the short-term interest rate by GMM.

1.11 Models with binary dependent variables

Review 8: The binomial distribution describes the probabilities associated with a sequence of n independent trials, where each trial has two possible outcomes (usually called success and failure). The probability of success p is the same in each trial. The probability of y successes in n trials is given by

$$f(y) = \binom{n}{y} p^y (1-p)^{(n-y)}.$$

Expected value and variance of a binomial random variable are given by np and $np(1-p)$, respectively. If the number of trials in a binomial experiment is large (e.g. $np \geq 5$), the binomial distribution can be approximated by the normal distribution. If $n=1$, the binomial distribution is a Bernoulli distribution.

We now consider the application of regression analysis to the case of binary dependent variables. This applies, for example, when the variable of interest is the result of a choice (e.g. brand choice or choosing means of transport), or an interesting event (e.g. the default of a company or getting unemployed). For simplicity we will only consider the binary case but the models discussed below can be extended to the multinomial case (see Greene (2003), section 21.7).

Observations of the dependent variable y indicate whether the event or decision has taken place or not ($y=1$ or $y=0$). The probability for the event is assumed to depend on regressors $\boldsymbol{X}$ and parameters $\boldsymbol{\beta}$, and is expressed in terms of a distribution function F. For a single observation i we specify the conditional probabilities

$$\mathrm{P}[y_i = 1] = F(\boldsymbol{x}_i, \boldsymbol{\beta}) \qquad \mathrm{P}[y_i = 0] = 1 - F(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

The conditional expectation of y_i is a weighted average of the two possible outcomes:

$$\mathrm{E}[y_i|\boldsymbol{x}_i] = \hat{y}_i = 1 \cdot \mathrm{P}[y_i = 1] + 0 \cdot \mathrm{P}[y_i = 0] = F(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

There are several options to formalize F. In the linear model $F(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}$, and the corresponding regression model is given by

$$y_i = \mathrm{E}[y_i|\boldsymbol{x}_i] + (y_i - \mathrm{E}[y_i|\boldsymbol{x}_i]) = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i = \hat{y}_i + \epsilon_i.$$

The linear model has three major drawbacks. First, $\boldsymbol{x}_i'\boldsymbol{\beta}$ is not constrained to the interval [0,1]. Second, the disturbances are not normal but Bernoulli random variables with two possible outcomes (*conditional* on $\boldsymbol{x}_i$):

$$\mathrm{P}[\epsilon_i = -\boldsymbol{x}_i'\boldsymbol{\beta}] = \mathrm{P}[y_i = 0] = 1 - \boldsymbol{x}_i'\boldsymbol{\beta} \qquad \mathrm{P}[\epsilon_i = 1 - \boldsymbol{x}_i'\boldsymbol{\beta}] = \mathrm{P}[y_i = 1] = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

This implies the third drawback that the disturbances are heteroscedastic with conditional variance

$$\mathrm{V}[\epsilon_i|\boldsymbol{x}_i] = \boldsymbol{x}_i'\boldsymbol{\beta}(1 - \boldsymbol{x}_i'\boldsymbol{\beta}).$$

Instead of specifying F as a linear function, in the **probit-model** F is assumed to be the standard normal distribution function

$$F(\boldsymbol{x}_i'\boldsymbol{\beta}) = \Phi(\hat{y}_i) = \int_{-\infty}^{\hat{y}_i} \phi(u)\, du = \int_{-\infty}^{\hat{y}_i} \frac{1}{\sqrt{2\pi}} \exp\{-0.5u^2\} du.$$

In the **logit-model** or **logistic regression** model the distribution function is given by

$$F(\boldsymbol{x}_i'\boldsymbol{\beta}) = L(\hat{y}_i) = \frac{1}{1+\exp\{-\boldsymbol{x}_i'\boldsymbol{\beta}\}} = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}.$$

The probit- and logit-models imply (slightly different) s-shaped forms of the conditional expectation $\mathrm{E}[y_i|\boldsymbol{x}_i]$. The logit-model assigns larger probabilities to y_i=0 than the probit-model if $\boldsymbol{x}_i'\boldsymbol{\beta}$ is very small. The difference between the two models will be large if the sample has only a few cases for which y_i=1 (or y_i=0), and if an important regressor has a large variance (see Greene, 2000, p.667).

The interpretation of the coefficients in the three models can be based on the partial derivatives with respect to regressor j:

$$\frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial x_{ij}} = \beta_j$$

$$\frac{\partial \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial x_{ij}} = \phi(\boldsymbol{x}_i'\boldsymbol{\beta})\beta_j$$

$$\frac{\partial L(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial x_{ij}} = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{(1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\})^2}\beta_j.$$

Hence, in the probit- and logit-model the effect of a change in regressor j depends on the probability at a given value of $\boldsymbol{x}_i'\boldsymbol{\beta}$. A convenient interpretation of the logit-model is based on the so-called **odds-ratio**, which is defined as

$$\frac{L(\hat{y}_i)}{1-L(\hat{y}_i)} \qquad L(\hat{y}_i) = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}.$$

The log odds-ratio is given by

$$\ln\left(\frac{L(\hat{y}_i)}{1-L(\hat{y}_i)}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

This implies that $\exp\{\beta_j \Delta x_j\}$ is the factor by which the odds-ratio is changed c.p. if regressor j is changed by Δx_j units. The effect of a change in a regressor on $L(\hat{y}_i)$ is low if $L(\hat{y}_i)$ is close to zero or one.

Binary choice models can be estimated by maximum-likelihood, whereas linear models are usually estimated by least squares. Each observation in the sample is treated as a random draw from a Bernoulli distribution. For a single observation the (conditional) probability of observing y_i is given by

$$\mathrm{P}[y_i|\boldsymbol{x}_i,\boldsymbol{\beta}] = F(\boldsymbol{x}_i,\boldsymbol{\beta})^{y_i}(1-F(\boldsymbol{x}_i,\boldsymbol{\beta}))^{(1-y_i)} = F_i^{y_i}(1-F_i)^{(1-y_i)}.$$

For the entire sample the joint probability is given by

$$\prod_{i=1}^{n} F_i^{y_i}(1-F_i)^{(1-y_i)},$$

and the log-likelihood function is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln F_i + (1-y_i)\ln(1-F_i).$$

Maximizing the log-likelihood requires an iterative procedure. Standard errors of estimated coefficients in the logit-model can be based on the Hessian

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} L(\hat{y}_i)(1-L(\hat{y}_i))\boldsymbol{x}_i \boldsymbol{x}_i',$$

and tests of the estimated coefficients make use of the asymptotic normality of ML estimators.

A likelihood ratio test of m restrictions $\boldsymbol{r}'\boldsymbol{\beta}=\mathbf{0}$ is based on comparing the restricted likelihood ℓ_r to the unrestricted likelihood ℓ_u:

$$2[\ell_u - \ell_r] \sim \chi^2_m.$$

The goodness of fit cannot be measured in terms of R^2. McFadden's R^2 is based on ℓ_u and the log-likelihood ℓ_0 of a model which only contains a constant term:

$$\text{McFadden } R^2 = 1 - \ell_u/\ell_0.$$

Example 25: We consider the choice among mortgages with fixed and adjustable interest rates analyzed by Dhillon et al. (1987), and use part of their data from Studenmund (2001), p.459[60]. The dependent variable is ADJUST (equal to 1 when an adjustable rate has been chosen). The regressors are the fixed interest rate (FIXED), the interest premium on the adjustable rate (PREMIUM), the net worth of the borrower (NET), the ratio of the borrowing costs (adjustable over fixed; POINTS), the ratio of the adjustable rate maturity to that of the fixed rate (MATURITY) and the difference between the 10-year and 1-year Treasury rate (YIELD). Details can be found in the files `mortgage.wf1` and `mortgage.xls`.

The estimation results from the linear and the logit-model are summarized in the table below. The z-values are based on the Hessian matrix. The p-values from the two models are only marginally different. The coefficients of PREMIUM, NET and YIELD are significant at the 5% level. The fitted probabilities from both models are very similar which is confirmed by the similarity of R^2 and McFadden's R^2. The linear model's probabilities are negative in only two cases, and never greater than one.

[60] The data is available from the Student Resources at `http://wps.aw.com/aw_studenmund_useecon_5`.

	coefficients		t-, z- and LR-test			p-values		
	linear	logit	linear	logit z	logit LR	linear	z	LR
constant	–0.083	–3.722	–0.064	–0.514	0.268	0.949	0.607	0.605
FIXED	0.161	0.902	1.963	1.859	3.699	0.054	0.063	0.054
PREMIUM	–0.132	–0.708	–2.643	–2.331	6.386	0.010	0.020	0.012
NET	0.029	0.149	2.437	1.906	4.824	0.017	0.057	0.028
POINTS	–0.088	–0.518	–1.242	–1.217	1.595	0.218	0.224	0.207
MATURITY	–0.034	–0.238	–0.179	–0.229	0.053	0.858	0.819	0.819
YIELD	–0.793	–4.110	–2.451	–2.159	5.378	0.017	0.031	0.020

R^2=0.314; McFadden R^2=0.26

The coefficients from the linear model have the usual interpretation. The coefficient -0.708 from the logit-model is transformed to $\exp\{-0.708\}$=0.49, and can be interpreted as follows: the odds-ratio is about one half of its original value if the premium increases c.p. by one unit. For the first observation in the sample we obtain a fitted probability of 0.8 which corresponds to an odds-ratio of 4:1. If the premium changes from its current value of 1.5 to 2.5 the odds-ratio will fall to 2:1. From the linear model the corresponding change yields a drop in $\hat{y}$ from 0.78 to 0.65.

1.12 Sample selection[61]

Consider two random variables $y \sim \mathrm{N}(\mu, \sigma^2)$ and $z \sim \mathrm{N}(\mu_z, \sigma_z^2)$ with correlation ρ_{yz}. Suppose y is only observed if $z > a$ (so-called incidental truncation). The expected value of y conditional on truncation is given by

$$\mathrm{E}[y|\text{truncation}] = \mu + \rho_{yz}\sigma\lambda(\alpha_z),$$

where (see Greene, 2003, p.781) $\alpha_z = (a - \mu_z)/\sigma_z$, and $\lambda(\alpha_z)$ is the so-called **inverse Mills ratio** given by $\lambda(\alpha_z) = f(\alpha_z)/[1 - F(\alpha_z)]$, where $f(\cdot)$ denotes the normal pdf and $F(\cdot)$ the normal cdf. For example, if we can only observe the income y of people whose wealth z is below a (and $\rho_{yz} > 0$), the average of the sample income is lower than the 'true' average income in the population.

A similar argument holds for a regression, i.e. for the conditional expectation

$$y_i = \hat{y}_i + \epsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

$$\mathrm{E}[\hat{y}_i|\text{truncation}] = \boldsymbol{x}_i'\boldsymbol{\beta} + \rho_{yz}\sigma_\epsilon\lambda(\alpha_z).$$

A similar result holds if z is not a (correlated) random variable but determined by an equation like

$$z_i = \boldsymbol{w}_i'\boldsymbol{\gamma} + u_i = \hat{z}_i + u_i.$$

If sample data can only be observed conditional on some mechanism related to z, the conditional mean of y (now subject to *selection*) is given by

$$\mathrm{E}[\hat{y}_i|\text{selection}] = \boldsymbol{x}_i'\boldsymbol{\beta} + \rho_{\epsilon u}\sigma_\epsilon\lambda_i(\alpha_{u_i}),$$

where $\alpha_{u_i} = -\hat{z}_i/\sigma_u$ and $\lambda_i(\alpha_{u_i}) = f(\hat{z}_i/\sigma_u)/F(\hat{z}_i/\sigma_u)$. This result is obtained by assuming bivariate normality of ϵ and u (rather than y and z). Note that the inverse Mills ratio $\lambda_i(\cdot)$ is not a constant, but depends on $\boldsymbol{w}_i'\boldsymbol{\gamma}$. Estimating the equation $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$ without $\lambda_i(\cdot)$ yields inconsistent estimates because of the omitted regressor, or, equivalently, as a result of sample selection.[62] Note that a non-zero correlation among ϵ and u determines the bias/inconsistency. Thus, a special treatment is required when the *unobservable* factors determining inclusion in the subsample are correlated with the *unobservable* factors affecting the variable of primary interest.

In many cases, z is not directly observed/observable, but only a *binary* variable d, indicating the consequence of the z-based selection rule. This offers the opportunity to estimate the so-called selection equation (using a logistic regression as described in section 1.11):

$$d_i = \boldsymbol{w}_i'\boldsymbol{\gamma} + v_i.$$

[61] Most of this section is based on Greene (2003), section 22.4.

[62] The resulting inconsistency cannot be 'argued away' by stating that the estimated incomplete equation is representative for the population corresponding to that available subsample. Since the estimated equation describes (only) the non-random subsample, such a viewpoint is rather useless as long as nothing is known about the mechanism that determines whether y (in the population) is non-zero.

This forms the basis for the so-called **Heckman correction**. Heckman (1979) suggested a two-step estimator[63] which first estimates the selection equation by probit to obtain $\lambda_i = f(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}})/F(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}})$ for every i, and then estimates the original equation after adding this auxiliary regressor to the equation.[64] The coefficient of λ_i can be interpreted by noting that it is an estimate of the term $\rho_{\epsilon u}\sigma_\epsilon$; i.e. it can be viewed as a scaled correlation coefficient.[65]

For the practical implementation of this two-step approach we note that y_i and $\boldsymbol{x}_i$ are only observed if $d_i=1$, while the regressors $\boldsymbol{w}_i$ must be observed for *all* cases. The information in $\boldsymbol{w}_i$ must be able to sufficiently discriminate among subjects who enter or do not enter the sample. More importantly, the selection equation requires at least one (additional) exogeneous regressor which is not included in $\boldsymbol{x}_i$. In other words, we impose an exclusion condition on the main equation, and this additional regressor plays a similar role as an instrument in case of IV regressions for treating endogeneity. Note that IV-estimation is *impossible* when the regressors in the first stage are identical to those in the main equation (because of perfect multicollinearity). In the Heckit approach it is feasible to set $\boldsymbol{w}_i=\boldsymbol{x}_i$ (because of the nonlinearity of the inverse Mills ratio, and the fact that a different number of observations is used in the two equations) but not recommended (see Wooldridge, 2002, p.564).

> **Example 26:** We consider a well-known and frequently used dataset about female labor force participation and wages, and replicate the results in Table 22.7 in Greene (2003).[66] A wage model can only be estimated for those 428 females who actually have a job, so that wage data can be observed.[67] One can view the absent wage information for another 325 females in this dataset as the result of truncation: if the offered wage is below the reservation wage, females are not actively participating in the labor market.
> Estimation results based on the available sample of 428 females may suffer from a selection bias if unobserved effects in the wage and selection equations are correlated (i.e. $\rho_{\epsilon u}\neq 0$; see above). Whether this bias results from so-called self-selection (i.e. women's deliberate choice to participate in the labor market), or other sampling effects is irrelevant for the problem, but may be important for choosing regressors $\boldsymbol{w}_i$.
> The estimated coefficient of the inverse Mills ratio is given by -1.1. This can be interpreted as follows: women who have above average willingness (interest or tendency) to work (i.e. z_i is above $\hat{z}_i$; $u_i>0$) tend to earn below average wage (i.e. y_i is below $\hat{y}_i$; $\epsilon_i<0$). This estimate is statistically insignificant which indicates that sample selection may not play an important role in this example.

[63] The procedure is often called 'Heckit', because of the combination of the name Heckman and logit/probit models.

[64] λ_i are also called 'generalized residuals' of a probit model. For the entire sample (not just the subsample for which y is observed) they have mean zero and are uncorrelated with the regressors $\boldsymbol{w}_i$.

[65] It may be possible to assign a 'physical' meaning to this coefficient upon recalling that the slope in a (simple) regression of y on x is given by $\rho_{yx}\sigma_y/\sigma_x$ (see p.2). Thus, the ratio's coefficient is proportional to the slope of a regression of ϵ_i on u_i (i.e. the slope is multiplied/scaled by σ_u).

[66] Source and description of variables: `https://rdrr.io/rforge/Ecdat/man/Mroz.html`; this dataset `Mroz87` is also included in the R-package `sampleSelection`. `sample-selection.R` contains code for Heckit estimates (two-stage and ML), as well as an extension which also deals with endogeneity.

[67] For the purpose of this example we ignore the potential endogeneity associated with estimating a wage equation (see example 20). See Wooldridge (2002, p.567) for a treatment of this case.

1.13 Duration models[68]

The purpose of **duration analysis** (also known as **event history analysis** or **survival analysis**) is to analyze the length of time (the duration) of some phenomenon of interest (e.g. the length of being unemployed or the time until a loan defaults). A straightforward application of regression models using observed durations as the dependent variable is inappropriate, however, because duration data are typically **censored**. This means that the *actual* duration cannot be recorded for some elements of the sample. For example, some people in the sample are still unemployed at the time of analysis, and it is unknown when they are going to become employed again (if at all). We can only record the length of the unemployment period at the time the observation is made. Such records are censored observations and this fact must be taken into account in the analysis (see below). Two cases are possible: the subject under study is still in the interesting state when measurements are made, and it is unknown how long it will continue to stay in that state (right censoring). Left censoring holds, if the subject has already been in the interesting state before the beginning of the study, and it is unknown for how long.

We define the (continuous) variable T which measures the length of time spent in the interesting state, or the time until the event of interest has occurred. The units of measurement will usually be days, weeks or months, but T is not constrained to integer values. The distribution of T is described by a cumulative distribution function

$$F(t) = \mathrm{P}[T \leq t] = \int_0^t f(s)\, ds.$$

The **survival** or **survivor function** is the probability of being in the interesting state for more than t units of time:

$$S(t) = 1 - F(t) = \mathrm{P}[T \geq t].$$

We now consider the conditional probability of leaving the state of interest between t and $t+h$ conditional on having 'survived' until t:

$$\mathrm{P}[t \leq T \leq t+h | T \geq t] = \frac{\mathrm{P}[t \leq T \leq t+h]}{\mathrm{P}[T \geq t]} = \frac{F(t+h) - F(t)}{1 - F(t)} = \frac{F(t+h) - F(t)}{S(t)}.$$

This probability is used to define the **hazard function**

$$\lambda(t) = \lim_{h \to 0} \frac{\mathrm{P}[t \leq T \leq t+h | T \geq t]}{h}.$$

$\lambda(t)$ does not have a straightforward interpretation. It may be viewed as an instantaneous probability of leaving the state. However, to view it as a probability is not quite appropriate, since $\lambda(t)$ can be greater than one (in fact, it has no upper bound). If we assume that the hazard rate is a constant λ and assume that the event is repeatable, then λ is the expected number of events per unit of time. Alternatively, a constant hazard rate implies $\mathrm{E}[T]=1/\lambda$, which is the expected number of periods until the state is left.

[68] Most of this section is based on Greene (2003), section 22.5 and Kiefer (1988).

The hazard rate can be expressed in terms of $F(t)$, $f(t)$ and $S(t)$. Since

$$\lim_{h\to 0} \frac{F(t+h)-F(t)}{h} = F'(t) = f(t),$$

the hazard rate can also be written as

$$\lambda(t) = \frac{f(t)}{S(t)} \quad \text{or} \quad f(t) = \lambda(t)S(t).$$

It can be shown that

$$F(t) = 1 - \exp\left\{-\int_0^t \lambda(s)\,ds\right\}.$$

A constant hazard rate $\lambda(t)$=γ corresponds to an exponential distribution $F(t)$=$1-e^{-\gamma t}$. It implies that the probability of leaving the interesting state during the next time interval does not depend on the time spent in the state. This may not always be a realistic assumption. Instead, assuming a **Weibull distribution** for T results in the hazard rate

$$\lambda(t) = \gamma\alpha t^{\alpha-1} \qquad \gamma > 0,\ \alpha > 0,$$

which is increasing if α>1. Assuming a lognormal or log-logistic distribution for T gives rise to a non-monotonic behavior of the hazard rate.
The parameters $\boldsymbol{\theta}$=$\{\alpha, \gamma\}$ can be estimated by maximum likelihood. The joint density for an i.i.d. sample of n *uncensored* durations t_i is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}) \qquad f(t) = \lambda(t)S(t).$$

When t_i is a right-censored observation, we only know that the actual duration t_i^* is at least t_i. As a consequence, the contribution to the likelihood is the probability that the duration is longer than t_i, which is given by the survivor function $S(t_i)$. Using the dummy variable d_i=1 to indicate uncensored observations, the log-likelihood is defined as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \ln f(t_i, \boldsymbol{\theta}) + (1-d_i)\ln S(t_i, \boldsymbol{\theta}).$$

Because $f(t)$=$\lambda(t)S(t)$ the log-likelihood can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \ln \lambda(t_i, \boldsymbol{\theta}) + \ln S(t_i, \boldsymbol{\theta}).$$

In other words, the likelihood of observing a duration of length t_i depends on survival until t_i, and exiting the interesting state at t_i. For censored cases, exiting cannot be accounted for, and only survival until t_i enters the likelihood.

In case of the Weibull distribution $\lambda(t)=\gamma\alpha t^{\alpha-1}$, $\ln S(t)=-\gamma t^{\alpha}$, and the log-likelihood is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \ln \gamma\alpha t_i^{\alpha-1} - \gamma t_i^{\alpha}.$$

An obvious extension of modeling durations makes use of explanatory variables[69]. This can be done by replacing the constant γ by the term $\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}$.[70] The resulting parameter estimates can be interpreted in terms of $\exp\{\Delta\beta_j\}$, which is the factor by which the hazard rate is multiplied if regressor j is increased ceteris paribus by Δ units.

The **proportional hazards model** (or **Cox regression model**) does not require any assumption about the distribution of T. Rather than modeling the hazard rate as

$$\lambda(t, \boldsymbol{x}_i) = \lambda_0(t) \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\},$$

where $\lambda_0(t)$ is the **baseline hazard function** (e.g. $\alpha t^{\alpha-1}$ for the Weibull model), the Cox regression assumes that the ratio of the hazard rates of two individuals does not depend upon time:

$$\frac{\lambda(t, \boldsymbol{x}_i)}{\lambda(t, \boldsymbol{x}_j)} = \frac{\lambda_0(t) \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\lambda_0(t) \exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}}.$$

Hence, there is no need to specify the baseline function $\lambda_0(t)$. Cox defines a **partial likelihood** estimator using the log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\boldsymbol{x}_i'\boldsymbol{\beta} - \ln \sum_{j \in R_i} \exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\} \right].$$

For a sample of n distinct exit times $t_1, \ldots, t_n$ (i.e. considering uncensored cases only), the risk set R_i contains all individuals whose exit time is at least t_i (which includes censored and uncensored cases).

> **Example 27:** We consider a dataset about lung cancer from the North Central Cancer Treatment Group.[71] Ignoring (available) covariates and assuming a Weibull distribution results in estimates of $\hat{\alpha}$=1.342 and $\hat{\gamma}$=0.0003. The function `survreg` in the R-package `survival` reports a constant term 6.054, which can be derived from $-\ln\hat{\gamma}/\hat{\alpha}$.
>
> Using the available regressors and maintaining the Weibull assumption shows that sex, ph.ecog and ph.karno are significant covariates. The estimate for sex can be converted to $\exp\{-0.56\}$=0.571, which implies that the hazard rate for otherwise identical observations is nearly halved when comparing a man (sex=1) to a female (sex=2). The estimate 0.0235 for ph.karno implies that an increase of this variable by (a typical change of) 10 units yields a factor of $\exp\{10 \cdot 0.0235\}$=1.265, i.e. an approximately 25% increase in the hazard rate. Running a Cox regression yields very similar parameter estimates.

[69] In the context of hazard rate models the regressors are frequently called **covariates**.

[70] Note that the function `survreg` in the R-package `survival` sets $\gamma=\exp\{-\boldsymbol{x}_i'\boldsymbol{\beta}\}$, and applies a scaling factor.

[71] Source and description of variables:
https://www.rdocumentation.org/packages/survival/versions/2.41-2/topics/lung. Computations and code can be found in `lung.xlsx` and `lung.R`.

2 Time Series Analysis

2.1 Financial time series

A financial time series is a chronologically ordered sequence of data observed on financial markets. These include stock prices and indices, interest rates, exchange rates (prices for foreign currencies), and commodity prices. Usually the subject of financial studies are **returns** rather than prices. Returns summarize an investment irrespective of the amount invested, and financial theories are usually expressed in terms of returns.

Log returns y_t are calculated from prices p_t using

$$y_t = \ln p_t - \ln p_{t-1} = \ln(p_t/p_{t-1}).$$

This definition corresponds to *continuous* compounding. p_t is assumed to include dividend or coupon payments. **Simple returns** r_t are computed on the basis of relative price changes:

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1.$$

This definition corresponds to *discrete* compounding. Log and simple returns are related as follows:

$$y_t = \ln(1 + r_t) \qquad r_t = \exp\{y_t\} - 1.$$

A Taylor series expansion of r_t shows that the two return definitions differ with respect to second and higher order terms:

$$r_t = \exp\{y_t\} - 1 = \sum_{i=0}^{\infty} \frac{y_t^i}{i!} - 1 = \sum_{i=1}^{\infty} \frac{y_t^i}{i!} = y_t + \sum_{i=2}^{\infty} \frac{y_t^i}{i!}.$$

The simple return of a **portfolio** of m assets is a weighted average of the simple returns of individual assets

$$r_{p,t} = \sum_{i=1}^{m} w_i r_{it},$$

where w_i is the weight of asset i in the portfolio. For log returns this relation only holds approximately:

$$y_{p,t} \approx \sum_{i=1}^{m} w_i y_{it}.$$

Some financial models focus on returns and their statistical properties aggregated over time. **Multi-period** log returns are the sum of single-period log returns. The h-period log return $(\ln p_t - \ln p_{t-h})$ is given by

$$\ln p_t - \ln p_{t-h} = \ln(p_t/p_{t-1}) + \ln(p_{t-1}/p_{t-2}) + \cdots + \ln(p_{t-h+1}/p_{t-h})$$

$$y_t(h) = y_t + y_{t-1} + \cdots + y_{t-h+1}.$$

The corresponding expression for simple returns is

$$p_t/p_{t-h} = (p_t/p_{t-1})(p_{t-1}/p_{t-2}) \cdots (p_{t-h+1}/p_{t-h})$$

$$1 + r_t(h) = (1 + r_t)(1 + r_{t-1}) \cdots (1 + r_{t-h+1}) = \prod_{j=0}^{h-1} (1 + r_{t-j}).$$

2.1.1 Descriptive statistics of returns

Basic statistical properties of returns are described by mean, standard deviation, skewness and kurtosis. The mean is estimated from a sample of log returns y_t (t=1, ..., n) using

$$\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t.$$

The mean $\bar{r}$ of simple returns (obtained from *the same* price series) is *not equal* to $\bar{y}$. An approximate[72] relation between the two means is

$$\bar{r} \approx \exp\{\bar{y} + 0.5s^2\} - 1 \qquad \bar{y} \approx \ln(1 + \bar{r}) - 0.5s^2, \tag{40}$$

where s^2 is the (sample) **variance** of log returns:

$$s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y})^2.$$

The square root of s^2 is the (sample) **standard deviation** or **volatility**[73]. Examples 28 and 29 document the well-known fact that the variance (or volatility) of returns is not constant over time (i.e. the heteroscedasticity of financial returns).

> **Example 28:** Figure 2 shows the stock prices of IBM[74] and its log returns. Log and simple returns cannot be distinguished in such graphs. Obvious features are the erratic, strongly oscillating behavior of returns around the more or less constant mean, and the increase in the volatility towards the end of the sample period.

> **Example 29:** Figure 3 shows the daily log returns of IBM[75] over a long period of time (1962–1997). This series shows that temporary increases in volatility as in Figure 2 are very common. This phenomenon is called **volatility clustering** and can be found in many return series.

[72] The relation is exact if log returns are normally distributed (see section 2.1.2).

[73] In the context of financial economics the term volatility is frequently used in place of the statistical term standard deviation. Volatility usually refers to the standard deviation expressed in *annual* terms.

[74] Source: Box and Jenkins (1976), p.526; see file `ibm.wf1`; daily data from 1961/5/17 to 1962/11/2; 369 observations.

[75] Source: Tsay (2002), p.257; daily data from 1962/7/3 to 1997/12/31; 8938 observations; available from `http://faculty.chicagobooth.edu/ruey.tsay/teaching/fts2/`.

Figure 2: Daily stock prices of IBM and its log returns 1961–1962.

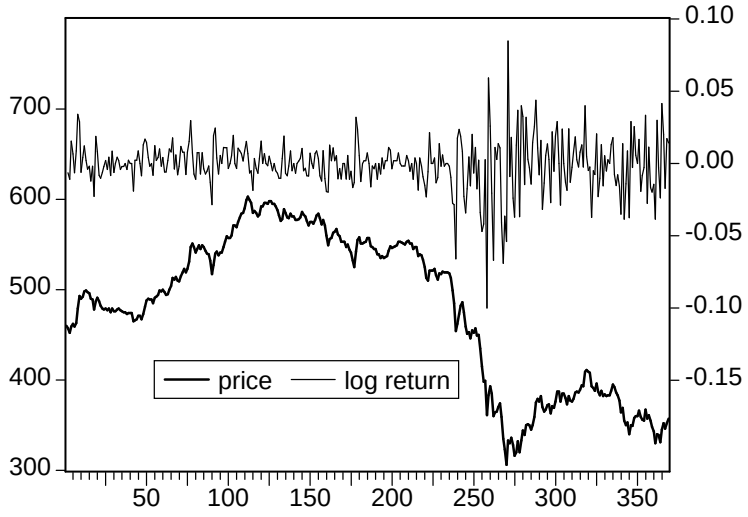
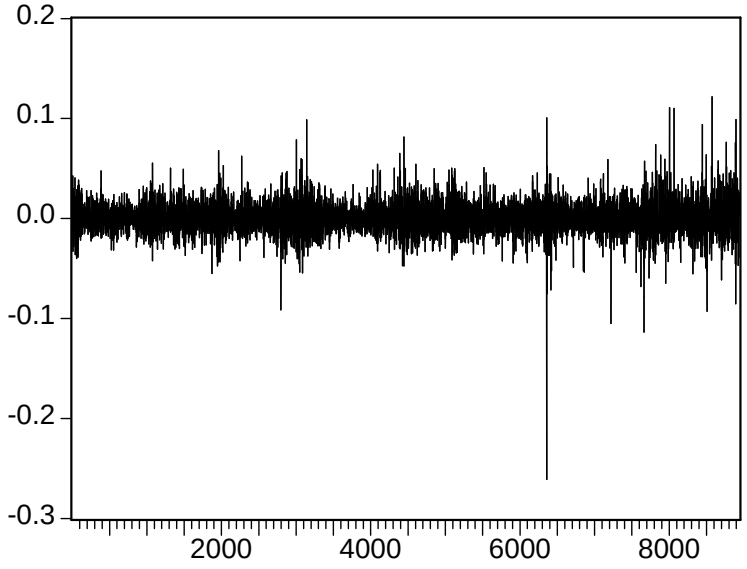


Figure 3: Daily IBM log returns 1962–1997.



Many financial theories and models assume that returns are normally distributed to facilitate theoretical derivations and applications. Deviations from normality can be measured by the (sample) **skewness**

$$S = \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \bar{y})^3}{\tilde{s}^3}$$

and (sample) **kurtosis**

$$U = \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \bar{y})^4}{\tilde{s}^4}.$$

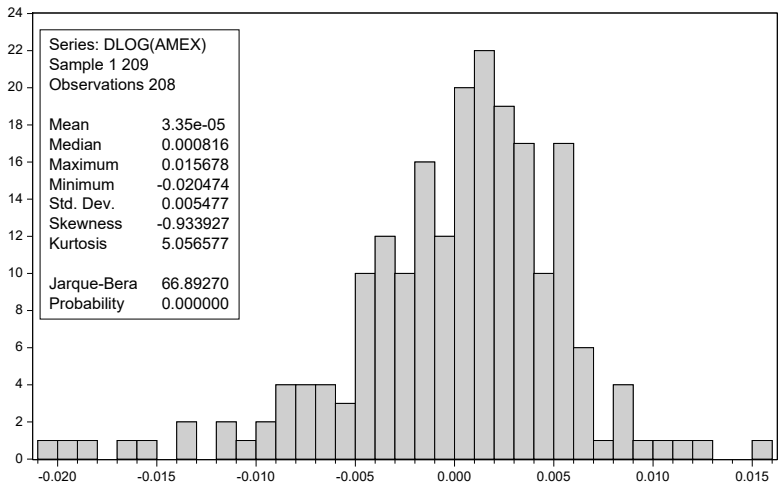
In large samples $S \overset{a}{\sim} \mathrm{N}(0, 6/n)$ and $U \overset{a}{\sim} \mathrm{N}(3, 24/n)$. Skewness is a measure of symmetry. If the skewness is negative, the left tail of the histogram is longer than the right tail. Simply speaking, the skewness is negative, if y_t has more negative than positive extreme values. The kurtosis[76] is a measure for the tail behavior. Financial returns typically have a kurtosis greater than 3. This is the case if the distribution is more strongly concentrated around the mean than the normal and assigns correspondingly higher probabilities to extreme values (positive or negative). Such distributions are **leptokurtic** and have so-called **fat** or **heavy tails**.

The **Jarque-Bera (JB)** test can be used to test for normality. It is based on the null hypothesis of a normal distribution and the test statistic takes skewness S and kurtosis U into account:

$$\mathrm{JB} = \frac{n}{6}\left[S^2 + \frac{1}{4}(U-3)^2\right] \qquad \mathrm{JB} \sim \chi^2_2.$$

Example 30: Figure 4 shows the histogram and descriptive statistics of log returns from the index of the American Stock Exchange (AMEX)[77]. The distribution is skewed and has fat tails. The JB-test rejects normality.

Figure 4: Histogram and descriptive statistics of AMEX log returns.



Exercise 17: Download a few financial time series from finance.yahoo.com or another website, or use another data source. Choose at least two different types of series (stock prices, indices, exchange rates or commodity prices) or at least two different frequencies (daily, weekly or monthly). Compute log and simple returns, obtain their descriptive statistics, and test for normality.

[76] $U-3$ is also called **excess kurtosis**.

[77] Source: SAS (1995) p.163; raw data: `http://ftp.sas.com/samples/A55217` (withdrawn by SAS); see files `amex.*`; daily data from 1993/8/2 to 1994/5/27.

2.1.2 Return distributions

Review 9:[78] A random variable X has a **lognormal distribution** if $Y{=}\ln X$ is normally distributed. Conversely, if $Y{\sim}\mathrm{N}(\mu,\sigma^2)$ then $X{=}\exp\{Y\}$ is lognormal. The density function of a lognormal random variable X is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\ln x-\mu)^2}{2\sigma^2}\right\} \quad x \geq 0,$$

where μ and σ^2 are mean and variance of $\ln X$, respectively. Mean and variance of X are given by

$$\mathrm{E}[X] = \mathrm{E}[\exp\{Y\}] = \exp\{\mu+0.5\sigma^2\} \qquad \mathrm{V}[X] = \exp\{2\mu+\sigma^2\}[\exp\{\sigma^2\}-1].$$

We now consider the log return in t and treat it as a random variable (denoted by Y_t; y_t is the corresponding sample value or realization). μ and σ^2 are mean and variance of the underlying population of log returns. Assuming that log returns are *normal* random variables with $Y_t{\sim}\mathrm{N}(\mu,\sigma^2)$ implies that $(1{+}R_t){=}\exp\{Y_t\}$, the simple, gross returns are *lognormal* random variables with

$$\mathrm{E}[R_t] = \exp\{\mu+0.5\sigma^2\}-1 \quad \text{and} \quad \mathrm{V}[R_t] = \exp\{2\mu+\sigma^2\}[\exp\{\sigma^2\}-1].$$

If the simple, gross return is lognormal $(1{+}R_t){\sim}\mathrm{LN}(1{+}m,v)$, mean and variance of the corresponding log return are given by

$$\mathrm{E}[Y_t] = \ln(1+m)-0.5\sigma_Y^2 \qquad \sigma_Y^2 = \mathrm{V}[Y_t] = \ln\left(1+\frac{v}{(1+m)^2}\right). \tag{41}$$

What are the implications for the corresponding prices? Normality of Y_t implies that prices given by $P_t{=}\exp\{Y_t\}P_{t-1}$ or $P_t{=}(1{+}R_t)P_{t-1}$ are lognormal (for given, non-random P_{t-1}). Thus, prices can never become negative if log returns are normal. Note that the computation of *expected* prices from expected returns differs from computing historical (ex-post) prices. Whereas $p_t{=}\exp\{y_t\}p_{t-1}$ holds for *observed* y_t and p_t, the *expected* price is given by $\mathrm{E}[p_t]{=}\exp\{\mu{+}0.5\sigma^2\}p_{t-1}$ if $y_t{\sim}\mathrm{N}(\mu,\sigma^2)$.

Example 31: The mean log return of the FTSE[79] is 0.00765, whereas the mean of simple returns is 0.009859. The standard deviation of log returns is 0.065256. Relation (40) holds pretty well since $\exp\{\bar{y}{+}0.5s^2\}{-}1{=}0.009827$.

We now compare the ex-post and ex-ante implications of $\bar{y}$ and $\bar{r}$. The value of the index was $p_0{=}105.4$ in January 1965. Given the average log return $\bar{y}$ and continuous compounding, the index at t is given by

$$p_t = p_0\exp\{\bar{y}t\} = 105.4\exp\{0.00765t\}.$$

This yields 1137.75 in December 1990 ($t{=}311$), which corresponds to the actual value of the FTSE. For comparison we use $\bar{r}$, discrete compounding and

$$p_t = p_0(1+\bar{r})^t = 105.4(1+0.009859)^t$$

[78] For details see Hastings and Peacock (1975), p.84.

[79] The Financial Times All Share Index (FTSE). Source: Mills (1993) p.225; see files `ftse.*`; monthly data from January 1965 to December 1990.

to compute p_{311}=2228.06. Thus, at hindsight only the average log return corresponds exactly to observed prices. We would need to use $\tilde{r}=(p_t/p_0)^{1/t}-1$ to get the correct ex-post implications based on discrete returns. However, from an *ex-ante* perspective, $\bar{y}$ and $\bar{r}$ imply roughly the same *expected* prices if log returns are assumed to be normal:

$$\mathrm{E}[p_t] = p_0 \exp\{t(\bar{y} + 0.5s^2)\} = p_0 \exp\{0.009779t\} \approx p_0(1 + 0.009859)^t.$$

Another attractive feature of normal log returns is their behavior under temporal aggregation. If single-period log returns are normally distributed $Y_t \sim \mathrm{N}(\mu, \sigma^2)$, the multi-period log returns are also normal with $Y_t(h) \sim \mathrm{N}(h\mu, h\sigma^2)$. This property is called **stability** (under addition). It does not hold for simple returns.

Many financial theories and models assume that *simple* returns are normal. There are several conceptual difficulties associated with this assumption. First, simple returns have a lower bound of -1, whereas the normal distribution extends to $-\infty$. Second, multi-period returns are not normal even if single-period (simple) returns are normal. Third, a normal distribution for simple returns implies a normal distribution for prices, since $P_t=(1+R_t)P_{t-1}$. Thus, a non zero probability may be assigned to negative prices which is generally not acceptable. These drawbacks can be overcome by using log returns rather than simple returns. However, empirical properties usually indicate strong deviations from normality for both simple and log returns.

As a consequence of the empirical evidence against the normality of returns various alternatives have been suggested. The class of **stable distributions** has the desirable properties of fat tails and stability under addition. One example is the **Cauchy distribution** with density

$$f(y) = \frac{1}{\pi} \frac{b}{b^2 + (y-a)^2}, \quad -\infty < y < \infty.$$

However, the variance of stable distributions does not exist, which causes difficulties for almost all financial theories and applications.[80] The **Student t-distribution** also has fat tails if its only parameter – the degrees of freedom – is set to a small value. The t-distribution is a frequently applied alternative to the normal distribution.[81]

The **mixture of normal distributions** approach assumes that returns are generated by two or more normal distributions, each with a different variance. For example, a mixture of two normal distributions[82] is given by

$$y_t \sim (1-x)\mathrm{N}(\mu, \sigma_1^2) + x\mathrm{N}(\mu, \sigma_2^2),$$

where x is a Bernoulli random variable with $\mathrm{P}[x=1]=\alpha$. This accounts for the observation that return volatility is not constant over time (see example 29). The normal mixture model is based on the notion that financial markets are processing information. The amount of information can be approximated by the variance of returns. As it turns out, the mixture also captures non-normality. For instance, a mixture of a low variance distribution (with high probability α) and a large variance distribution (with low probability α) results

[80] For details see Fielitz and Rozelle (1983).
[81] For details see Blattberg and Gonedes (1974) or Kon (1984).
[82] An example of simulated returns based on a mixture of three normal distributions can be found in the file `mixture of normal distributions.xls`.

in a non-normal distribution with fat tails. Thus, if returns are assumed to be conditionally normal given a certain amount of information, the implied unconditional distribution is non-normal. Kon (1984) has found that between two and four normal distributions are necessary and provide a better fit than t-distributions with degrees of freedom ranging from 3.1 to 5.5.

2.1.3 Abnormal returns and event studies[83]

Financial returns can be viewed as the result of processing information. The purpose of **event studies** is to test the statistical significance of events (mainly announcements) on the returns of one or several assets. For example, a frequently analyzed event is the announcement of a (planned) merger or takeover. This may be a signal about the value of the firm which may be reflected in its stock price. Comparing returns before and after the information becomes publicly available can be used to draw conclusions about the relevance of this information.

Event studies typically consist of analyzing the effects of a particular type of information or event across a large number of companies. This requires an alignment of individual security returns relative to an event date (denoted by τ=0). In other words, a new time index τ replaces the calendar time t such that τ=0 corresponds to the event date in each case. The **event window** covers a certain time period around τ=0 and is used to make comparisons with pre-event (or post-event) returns.

The effects of the event have to be isolated from effects that would have occurred irrespective of the event. For this purpose it is necessary to define normal and abnormal returns. 'Normal' refers to the fact that these returns would 'normally' be observed, either because of other reasons than the event under study or if the event has no relevance. Normal returns can be defined either on the basis of average historical returns or a regression model. These estimates are obtained from the **estimation window**, which is a time period preceding the event window. They serve as the expected or predicted returns during the event window. **Abnormal returns** are the difference between normal and observed returns during the event window.

Suppose that the estimation window ranges from τ=τ_0+1 to τ_1 (n_1 observations), and the event window ranges from τ=τ_1+1 to τ_2 (n_2 observations) and includes the event date τ=0. We will consider estimating abnormal returns for company i based on the market model

$$y_\tau^i = \alpha_i + \beta_i y_\tau^{im} + \epsilon_\tau^i \qquad \tau=\tau_0+1,\ldots,\tau_1,$$

where the market return y_τ^{im} has a firm-specific superscript to indicate that the market returns have been aligned to match the firm's event date. Given OLS estimates a_i and b_i and observations for the market returns in the event window, we can compute n_2 abnormal returns

$$e_\tau^i = y_\tau^i - a_i - b_i y_\tau^{im} \qquad \tau=\tau_1+1,\ldots,\tau_2.$$

We define the $n_1\times 2$ matrix $\boldsymbol{X}$ for firm i using n_1 observations from the estimation window. Each of its rows is given by $(1\ y_\tau^{im})$. A corresponding $n_2\times 2$ matrix $\boldsymbol{X}_0$ is defined for the event window and the subscript 0 refers to the index set $(\tau_1+1,\ldots,\tau_2)$. Given the OLS estimates $\boldsymbol{b}_i=(a_i\ \ b_i)'$ for the parameters of the market model, the vector of abnormal returns for firm i is defined as

$$\boldsymbol{e}_0^i = \boldsymbol{y}_0^i - \boldsymbol{X}_0\boldsymbol{b}_i,$$

[83] Most of this section is based on Chapter 4 in Campbell et al. (1997) where further details and references on the event study methodology can be found. Other useful sources of information are the Event Study Webpage http://web.mit.edu/doncram/www/eventstudy.html by Don Cram and the lecture notes by Frank de Jong.

where $\boldsymbol{y}_0^i$ is the vector of observed returns. From section 1.2.6 we know that $\mathrm{E}[\boldsymbol{e}_0^i]=0$ (since $\boldsymbol{X}_0\boldsymbol{b}_i$ is an unbiased estimate of $\boldsymbol{y}_0^i$), and its variance is given by

$$\mathrm{V}[\boldsymbol{e}_0^i] = \boldsymbol{V}_i = \sigma_i^2\boldsymbol{I} + \sigma_i^2\boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0'.$$

$\boldsymbol{I}$ is the $n_2\times n_2$ identity matrix and σ_i^2 is the variance of disturbances ϵ^i in the market model. The estimated variance $\hat{\boldsymbol{V}}_i$ is obtained by using the error variance from the estimation period

$$s_i^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n_1 - 2} \qquad \boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$$

in place of σ_i^2.

Event studies are usually based on the null hypothesis that the event under consideration has *no impact* on (abnormal) returns. Statistical tests can be based on the assumption that abnormal returns are normally distributed and the properties just derived: $\boldsymbol{e}_0^i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_i)$. However, the information collected must be aggregated to be able to make statements and draw conclusions about the event (rather than individual cases or observations). It is not always known *when* an event will have an effect and *how long* it will last. Therefore abnormal returns are *cumulated* across time in the event window. In addition, the implications of the event are expressed in terms of averages across several firms which may potentially be affected by the event. We start by considering the temporal aggregation.

If the event window consists of more than one observation we can define the cumulative abnormal return for firm i by summing all abnormal returns from $\tau_1{+}1$ to τ

$$c_\tau^i = \boldsymbol{\iota}_\tau'\boldsymbol{e}_0^i,$$

where the $n_2\times 1$ vector $\boldsymbol{\iota}_\tau$ has ones from row one to row τ, and zeros elsewhere. The estimated variance of c_τ^i is given by

$$\boldsymbol{\iota}_\tau'\hat{\boldsymbol{V}}_i\boldsymbol{\iota}_\tau.$$

This variance is firm specific. To simplify the notation we define the variance in terms of

$$\boldsymbol{H} = \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0' \qquad \boldsymbol{V}_i = \sigma_i^2\boldsymbol{I} + \sigma_i^2\boldsymbol{H}.$$

Note that $\boldsymbol{H}$ is firm-specific since $\boldsymbol{X}$ is different for each firm (the market returns contained in $\boldsymbol{X}$ have to be aligned with the event time of firm i). The standard error of cumulative abnormal returns across n_2 periods for firm i is given by

$$\mathrm{se}[c_\tau^i] = \sqrt{n_2 s_i^2 + s_i^2(\boldsymbol{\iota}_\tau'\boldsymbol{H}\boldsymbol{\iota}_\tau)}.$$

The null hypothesis of zero abnormal returns can be tested using the standardized test statistic

$$t_c^i = \frac{c_\tau^i}{\mathrm{se}[c_\tau^i]}.$$

Under the assumption that abnormal returns are jointly normal and serially uncorrelated the test statistic t_c has a t-distribution with df=n_1-2.

Event studies are frequently based on analyzing many firms which are all subject to the same kind of event (usually at different points in calendar time). Under the assumption that abnormal returns for individual firms are uncorrelated (i.e. the event windows do not overlap) tests can be based on averaging cumulative abnormal returns across m firms and the test statistic

$$t_1 = \frac{\bar{c}}{\text{se}[\bar{c}]} \overset{a}{\sim} \text{N}(0,1),$$

where

$$\bar{c} = \frac{1}{m}\sum_{i=1}^{m} c_\tau^i \qquad \text{se}[\bar{c}] = \sqrt{\frac{1}{m^2}\sum_{i=1}^{m} \text{se}[c_\tau^i]^2}.$$

Alternatively, the test statistic t_c^i can be averaged to obtain the test statistic

$$t_2 = \sqrt{\frac{m(n_1-4)}{n_1-2}}\left(\frac{1}{m}\sum_{i=1}^{m} t_c^i\right) \overset{a}{\sim} \text{N}(0,1).$$

Example 32: We consider the case of two Austrian mining companies Radex and Veitscher who where the subject of some rumors about a possible takeover. The first newspaper reports about a possible 'cooperation' appeared on March 8, 1991. Similar reports appeared throughout March 29. On April 16 it was officially announced that Radex will buy a 51% share of Veitscher. The purpose of the analysis is to test for abnormal returns associated with this event. Details can be found in the file `event.xls`.

The estimation window consists of the three year period from January 25, 1988 to January 24, 1991. We use daily log returns for the two companies and the ATX to estimate the market model. The event window consists of 51 days (January 25 to April 10). The cumulative abnormal returns start to increase strongly about 14 days before March 8 and reach their peak on March 7. After that day cumulative abnormal returns are slightly decreasing. Based on 51 days of the event period we find c_τ^1=0.25 and c_τ^1=0.17 for Radex and Veitscher, respectively. The associated t-statistics are t_c^1=3.29 and t_c^2=2.54 which are both highly significant. Tests based on an aggregation across the two companies are not appropriate in this example since they share the same event window.

Exercise 18: Use the data from example 32 to test the significance of cumulative abnormal returns for event windows ranging from January 25 to March 7 and March 15, respectively. You may also use other event windows that allow for interesting conclusions.

2.1.4 Autocorrelation analysis of financial returns

The methods of time series analysis are used to investigate the dynamic properties of a single realization y_t, in order to draw conclusions about the nature of the underlying stochastic process Y_t, and to estimate its parameters. Before we define specific time series models, and consider their estimation and forecasts, we briefly analyze the dynamic properties of some financial time series.

Autocorrelation analysis is a standard tool for that purpose. The sample autocovariance[84] and the sample autocorrelation

$$c_\ell = \frac{1}{n}\sum_{t=\ell+1}^{n}(y_t - \bar{y})(y_{t-\ell} - \bar{y})$$

$$r_\ell = \frac{c_\ell}{c_0} = \frac{c_\ell}{s^2}$$

can be used to investigate *linear* temporal dependencies in an observed series y_t. c_ℓ and r_ℓ are sample estimates of γ_ℓ (13) and ρ_ℓ (14). If the underlying process Y_t is i.i.d., the sampling distribution of r_ℓ is $r_\ell \simeq \mathrm{N}(-1/n, 1/n)$. This can be used to test individual autocorrelations for significance (e.g. using the 95% confidence interval[85] $-1/n \pm 1.96/\sqrt{n}$.).

Rather than testing individual autocorrelations the **Ljung-Box** statistic can be used to test jointly that *all* autocorrelations up to lag p are zero:

$$Q_p = n(n+2)\sum_{\ell=1}^{p}\frac{r_\ell^2}{n-\ell}.$$

Under the null hypothesis of zero autocorrelation in the population ($\rho_1 = \cdots = \rho_p = 0$): $Q_p \sim \chi^2_p$.

> **Example 33:** The autocorrelations of IBM log returns in Figure 5 are negligibly small (except for lags 6 and 9). The p-values of the Q-statistic (Prob and Q-Stat in Figure 5) indicate that the log returns are uncorrelated. The situation is slightly different for FTSE log returns. These autocorrelations are rather small but the correlations at lags one, two and five are slightly outside the 95%-interval. The p-values of the Q-statistic are between 0.01 and 0.05. Depending on the significance level we would either reject or accept the null hypothesis of no correlation. We conclude that the FTSE log returns are weakly correlated.

Assuming that returns are independent is stronger than assuming uncorrelated returns[86]. However, testing for independence is not straightforward because it usually requires to specify a particular type of dependence. Given that the variance of financial returns is typically not constant over time, a simple test for independence is based on the autocorrelations of squared or absolute returns.

> **Example 34:** Figure 6 shows the autocorrelations of squared and absolute log returns of IBM. There are many significant autocorrelations even at long lags. Thus we

[84] This is a biased estimate of the autocovariance which has the advantage of yielding a positive semi-definite autocovariance matrix. The unbiased estimated is obtained if the sum is divided by $n-1$.

[85] Usually the mean $-1/n$ is ignored and $\pm 1.96/\sqrt{n}$ is used as the 95% confidence interval.

[86] Independence and uncorrelatedness are only equivalent if returns are normally distributed.

Figure 5: Autocorrelations of IBM (left panel) and FTSE (right panel) log returns.

Included observations: 368

Autocorrelation		AC	Q-Stat	Prob
	1	0.023	0.2017	0.653
	2	0.006	0.2149	0.898
	3	-0.036	0.6889	0.876
	4	-0.055	1.8366	0.766
	5	-0.027	2.1125	0.833
	6	0.139	9.4080	0.152
	7	0.070	11.267	0.127
	8	0.041	11.913	0.155
	9	-0.090	15.011	0.091
	10	0.002	15.014	0.132

Sample: 1965:01 1990:12
Included observations: 311

Autocorrelation		AC	Q-Stat	Prob
	1	0.113	4.0342	0.045
	2	-0.103	7.3859	0.025
	3	0.093	10.118	0.018
	4	0.061	11.304	0.023
	5	-0.102	14.589	0.012
	6	-0.036	15.001	0.020
	7	0.043	15.599	0.029
	8	-0.047	16.312	0.038
	9	0.076	18.152	0.033
	10	0.022	18.303	0.050

Figure 6: Autocorrelations of squared (left panel) and absolute (right panel) log returns of IBM.

Autocorrelation		AC	Q-Stat	Prob
	1	0.303	34.004	0.000
	2	0.188	47.096	0.000
	3	0.321	85.576	0.000
	4	0.306	120.61	0.000
	5	0.040	121.20	0.000
	6	0.158	130.55	0.000
	7	0.111	135.19	0.000
	8	0.121	140.77	0.000
	9	0.298	174.39	0.000
	10	0.265	201.12	0.000
	11	0.146	209.22	0.000
	12	0.242	231.62	0.000
	13	0.372	284.77	0.000
	14	0.067	286.48	0.000
	15	0.111	291.25	0.000
	16	0.164	301.71	0.000
	17	0.200	317.29	0.000
	18	0.065	318.96	0.000
	19	0.251	343.51	0.000
	20	0.196	358.54	0.000

Autocorrelation		AC	Q-Stat	Prob
	1	0.405	60.964	0.000
	2	0.294	93.200	0.000
	3	0.398	152.34	0.000
	4	0.340	195.55	0.000
	5	0.143	203.22	0.000
	6	0.258	228.33	0.000
	7	0.235	249.16	0.000
	8	0.212	266.10	0.000
	9	0.327	306.66	0.000
	10	0.334	349.13	0.000
	11	0.252	373.39	0.000
	12	0.270	401.18	0.000
	13	0.323	441.12	0.000
	14	0.178	453.34	0.000
	15	0.225	472.81	0.000
	16	0.283	503.71	0.000
	17	0.314	541.92	0.000
	18	0.211	559.26	0.000
	19	0.323	600.00	0.000
	20	0.297	634.61	0.000

conclude that the IBM log returns are uncorrelated but not independent. At the same time the significant autocorrelations among squared and absolute returns point at dependencies in (the variance of) returns.

In section 2.1 we have presented examples of volatility clustering. If the sign of returns is ignored (either by considering squared or absolute returns), the correlation within clusters is high. If the variance has moved to a high level it tends to stays there; if it is low it tends to stay low. This explains that autocorrelations of absolute and squared returns are positive for many lags.

Significant autocorrelation in squared or absolute returns is evidence for heteroscedasticity. In this case the standard errors $1/\sqrt{n}$ are not appropriate to test the regular autocorrelations r_ℓ for significance. Corrected confidence intervals can be based on the modified

variance of the autocorrelation coefficient at lag ℓ:

$$\frac{1}{n}\left(1+\frac{c_{y^2}(\ell)}{s^4}\right),$$

where $c_{y^2}(\ell)$ is the autocovariance of y_t^2 and s^4 is the squared variance of y_t. The resulting standard errors are larger than $1/\sqrt{n}$ if squared returns are positively autocorrelated which is typical for financial returns. This leads to wider confidence intervals and to more conservative conclusions about the significance of autocorrelations. If the modified standard errors are used for testing log returns of the FTSE no autocorrelations in Figure 5 are significant (α=0.05).

Exercise 19: Use the log returns defined in exercise 17. Estimate and test autocorrelations of regular, squared and absolute returns.

2.1.5 Stochastic process terminology

We briefly define some frequently used stochastic processes:

A **white-noise process** ϵ_t is a stationary and uncorrelated sequence of random numbers. It may have mean zero (which is mainly assumed for convenience), but this is not essential. The key requirement is that the series is serially uncorrelated; i.e. $\gamma_\ell{=}\rho_\ell{=}0$ $(\forall\ell{\neq}0)$. If ϵ_t is normally distributed and white-noise it is independent (Gaussian white-noise). If ϵ_t is white-noise with constant mean and constant variance with a fixed distribution it is an i.i.d. sequence[87] (also called independent white-noise).

A **martingale difference sequence (m.d.s.)** Y_t is defined with respect to the information I_t available at t. This could include any variables but typically only includes Y_t: $I_t{=}\{Y_t, Y_{t-1}, \ldots\}$. $\{Y_t\}_{t=1}^{\infty}$ is a m.d.s. (with respect to I_{t-1}) if $\mathrm{E}[Y_t|Y_{t-1}, Y_{t-2}, \ldots]{=}0$ (which implies $\mathrm{E}[Y_t]{=}0$). Since white-noise restricts the conditional expectation to linear functions, a m.d.s. implies stronger forms of independence than white-noise.

A **random walk** with drift δ is defined as

$$Y_t = Y_{t-1} + \delta + \epsilon_t \qquad \epsilon_t \ldots \text{white noise.}$$

In other words, the time increments of a random walk are white-noise.[88]

If Y_t is an element of I_t and $\mathrm{E}[Y_t|I_{t-1}]{=}Y_{t-1}$ then Y_t is a **martingale** (or **martingale sequence**) with respect to I_{t-1}. A random walk is an example of a martingale.

A **mean reverting** process is a stationary process with non-zero autocorrelations. It is expected to revert to its (unconditional) mean[89] μ from below (above) if $Y_t{<}\mu$ $(Y_t{>}\mu)$. Since the process is stationary, it reverts to the mean relatively fast compared to a non-stationary process without drift (see section 2.3).

An autocorrelated process[90] can be written as

$$Y_t = \hat{Y}_t + \epsilon_t \qquad \hat{Y}_t = \mathrm{E}[Y_t|Y_{t-1}, Y_{t-2}, \ldots,] \qquad \sigma_Y^2 \neq \sigma_\epsilon^2,$$

where $\hat{Y}_t$ denotes the conditional mean. If the variance of σ_ϵ^2 is not constant over time, the conditional variance is defined in a similar way

$$\mathrm{E}[(Y_t - \hat{Y}_t)^2|Y_{t-1}, Y_{t-2}, \ldots,] = \mathrm{V}[\epsilon_t|Y_{t-1}, Y_{t-2}, \ldots,] = \sigma_t^2.$$

In this case ϵ_t is uncorrelated (white noise) but not i.i.d. σ_t is the conditional variance of ϵ_t (i.e. the conditional expectation of ϵ_t^2).

[87] We will use the stronger i.i.d. property for a white-noise with constant variance (and distribution). White-noise only refers to zero autocorrelation and need not have constant variance.

[88] Campbell et al. (p.31 1997) distinguish three types of random walks depending on the nature of ϵ_t: i.i.d. increments, independent (but not identically distributed) increments and uncorrelated increments.

[89] Strictly speaking this definition also applies to white-noise, but the term mean reversion is mainly used in the context of autocorrelated stationary processes.

[90] An uncorrelated process would be written as $Y_t{=}\mu{+}\epsilon_t$.

2.2 ARMA models

We now introduce some important linear models for the conditional mean. An **autoregressive moving-average (ARMA)** process is a linear stochastic process which is completely characterized by its autocovariances γ_ℓ (or autocorrelations ρ_ℓ). Thus, various ARMA models can be defined and distinguished by their (estimated) autocorrelations. In practice the (estimated) autocorrelations r_ℓ from an observed time series are compared to the known theoretical autocorrelations of ARMA processes. Based on this comparison a time series model is specified. This is also called the **identification step** in the model building process. After estimating its parameters diagnostic checks are used to confirm that a suitable model has been chosen (i.e. the underlying stochastic process conforms to the estimated model). ARMA models are only appropriate for *stationary* time series.

2.2.1 AR models

The first order **autoregressive process** AR(1)

$$Y_t = \nu + \phi_1 Y_{t-1} + \epsilon_t \qquad |\phi_1| < 1$$

has exponentially decaying autocorrelations $\rho_\ell{=}\phi_1^\ell$. ϵ_t is a white-noise process with mean zero and constant variance σ_ϵ^2. The condition $|\phi_1|{<}1$ is necessary and sufficient for the AR(1) process to be weakly stationary.

The unconditional mean of an AR(1) process is given by

$$\mathrm{E}[Y_t] = \mu = \frac{\nu}{1-\phi_1}.$$

An equivalent formulation of the AR(1) process is given by

$$Y_t - Y_{t-1} = \Delta Y_t = (1-\phi_1)(\mu - Y_{t-1}) + \epsilon_t.$$

Thus, deviations from the unconditional mean imply expected changes in Y_t which depend on the extent of the deviation and the degree of mean reversion $1{-}\phi_1$.

The unconditional variance of an AR(1) process is derived as follows:

$$\mathrm{V}[Y_t] = \sigma_Y^2 = \mathrm{V}[\nu + \phi_1 Y_{t-1} + \epsilon_t] = \phi_1^2 \mathrm{V}[Y_{t-1}] + \mathrm{V}[\epsilon_t].$$

If Y_t is stationary $\mathrm{V}[Y_t]{=}\mathrm{V}[Y_{t-1}]$ and

$$\mathrm{V}[Y_t] = \frac{\sigma_\epsilon^2}{1-\phi_1^2},$$

which is bounded and non-negative only if $|\phi_1|{<}1$.

The exponential decay of the autocorrelations of an AR(1) process can be derived by multiplying the model equation by Y_{t-1} and taking the expected value (assuming $\nu{=}0$ for the sake of simplicity):

$$\gamma_1 = \mathrm{E}[Y_t Y_{t-1}] = \mathrm{E}[\phi_1 Y_{t-1} Y_{t-1}] + \mathrm{E}[\epsilon_t Y_{t-1}] = \phi_1 \mathrm{E}[Y_{t-1}^2] = \phi_1 \gamma_0 = \phi_1 \sigma_Y^2. \tag{42}$$

Therefore $\rho_1{=}\gamma_1/\gamma_0{=}\phi_1$ since $\gamma_0{=}\sigma_Y^2$. Repeating this procedure for Y_{t-2} gives

$$\gamma_2 = \mathrm{E}[Y_t Y_{t-2}] = \phi_1 \mathrm{E}[Y_{t-1} Y_{t-2}] = \phi_1^2 \sigma_Y^2,$$

so that $\rho_2{=}\phi_1^2$, and in general $\rho_\ell{=}\phi_1^\ell$.
A generalization of the AR(1) process is the AR(p) process

$$Y_t = \nu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \qquad \mathrm{E}[Y_t] = \mu = \frac{\nu}{(1 - \phi_1 - \cdots - \phi_p)}.$$

It is convenient to make use of the **backshift operator** B with the property

$$B^\ell Y_t = Y_{t-\ell}.$$

Using this operator an AR(p) process can be formulated as follows:

$$Y_t = \nu + (\phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p) Y_t + \epsilon_t$$

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Y_t = \nu + \epsilon_t$$

$$\phi(B) Y_t = \nu + \epsilon_t \qquad \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p.$$

$\phi(B)$ is a polynomial of order p in B. Using $\phi(B)$ we can rewrite the AR process as

$$Y_t = \frac{\nu}{\phi(B)} + \frac{\epsilon_t}{\phi(B)} = \mu + \frac{\epsilon_t}{\phi(B)}.$$

For the AR(1) model $\phi(B) = (1 - \phi_1 B)$ we have

$$\frac{1}{(1 - \phi_1 B)} = (1 + \phi_1 B + \phi_1^2 B^2 + \cdots),$$

and Y_t can be written as

$$Y_t = \mu + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \cdots.$$

The resulting process is an infinite order moving-average (see below).
Since ϵ_t is stationary (by definition) Y_t will only be stationary if the weighted sum of disturbances converges, i.e. if $|\phi_1|{<}1$. In general, the stationarity of an AR(p) model depends on the properties of the polynomial $\phi(B)$.

Review 10: Given a polynomial of degree n

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

the constants $z_1,\ldots,z_n$ (real or complex) are called **zeros** of $f(x)$ or **roots** of $f(x){=}0$ such that

$$f(x) = a_n (x - z_1) \cdots (x - z_n).$$

The stationarity of an AR process depends on the roots of the AR polynomial $\phi(B)$ which can be factored as follows:

$$\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p) = \prod_{i=1}^{p}(1 - w_i B),$$

where w_i are the inverted roots of the polynomial, which may be complex valued. The AR model is stationary if all inverted roots are less than one in absolute value (or, all inverted roots are *inside* the unit circle).

Various autocorrelation patterns of an AR(p) process are possible. For instance, the autocorrelations of an AR(2) model show sinusoidal decay if its inverted roots are complex. In this case the underlying series has stochastic cycles.[91]

AR models imply non-zero autocorrelations for many lags. Nevertheless it may be sufficient to use one or only a few lags of Y_t to define $\hat{Y}_t$. The number of necessary lags p can be determined on the basis of **partial autocorrelations** $\phi_{\ell\ell}$. $\phi_{\ell\ell}$ is the ℓ-th coefficient of an AR(ℓ) model. It measures the effect of $Y_{t-\ell}$ on Y_t under the condition that the effects from *all other* lags are held constant.

Partial autocorrelations can be determined from the solution of the **Yule-Walker equations** of an AR(p) process:

$$\gamma_\ell = \phi_1\gamma_{\ell-1} + \phi_2\gamma_{\ell-2} + \cdots + \phi_p\gamma_{\ell-p} \qquad \ell = 1, \ldots, p.$$

For example, the Yule-Walker equations of an AR(2) process are given by

$$\begin{aligned} \ell = 1: \ & \gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1 \\ \ell = 2: \ & \gamma_2 = \phi_1\gamma_1 + \phi_2\gamma_0. \end{aligned}$$

In this case the solutions are given by (see Box and Jenkins, 1976, p.83)

$$\phi_1 = \frac{\gamma_1(1-\gamma_2)}{1-\gamma_1^2} \qquad \phi_2 = \phi_{22} = \frac{\gamma_2 - \gamma_1^2}{1-\gamma_1^2}.$$

AR coefficients (and thereby, partial autocorrelations) can be obtained by solving the Yule-Walker equations recursively for AR models of increasing order. The recursions (in terms of autocorrelations ρ_ℓ) are given by (see Box and Jenkins, 1976, p.83)

$$\phi_{p+1,p+1} = \frac{\rho_{p+1} - \sum_{\ell=1}^{p}\phi_{p,\ell}\rho_{p+1-\ell}}{1 - \sum_{\ell=1}^{p}\phi_{p,\ell}\rho_\ell}$$

$$\phi_{p+1,\ell} = \phi_{p,\ell} - \phi_{p+1,p+1}\phi_{p,p-\ell+1} \qquad \ell = 1, \ldots, p.$$

[91] An example of such a process is obtained if ϕ_1=1.5 and ϕ_2=−0.7. The file `arma.xls` can be used to obtain simulated paths of ARMA models.

If the theoretical autocovariances are replaced by estimated autocovariances or autocorrelations, (preliminary) AR parameters can be estimated from the solution of the equation system.

The partial autocorrelations of an AR(p) process *cut off* at p: $\phi_{\ell\ell}$=0 for ℓ>p. This is the basis for identifying AR(p) models empirically. Significant partial autocorrelations up to lag p and decaying autocorrelations (theoretically) suggest to estimate an AR(p) model. In practice this identification may be difficult.

2.2.2 MA models

The **moving-average process** of order q denoted by MA(q) is defined as follows:

$$Y_t = \mu + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} + \epsilon_t \qquad \epsilon_t \ldots \text{white-noise.}$$

Its unconditional mean and variance are given by

$$\mathrm{E}[Y_t] = \mu \qquad \mathrm{V}[Y_t] = (1 + \theta_1^2 + \cdots \theta_q^2)\sigma_\epsilon^2.$$

The autocovariance at lag 1 is given by (assuming ν=0 for the sake of simplicity):

$$\begin{aligned}\gamma_1 = \mathrm{E}[Y_tY_{t-1}] &= \mathrm{E}[(\theta_1\epsilon_{t-1} + \epsilon_t)(\theta_1\epsilon_{t-2} + \epsilon_{t-1})] \\ &= \mathrm{E}[\theta_1^2\epsilon_{t-1}\epsilon_{t-2} + \theta_1\epsilon_{t-1}^2 + \theta_1\epsilon_t\epsilon_{t-2} + \epsilon_t\epsilon_{t-1}] \\ &= \theta_1\mathrm{E}[\epsilon_{t-1}^2] = \theta_1\sigma_\epsilon^2,\end{aligned}$$

since for a white-noise process $\mathrm{E}[\epsilon_t\epsilon_s]$=0 ($\forall t{\neq}s$). In a similar manner it can be shown that γ_ℓ=0 ($\forall\ \ell$>1). For the general MA(q) process the autocorrelation function is given by

$$\rho_\ell = \begin{cases} \dfrac{\sum_{i=0}^{q-\ell}\theta_i\theta_{i+\ell}}{1+\sum_{i=1}^{q}\theta_i^2} & \ell = 1,\ldots,q \\ 0 & \ell > q. \end{cases} \tag{43}$$

Thus a MA(q) is characterized by autocorrelations that *cut off* at lag q.

The choice of the term 'moving-average' can be derived from the correspondence between MA(q) and AR(∞) models. We consider a MA(1) model and rewrite it as an AR model:

$$Y_t = \mu + (1 + \theta_1 B)\epsilon_t \qquad \frac{Y_t}{(1+\theta_1 B)} = \frac{\mu}{(1+\theta_1 B)} + \epsilon_t$$

$$(1 - \theta_1 B + \theta_1^2 B^2 - \cdots)Y_t = \nu + \epsilon_t.$$

This shows that the corresponding AR coefficients imply a weighted average of lagged Y_t (with decreasing weights, provided $|\theta_1|$<1). This transformation is possible if the MA model is **invertible**, which is the case if all inverted roots of the MA polynomial are less than one.

The relation between AR and MA models is the foundation for identifying MA(q) models empirically. Significant autocorrelations up to lag q and decaying partial autocorrelations suggest to estimate an MA(q) model.

Table 1: Theoretical patterns of (partial) autocorrelations.

	autocorrelations	partial autocorrelations
AR(1) $\phi_1>0$	exponential decay; $\rho_\ell=\phi_1^\ell$	cut off at lag 1; $\phi_{11}>0$
AR(1) $\phi_1<0$	oscillating decay; $\rho_\ell=\phi_1^\ell$	cut off at lag 1; $\phi_{11}<0$
AR(p)	exponential decay (oscillating)	cut off at lag p
MA(1) $\theta_1>0$	cut off at lag 1; $\rho_1>0$	oscillating decay; $\phi_{11}>0$
MA(1) $\theta_1<0$	cut off at lag 1; $\rho_1<0$	exponential decay; $\phi_{11}<0$
MA(q)	cut off at lag q	exponential decay (oscillating)
ARMA(p,q)	decay starting at lag q	decay starting at lag p

2.2.3 ARMA models

ARMA(p,q) models combine AR and MA models:

$$Y_t = \nu + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

$$\phi(B) Y_t = \nu + \theta(B)\epsilon_t.$$

Table 1 summarizes the theoretical patterns of autocorrelations and partial autocorrelations of ARMA(p,q) models. These can be used as a rough guideline for model identification. In practice, the identification of AR and MA models can be difficult because estimated (partial) autocorrelations are subject to estimation errors and the assignment to theoretical patterns may be ambiguous. However, several possible models may be temporarily selected. For instance, a low-order ARMA model can be specified and extended step-by-step. Subsequent model estimation is carried out for all selected models. After estimation and diagnostic checking a final choice among the models can be made.

Low-order ARMA models can substitute high-order AR or MA models with a few parameters only. Provided that all inverted roots of $\phi(B)$ are less than one in absolute terms, an ARMA(p,q) model can be formulated as follows:

$$Y_t = \frac{\nu}{\phi(B)} + \frac{\theta(B)}{\phi(B)}\epsilon_t \qquad Y_t = \mu + \psi(B)\epsilon_t.$$

This is equivalent to a MA(∞) model:

$$Y_t = \mu + (1 + \psi_1 B + \psi_2 B^2 + \cdots)\epsilon_t = \mu + \sum_{\ell=0}^{\infty} \psi_\ell \epsilon_{t-\ell} \qquad (\psi_0 = 1).$$

This representation not only holds for ARMA models. According to the **Wold decomposition** any covariance stationary process (with mean zero) can be written as

$$Y_t = U_t + V_t = \sum_{\ell=0}^{\infty} \psi_\ell \epsilon_{t-\ell} + V_t,$$

where U_t and V_t are uncorrelated, ψ_0=1 and $\sum_{\ell=0}^{\infty} \psi_\ell^2 < \infty$, ϵ_t is white-noise defined as

$$\epsilon_t = Y_t - \mathrm{E}[Y_t | Y_{t-1}, Y_{t-2}, \ldots],$$

$\mathrm{E}[\epsilon_t V_t]$=0, and V_t can be predicted from V_{t-1},V_{t-2},... with zero prediction variance.

2.2.4 Estimating ARMA models

ARMA models can be viewed as a special case of a linear regression model with lagged dependent variables. Estimating ARMA models leads to biased estimates in small sample since assumption **AX** $\mathrm{E}[\epsilon|\boldsymbol{X}]=\mathbf{0}$ is violated. This can be shown using the AR(1) model $Y_t=\phi_1 Y_{t-1}+\epsilon_t$. Assuming that $\overline{\mathbf{AX}}$ holds (i.e. ϵ_t is orthogonal to (the stochastic regressor) Y_{t-1}; $\mathrm{E}[Y_{t-1}\epsilon_t]=0$) we have

$$\mathrm{E}[Y_t\epsilon_t] = \mathrm{E}[(\phi_1 Y_{t-1} + \epsilon_t)\epsilon_t] = \mathrm{E}[\epsilon_t^2].$$

This implies that the regressor Y_{t-1} is not orthogonal to the disturbance ϵ_{t-1} (i.e. there is correlation between regressors and disturbances *across* observations). This violates **AX** and estimates of ϕ will be biased. As shown in section 1.3 the ARMA parameters can be consistently estimated if $\overline{\mathbf{AX}}$ (i.e. Y_{t-1} and ϵ_t are uncorrelated) holds. In other words, the bias – due to the unavoidable violation of **AX** by the lagged dependent variable – disappears in large samples. We use again the simple AR(1) model with $\nu=0$ and consider

$$\mathrm{cov}[Y_{t-1},\epsilon_t] = \mathrm{E}[Y_{t-1}\epsilon_t] - \mathrm{E}[Y_{t-1}]\mathrm{E}[\epsilon_t] = \mathrm{E}[Y_{t-1}(Y_t - \phi_1 Y_{t-1})].$$

We can use (42) to obtain

$$\mathrm{cov}[Y_{t-1},\epsilon_t] = \mathrm{E}[Y_t Y_{t-1}] - \phi_1 \mathrm{E}[Y_{t-1}^2] = \gamma_1 - \phi_1\gamma_0 = 0.$$

Thus we can estimate the parameters of ARMA models consistently by OLS. Note that the presence of MA terms does not cause any problems if ϵ_t is white-noise (see section 2.2.5 below), and thus does not violate $\overline{\mathbf{AX}}$.

Table 2 illustrates the magnitude of the bias associated with estimating AR coefficients. The true model is an AR(1), the estimated model is $y_t=c+f_1 y_{t-1}+e_t$. The table shows the means and standard errors of the OLS estimates $\bar{c}=c/(1-f_1)$, $\bar{y}$ and f_1 obtained from 10000 simulated realizations of AR(1) processes with $\mu=0.5$. For $\phi_1\geq 0.9$ and in small samples there are strong biases and large standard errors. While $\bar{c}$ and $\bar{y}$ are almost unbiased for $\phi_1<0.9$, f_1 remains biased but the bias is reduced as n grows.

Given an observed time series y_t ($t=1,\ldots,n$) we formulate the model

$$y_t = c + f_1 y_{t-1} + \cdots + f_p y_{t-p} + h_1 e_{t-1} + \cdots + h_q e_{t-q} + e_t.$$

The parameters c, f_i and h_i are estimated such that the sum of squared residuals (errors) is minimized:[92]

$$\sum_{t=\max\{p,q\}+1}^{n} e_t^2 \longrightarrow \min.$$

Before the model is estimated it is necessary to determine p and q. This choice can be based upon comparing estimated (partial) autocorrelations to the theoretical (partial)

[92] In general, the estimation procedure is iterative. Whereas lagged y_t are fixed explanatory variables (in the sense that they do not depend on the coefficients to be estimated) the lagged values of e_t depend on the parameters to be estimated. For details see Box and Jenkins (1976), p.208.

Table 2: Means and standard errors (in parentheses) across 10000 estimates of AR(1) series for different sample sizes.

ϕ_1	n	$\bar{c}$		$\bar{y}$		f_1	
	50	0.413	(142)	0.497	(3.79)	0.886	(0.08)
0.990	100	0.373	(93.3)	0.481	(4.33)	0.938	(0.04)
	200	0.538	(31.1)	0.410	(4.47)	0.964	(0.02)
	50	0.552	(3.75)	0.514	(1.24)	0.818	(0.09)
0.900	100	0.501	(1.00)	0.505	(0.93)	0.859	(0.06)
	200	0.513	(0.70)	0.512	(0.68)	0.881	(0.04)
	50	0.502	(0.28)	0.502	(0.28)	0.448	(0.13)
0.500	100	0.499	(0.20)	0.499	(0.20)	0.475	(0.09)
	200	0.499	(0.14)	0.499	(0.14)	0.487	(0.06)
	50	0.500	(0.18)	0.500	(0.18)	0.169	(0.14)
0.200	100	0.499	(0.13)	0.499	(0.13)	0.182	(0.10)
	200	0.500	(0.09)	0.500	(0.09)	0.192	(0.07)
	50	0.500	(0.12)	0.500	(0.12)	–0.208	(0.14)
–0.200	100	0.501	(0.09)	0.501	(0.09)	–0.204	(0.10)
	200	0.500	(0.06)	0.500	(0.06)	–0.203	(0.07)
	50	0.501	(0.09)	0.501	(0.10)	–0.491	(0.12)
–0.500	100	0.500	(0.07)	0.500	(0.07)	–0.496	(0.09)
	200	0.500	(0.05)	0.500	(0.05)	–0.497	(0.06)
	50	0.501	(0.07)	0.501	(0.08)	–0.871	(0.08)
–0.900	100	0.501	(0.05)	0.500	(0.05)	–0.884	(0.05)
	200	0.500	(0.04)	0.500	(0.03)	–0.892	(0.04)

autocorrelations in Table 1. Alternatively, model selection criteria like the **Akaike information criterion** (AIC) or the **Schwarz criterion** (SC) can be used. AIC and SC are based on the log-likelihood[93] ℓ=ln L and the number of estimated parameters K (for an ARMA(p,q) model with a constant term K=p+q+1):

$$\text{AIC} = -\frac{2\ell}{n} + \frac{2K}{n} \qquad \text{SC} = -\frac{2\ell}{n} + \frac{K \ln n}{n}.$$

If the type of model cannot be uniquely determined from Table 1, several models are estimated and the model with *minimal* AIC or SC is selected.

2.2.5 Diagnostic checking of ARMA models

ARMA model building is not complete unless the residuals are white-noise. The consequence of residual autocorrelation is inconsistency (see section 1.7.3). This can be shown in terms of the simple model

$$Y_t = \nu + \phi Y_{t-1} + u_t \qquad u_t = \rho u_{t-1} + \epsilon_t.$$

[93] ℓ is defined as

$$\ell = -\frac{n}{2}\left[1 + \ln(2\pi) + \ln\left(\frac{1}{n}\sum_t e_t^2\right)\right].$$

To derive $\mathrm{E}[Y_{t-1}u_t]$ we use

$$Y_{t-1} = \frac{\nu}{1-\phi} + \frac{u_{t-1}}{1-\phi B} = \mu + u_{t-1}(1 + \phi B + \phi^2 B^2 + \cdots) = \mu + \sum_{i=0}^{\infty} u_{t-1-i}.$$

Hence

$$\mathrm{E}[Y_{t-1}u_t] = \mathrm{E}\left[\sum_{i=0}^{\infty} u_{t-1-i}u_t\right]$$

depends on the autocorrelations of u_t. $\mathrm{E}[Y_{t-1}u_t]$ will be non-zero as long as $\rho \neq 0$, and this will give rise to inconsistent estimates. Thus, it is essential that the model is specified such that the residuals are white-noise. This requirement can also be derived from an alternative viewpoint. The main purpose of a time series model is to extract all dynamic features from a time series. This objective is achieved if the residuals are white-noise.

Autocorrelation of residuals can be removed by changing the model specification (mainly by including additional AR or MA terms). The choice may be based on patterns in (partial) autocorrelations of the residuals. AIC and SC can also be used to support the decision about including additional lags. However, it is not recommended to include lags that cannot be meaningfully interpreted. For instance, even if the coefficient of y_{t-11} is 'significant' in a model for daily returns this is a highly questionable result.

Indications about possible misspecifications can be derived from the inverted roots of the AR and MA polynomials. If two inverted roots of the AR and the MA polynomial are similar in magnitude, the model possibly contains redundant terms (i.e. the model order is too large). This situation is known as **overfitting**. If the absolute value of one of the inverted AR roots is close to or above 1.0 then the autoregressive term implies non-stationary behavior. This indicates the need to take differences of the time series (we will return to that point in section 2.3.3). An absolute value of one of the inverted MA roots close to or above 1.0 indicates that the series is **overdifferenced**. Taking first differences of a white-noise series $y_t = \epsilon_t$ leads to

$$\Delta y_t = \epsilon_t - \epsilon_{t-1}.$$

The resulting series Δy_t is 'best' described by an MA(1) model with $\theta_1 = -1.0$. Its autocorrelations can be shown to be decaying. However, a white-noise must not be differenced at all, and it does not make sense to fit a model to Δy_t. Similar considerations hold for stationary series in general: they must not be differenced.

If residuals are white-noise but not homoscedastic, modifications of the ARMA model equation are not meaningful. Heteroscedasticity of the disturbances cannot be removed with a linear time series model for the conditional mean. Models to account for residuals that are not normally distributed and/or heteroscedastic will be introduced in section 2.5.

2.2.6 Example 35: ARMA models for FTSE and AMEX returns

Box and Jenkins (1976) have proposed a modeling strategy which consists of several steps. In the identification step one or several preliminary models are chosen on the basis of (partial) autocorrelations and the patterns in Table 1. After estimating each model the

Table 3: Results of fitting various ARMA models to FTSE log returns.

model	$f_{\ell\ell}$	p-value	AIC	SC	model	AIC	SC
null model			–2.618	–2.606			
AR(1)	0.113	0.046	–2.622	–2.597	MA(1)	–2.629	–2.605
AR(2)	–0.118	0.039	–2.626	–2.590	MA(2)	–2.636	–2.600
AR(3)	0.123	0.032	–2.632	–2.583	MA(3)	–2.634	–2.586
AR(4)	0.022	0.708	–2.622	–2.562	MA(4)	–2.632	–2.572

residuals are analyzed – mainly to test for any remaining autocorrelation. If necessary, the models are modified and estimated again. A model is used for forecasting if its residuals are white-noise and its coefficients are significant. If there are several competing models which fulfill these requirements, model selection criteria can be used to make a final choice. We illustrate this procedure by considering FTSE and AMEX log returns.

The autocorrelations of FTSE log returns are rather small (see Figure 5) and cannot be easily associated with the theoretical patterns in Table 1. To determine p and q we fit several models with increasing order and observe AIC, SC and the partial autocorrelations. Table 3 summarizes the results. Partial autocorrelations and AIC indicate that a AR(3) model is appropriate. The estimated model is

$$y_t = \underset{(0.075)}{0.0067} + \underset{(0.014)}{0.14}\ y_{t-1} - \underset{(0.02)}{0.13}\ y_{t-2} + \underset{(0.03)}{0.12}\ y_{t-3} + e_t \quad s_e = 0.06449.$$

Note that the p-values may be biased if the residuals e_t do not have the properties reviewed in section 1.7.

The overall minimum AIC indicates an MA(2) model:

$$y_t = \underset{(0.044)}{0.0077} + \underset{(0.001)}{0.15}\ e_{t-1} - \underset{(0.05)}{0.11}\ e_{t-2} + e_t \quad s_e = 0.06445.$$

In both models the standard deviation of residuals (the standard error) s_e is not very different from the standard deviation of log returns s_y 0.06526 (see example 31). This indicates that the conditional mean $\hat{y}_t$ from these models is very close to the unconditional mean $\bar{y}$.

p=0 would be chosen on the basis of SC. In this case the conditional and unconditional mean are identical and the standard error is equal to the standard deviation of returns:

$$y_t = 0.00765 + e_t \quad s_e = 0.06526.$$

The ARMA(1,1) model (AIC=–2.633, SC=–2.596)

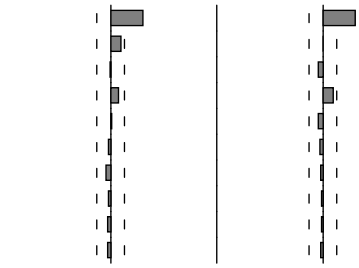
$$y_t = \underset{(0.073)}{0.011} - \underset{(0.039)}{0.47}\ y_{t-1} + \underset{(0.002)}{0.62}\ e_{t-1} + e_t \quad s_e = 0.06457,$$

is not supported by AIC or SC. It is worth mentioning that all p-values of the ARMA(2,2) model (AIC=–2.65, SC=–2.589)

$$y_t = 0.019 - 1.04y_{t-1} - 0.84y_{t-2} + 1.17e_{t-1} + 0.88e_{t-2} + e_t \quad s_e = 0.0638$$

Figure 7: Autocorrelogram of AMEX log returns.

Sample: 1 209
Included observations: 208



	AC	PAC	Q-Stat	Prob
1	0.318	0.318	21.325	0.000
2	0.103	0.002	23.566	0.000
3	-0.006	-0.044	23.574	0.000
4	0.074	0.098	24.745	0.000
5	0.010	-0.042	24.767	0.000
6	-0.024	-0.030	24.890	0.000
7	-0.043	-0.018	25.293	0.001
8	-0.024	-0.010	25.414	0.001
9	-0.025	-0.015	25.555	0.002
10	-0.032	-0.019	25.783	0.004

are equal to zero. This would be the 'optimal' model according to AIC. However, the model has inverted AR roots $-0.52 \pm 0.75i$ and MA roots $-0.59 \pm 0.73i$ which are very similar. This situation is known as overfitting: Too many, redundant parameters have been estimated and a model with less coefficients is more appropriate. The ratio of MA and AR polynomials $(1+0.135B-0.095B^2-0.014B^3+0.094B^4-0.086B^5+\cdots)$ has coefficients which are rather small and similar to the MA(2) model. The significance tests of individual coefficients for an overfitted model have very limited value.

We apply diagnostic checking to the residuals from the MA(2) and the AR(3) model. The p-values of Q_{10} to test for autocorrelation in residuals are 0.226 and 0.398, which indicates that residuals are white-noise. For squared residuals the p-values of Q_5 are 0.03 (MA) and 0.35 (AR); p-values of Q_{10} are 0.079 (MA) and 0.398 (AR). Thus MA residuals are not quite homoscedastic. The p-values of the JB-test are 0.0 for both models ($S\approx 0.5$ and $U\approx 13$) which rejects normality. Thus the significance tests of the estimated parameters may be biased.

The (partial) autocorrelations of AMEX log returns (see Figure 7) may be viewed to suggest a MA(1) model. The estimated model is

$$y_t = \underset{(0.94)}{3.6 \cdot 10^{-5}} + \underset{(0.0)}{0.28}\, e_{t-1} + e_t \qquad s_e = 0.005239.$$

The residuals are white-noise but not normally distributed. The squared residuals are correlated and indicate heteroscedasticity (i.e. the residuals are not independent).

> **Exercise 20:** Use the log returns defined in exercise 17. Identify and estimate suitable ARMA models and check their residuals.

2.2.7 Forecasting with ARMA models

Forecasting makes statements about the process Y_t at a future date $t+\tau$ on the basis of information available at date t. The forecast $\hat{Y}_{t,\tau}$ is the conditional expected value

$$\hat{Y}_{t,\tau} = \mathrm{E}[Y_{t+\tau}|Y_t, Y_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots] = \mathrm{E}[Y_{t+\tau}|I_t] \qquad \tau = 1, 2, \ldots$$

using the model equation. τ is the **forecasting horizon**.

Forecasts for future dates $t+\tau$ (τ=1,2,...) from the *same* date t are called **dynamic** (or **multi-step**) forecasts. The one-step ahead forecast $\hat{Y}_{t,1}$ is the starting point. The next dynamic forecast $\hat{Y}_{t,2}$ (for $t+2$) is also made in t and uses $\hat{Y}_{t,1}$. In general, a dynamic forecast $\hat{Y}_{t,\tau}$ depends on all previous dynamic forecasts (see below). **Static** forecasts are a sequence of one-step ahead forecasts $\hat{Y}_{t,1}$ $\hat{Y}_{t+1,1}$ ... made at different points in time.

AR model forecasts

Dynamic AR(1) model forecasts are given by:

$$\begin{aligned}
\hat{Y}_{t,1} &= \nu + \phi_1 Y_t \\
\hat{Y}_{t,2} &= \nu + \phi_1 \mathrm{E}[Y_{t+1}|I_t] = \nu + \phi_1 \hat{Y}_{t,1} = \nu + \phi_1(\nu + \phi_1 Y_t) \\
\hat{Y}_{t,\tau} &= \nu(1 + \phi_1 + \phi_1^2 + \cdots + \phi_1^{\tau-1}) + \phi_1^\tau Y_t
\end{aligned}$$

$$\lim_{\tau \longrightarrow \infty} \hat{Y}_{t,\tau} = \frac{\nu}{(1-\phi_1)} = \mu.$$

Unknown future values Y_{t+1} are replaced by the forecasts $\hat{Y}_{t,1}$. Forecasts of AR(1) models decay exponentially to the unconditional mean of the process μ. The rate of decay depends on $|\phi_1|$. Dynamic forecasts of stationary AR(p) models show a more complicated pattern but also correspond to the autocorrelations. The forecasts converge to

$$\frac{\nu}{(1-\phi_1-\cdots-\phi_p)} = \mu.$$

Note that ν is estimated by the constant term c in the model

$$\hat{y}_t = c + f_1 y_{t-1} + \cdots + f_p y_{t-p}.$$

Forecasts from an estimated AR model are determined by the estimated parameters in the same way as described above. μ is estimated by $\bar{c}$=c/(1−f_1−⋯−f_p) which need not agree exactly with the sample mean $\bar{y}$.[94]

MA model forecasts

Dynamic MA(q) model forecasts are given by:

$$\begin{aligned}
\hat{Y}_{t,1} &= \mu + \theta_1 \epsilon_t + \cdots + \theta_q \epsilon_{t-q+1} \\
\hat{Y}_{t,2} &= \mu + \theta_1 \mathrm{E}[\epsilon_{t+1}|I_t] + \theta_2 \epsilon_t + \cdots = \mu + \theta_2 \epsilon_t + \cdots \\
\hat{Y}_{t,\tau} &= \mu \qquad (\tau > q).
\end{aligned}$$

[94] In EViews AR models can be estimated with two different specifications. The *lag specification* LS Y C Y(-1) Y(-2) ... estimates the model

$$y_t = c + f_1 y_{t-1} + f_2 y_{t-2} + \cdots + e_t.$$

c is an estimate of ν. Using the *AR specification* LS Y C AR(1) AR(2) ... however, EViews estimates the model

$$y_t = \bar{c} + u_t \qquad u_t = f_1 u_{t-1} + f_2 u_{t-2} + \cdots + e_t,$$

where $\bar{c}$=c/(1−f_1−f_2−⋯) (using c from the lag specification) is an estimate of μ. The estimated coefficients f_i from the two specifications are identical.

The unknown future disturbance terms $\epsilon_{t+\tau}$ are replaced by their expected value zero. Forecasts of MA(q) processes cut off to μ after q periods. Thus the forecasting behavior corresponds to the autocorrelation pattern.

Forecasts based on an estimated MA model are determined in the same way. The unconditional mean μ is estimated by the constant term c in the model

$$\hat{y}_t = c + h_1 e_{t-1} + \cdots + h_q e_{t-q}$$

which need not agree exactly with the sample mean $\bar{y}$.

ARMA model forecasts

Forecasts using ARMA models can be derived in a similar fashion as described for AR and MA models. The behavior of dynamic forecasts corresponds to the autocorrelations (see Table 1). Once the contribution from the MA part has vanished the forecasting behavior is driven by the AR part.

To investigate the behavior of ARMA forecasts for $\tau \to \infty$ we make use of Wold's decomposition. It implies that the coefficients ψ_i of the MA(∞) representation approach zero as $i \to \infty$. This has two consequences:

1. Dynamic forecasts $\hat{Y}_{t,\tau}$ converge to μ if Y_t is stationary. Since $\hat{Y}_{t,\tau}$ is the conditional expected value of $Y_{t+\tau}$ this implies that returns are expected to approach their unconditional mean. This property is called **mean reversion** which requires stationarity. The speed of mean reversion depends on the coefficients ψ_ℓ (i.e. on the autocorrelations of the process).

2. The MA(∞) representation implies that the variance of the forecast errors $\epsilon_{t,\tau} = Y_t - \hat{Y}_{t-\tau,\tau}$ converges to the variance of Y_t.[95] The variance of $\epsilon_{t,\tau}$ can be used to compute forecast (confidence) intervals.

If the process is non-stationary (see section 2.3) $\psi(B)$ can be written as

$$\psi(B) = \frac{\theta(B)}{(1-B)\phi(B)} = (1 + B + B^2 + \cdots)\frac{\theta(B)}{\phi(B)}. \tag{44}$$

Thus, the polynomial $\psi(B)$ does not converge. This implies that the (non-stationary) process is not mean reverting and the forecast variance does not converge.

2.2.8 Properties of ARMA forecast errors

The properties of forecast errors are also based on the MA(∞) representation. The τ-step ahead forecast error is given by $Y_{t+\tau} - \hat{Y}_{t,\tau}$. The following properties hold if the forecast (the conditional expectation) is based on the correct process definition:

1. Expected value of the τ-step ahead forecast error:

$$\mathrm{E}[Y_{t+\tau} - \hat{Y}_{t,\tau}] = 0.$$

[95] For details see Tsay (2002), p.53.

2. Variance of the τ-step ahead forecast error[96]:

$$\mathrm{V}[\hat{Y}_{t,\tau}] = \sigma_\tau^2 = \sigma_\epsilon^2 \sum_{i=0}^{\tau-1} \psi_i^2.$$

3. Variance of the one-step ahead forecast error: $\sigma_1^2{=}\sigma_\epsilon^2$.

4. Forecast errors for a forecasting horizon τ behave like a MA($\tau{-}1$) process.

5. One-step ahead forecast errors are white-noise.

6. For $\tau{\longrightarrow}\infty$ the variance of forecast errors converges to the variance of the process:

$$\lim_{\tau \longrightarrow \infty} \sigma_\tau^2 \longrightarrow \sigma^2.$$

The forecast variance of integrated processes (see section 2.3) tends to ∞ because $\psi(B)$ defined in (44) implies *cumulating* the variance of forecast errors.

These properties may be used to determine a $(1{-}\alpha)$ confidence interval for forecasts that are calculated from an estimated ARMA model using n observations. The $(1{-}\alpha)$ forecast interval is given by

$$\hat{y}_{t,\tau} \pm T(\alpha/2, n) s_\tau,$$

where $T(\alpha, n)$ is the α-quantile of the t-distribution with n degrees of freedom, and s_τ is the estimated standard deviation of the τ-step ahead forecast error:

$$s_\tau^2 = s_e^2 \sum_{i=0}^{\tau-1} g_i^2 \qquad g(B) = \frac{h(B)}{f(B)}.$$

Example 36: Long-horizon returns revisited: In example 1.8.3 we have considered regressions with long-horizon returns. Now we will have a close look at the (partial) autocorrelations of such returns. For simplicity we assume that single-period returns are white-noise $\epsilon_t{\sim}\mathrm{N}(0, \sigma^2)$, and we consider the sum of only three consecutive, single-period returns:

$$y_t = \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2}.$$

Thus, y_t is a MA(2) process with parameters $\theta_1{=}\theta_2{=}1$. The autocovariances of y_t are $\gamma_1{=}2\sigma^2$, $\gamma_2{=}\sigma^2$, and $\gamma_\ell{=}0$ ($\ell{>}2$), so that the only non-zero autocorrelations are $\rho_1{=}2/3$ and $\rho_2{=}1/3$. These autocorrelations can be derived from (43). Partial autocorrelations can be obtained recursively and follow a very specific pattern. $\phi_{\ell\ell}$ at the 'seasonal' lags $\ell{=}1{,}4{,}7{,}\ldots$ are given by ρ_1/j ($j{=}1{,}2{,}\ldots$); for example, $\phi_{4,4}{=}1/3$. Partial autocorrelations at 'non-seasonal' lags $2{,}3{,}5{,}6{,}\ldots$ are all negative, and converge exponentially to zero. Empirically, the appropriate MA(2) model may be (easily) identified from the pattern of (partial) autocorrelations. However, the dynamic features captured by this model cannot be exploited in out-of-sample predictions of three-months returns for more than two periods ahead. These forecasts would be equal to the unconditional mean implied by the model parameters.

[96] There is no difference between the variance of the forecast and the variance of the forecast error if the expected value of the forecast error equals zero.

Exercise 21: Use the ARMA models from exercise 20. Estimate the *same* model for a subset of the available sample (omit about 10% of the observations at the end of the sample). Compute static and dynamic out-of-sample forecasts of returns and prices and compare them to the actual observations. Describe the behavior of forecasts and evaluate their quality.

2.3 Non-stationary models

2.3.1 Random-walk and ARIMA models

Consider an AR(1) process with parameter ϕ_1=1. The resulting non-stationary process is called **random-walk**:

$$Y_t = Y_{t-1} + \epsilon_t \qquad \epsilon_t \ldots \text{white-noise.}$$

The random-walk can be transformed into the stationary white-noise process ϵ_t by **differencing**:

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - B)Y_t = \epsilon_t.$$

A process that becomes stationary after differencing is also called **integrated** or **difference-stationary**. A random-walk can be written as the sum of all lags of ϵ_t

$$Y_t = \frac{\epsilon_t}{(1 - B)} = (1 + B + B^2 + \cdots)\epsilon_t = \sum_{i=-\infty}^{t} \epsilon_i,$$

which corresponds to *integrating* over ϵ_t.
If the first differences of a random-walk are white-noise but the mean of ϵ_t is different from zero then Y_t is a **random-walk with drift**:

$$Y_t = \nu + Y_{t-1} + \epsilon_t = \nu t + Y_0 + \sum_{i=0}^{t} \epsilon_i = \nu t + Y_0 + \omega_t.$$

This process has two trend components: the deterministic trend νt and the stochastic trend ω_t.
For a fixed, non-random initial value Y_0 the random-walk (with drift or without drift) has the following properties:

1. $\mathrm{E}[Y_t] = \nu t + Y_0$
2. $\mathrm{V}[Y_t] = t\sigma_\epsilon^2$
3. $\gamma_k = (t - k)\sigma_\epsilon^2$
4. r_k decay very slowly (approximately linearly).

A random-walk is non-stationary since mean, variance and autocovariance depend on t. Thus it is *not* mean-reverting and its (long-term) forecasts are given by

$$\mathrm{E}[\hat{Y}_{t,\tau}|Y_t] = \nu\tau + Y_t.$$

A general class of integrated processes can be defined, if the differences $Y_t - Y_{t-1}$ follow an ARMA(p, q) process:

$$Y_t = Y_{t-1} + U_t$$

$$U_t = \nu + \phi_1 U_{t-1} + \cdots + \phi_p U_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t.$$

In this case Y_t is an ARIMA$(p, 1, q)$ (integrated ARMA) process and Y_t is called integrated of order 1: $Y_t \sim I(1)$.

If Y_t is an ARMA(p, q) process after differencing d times so that

$$(1 - B)^d Y_t = U_t$$

$$U_t = \nu + \phi_1 U_{t-1} + \cdots + \phi_p U_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t,$$

Y_t is an ARIMA(p, d, q) process and $Y_t \sim I(d)$. Obviously, an ARIMA model for log prices is equivalent to an ARMA model for log returns.

Forecasts of the ARIMA(0,1,1) process $(1-B)Y_t = \nu + \theta_1 \epsilon_{t-1} + \epsilon_t$ are obtained by using the same procedure as in section 2.2.7:

$$\begin{aligned}
\hat{Y}_{t,1} &= Y_t + \nu + \theta_1 \epsilon_t \\
\hat{Y}_{t,2} &= \hat{Y}_{t,1} + \nu = Y_t + 2\nu + \theta_1 \epsilon_t \\
\hat{Y}_{t,\tau} &= Y_t + \nu\tau + \theta_1 \epsilon_t.
\end{aligned}$$

Forecasts of ARIMA(0,1,q) processes converge to a straight line with slope ν, where ν corresponds to the expected value of ΔY_t. The transition to the straight line is described by the MA parameters and corresponds to the cut off pattern of autocorrelations.

The ARIMA(1,1,0) process $(1-\phi_1 B)(1-B)Y_t = \nu + \epsilon_t$ can be written as

$$\Delta Y_t = \nu + \phi_1 \Delta Y_{t-1} + \epsilon_t \qquad Y_t = Y_{t-1} + \nu + \phi_1 \Delta Y_{t-1} + \epsilon_t = Y_{t-1} + \Delta Y_t,$$

and dynamic forecasts are obtained as follows:

$$\begin{aligned}
\hat{Y}_{t,1} &= Y_t + \Delta\hat{Y}_{t,1} = Y_t + \nu + \phi_1 \Delta Y_t \\
\hat{Y}_{t,2} &= \hat{Y}_{t,1} + \Delta\hat{Y}_{t,2} \\
&= Y_t + \Delta\hat{Y}_{t,1} + \Delta\hat{Y}_{t,2} = Y_t + [\nu + \phi_1 \Delta Y_t] + [\nu(1 + \phi_1) + \phi_1^2 \Delta Y_t] \\
\hat{Y}_{t,3} &= Y_t + \nu + \nu(1 + \phi_1) + \nu(1 + \phi_1 + \phi_1^2) + (\phi_1 + \phi_1^2 + \phi_1^3)\Delta Y_t.
\end{aligned}$$

Box and Jenkins (1976, p.152) show that the forecasts approach a straight line:

$$\lim_{\tau \longrightarrow \infty} \hat{Y}_{t,\tau} = Y_t + \mu\tau + (Y_t - Y_{t-1} - \mu)\frac{\phi_1}{(1 - \phi_1)} \qquad \mu = \frac{\nu}{(1 - \phi_1)}.$$

In general, forecasts of ARIMA(p,1,0) processes approach a straight line with slope μ, which is the expected value of ΔY_t. The transition to the straight line is described by the AR parameters, and corresponds to the pattern of autocorrelations.

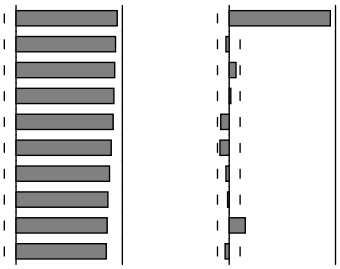
The process

$$Y_t = \nu_0 + \nu t + U_t$$

is a **trend-stationary** process. U_t is stationary but need not be white-noise. The process Y_t evolves around a linear, deterministic trend in a stationary way. The appropriate

Figure 8: Autocorrelogram of the FTSE.

Sample: 1965:01 1990:12
Included observations: 312



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.988	0.988	307.47	0.000
		2	0.975	-0.027	608.14	0.000
		3	0.965	0.068	903.16	0.000
		4	0.955	0.023	1193.0	0.000
		5	0.943	-0.078	1476.6	0.000
		6	0.929	-0.086	1752.8	0.000
		7	0.915	-0.026	2021.5	0.000
		8	0.901	-0.008	2283.0	0.000
		9	0.891	0.160	2539.6	0.000
		10	0.880	-0.040	2790.9	0.000

transformation to make this process stationary is to subtract the trend term $\nu_0+\nu t$ from Y_t. Note that differencing a trend stationary process does not only eliminate the trend but also affects the autocorrelations of ΔY_t:

$$Y_t - Y_{t-1} = \nu + U_t - U_{t-1}.$$

In general, the autocorrelations of $U_t - U_{t-1}$ are not zero. For instance, if U_t is white-noise $Y_t - Y_{t-1}$ is a MA(1) process with parameter $\theta_1 = -1$ and $\rho_1 = -0.5$ (see (43), p.104).

Example 37: Many financial time series (prices, indices, rates) or their logarithm are non-stationary. The autocorrelations of a non-stationary series decay very slowly (approximately linearly) and r_1 is close to 1.0. The autocorrelogram of the FTSE in Figure 8 shows this typical pattern.

2.3.2 Forecasting prices from returns

Frequently, a model fitted to returns is also used to obtain fitted values or forecasts of the corresponding prices. Suppose $\hat{y}_t$ is the conditional mean of log returns y_t=ln(p_t/p_{t-1}) derived from a ARMA model, and assume that the residuals are normal. Given the properties of the lognormal distribution (see section 2.1.2) the expected value of the price is given by

$$\hat{p}_t = p_{t-1} \exp\{\hat{y}_t + 0.5 s_e^2\}, \tag{45}$$

where s_e^2 is the variance of the residuals from the estimated model which determines $\hat{y}_t$.

Example 38: We fit the ARMA(1,1) model

$$y_t = \underset{(0.0698)}{0.0078} + u_t \qquad u_t = \underset{(0.049)}{-0.473}\, u_{t-1} + \underset{(0.03)}{0.626}\, e_{t-1} + e_t \qquad s_e = 0.06547$$

to the FTSE log returns using the period 1965:01 to 1988:12 (see file `ftse.wf1` for details). The one-step ahead (static) out-of-sample forecasts of the index are close to the actual values of the index and the dynamic forecasts quickly converge to a line with almost constant slope. The slope of this line can be determined as follows. Dynamic τ-period ahead forecasts of the index are given by

$$\hat{p}_{t,\tau} = p_t \exp\left\{\sum_{i=1}^{\tau} \hat{y}_{t,i} + \tau 0.5 s_e^2\right\},$$

where $\hat{y}_{t,i}$ are out-of-sample forecasts from the ARMA model.[97] Dynamic forecasts $\hat{y}_{t,\tau}$ converge to the constant $\bar{c}$=0.0078 but the changes in the index do not converge to a constant:

$$\hat{p}_{t,\tau} - \hat{p}_{t,\tau-1} = p_t[\exp\{\tau(\bar{c} + 0.5 s_e^2)\} - \exp\{(\tau-1)(\bar{c} + 0.5 s_e^2)\}].$$

[97] Note: EViews does not include the term $0.5 s_e^2$.

2.3.3 Unit-root tests

An AR process is stationary, if the AR polynomial $\phi(B)$ has *no* inverted root on or outside the unit circle. If there is a root *on* the unit circle (a so-called **unit-root**), $\phi(B)$ can be decomposed as follows:

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_{p-1} B^{p-1})(1 - B).$$

The term $(1-B)$ corresponds to the unit-root and implies taking first differences. The existence of a unit-root has considerable consequences for the behavior of the process and its forecasts. The dynamic forecasts of Y_t converge to a straight line with a slope equal to the expected value of ΔY_t, and the forecast interval (which is based on the variance of the forecast error) diverges. If there is no unit-root, forecasts of Y_t converge to the (unconditional) mean of Y_t, and the variance of forecast errors converges to the variance of Y_t.

> **Example 39:** The polynomial of the AR(2) model $(1-1.8B+0.8B^2)Y_t=\epsilon_t$ has two inverted roots (1.0 and 0.8) and can be decomposed into $\phi(B)=(1-0.8B)(1-B)$. Thus Y_t is an ARIMA(1,1,0) process and integrated $Y_t \sim I(1)$.
>
> The AR(2) model $(1-1.8B+0.81B^2)Y_t=\epsilon_t$ is only marginally different. However, its inverted roots are both equal to 0.9. There is no unit-root and the process is stationary $Y_t \sim I(0)$.

ARMA models are only suitable for stationary time series. One way to deal with integrated time series is to take first (or higher order) differences, which is not appropriate if the series is trend-stationary. Empirically, it is very difficult to distinguish trend-stationary and difference-stationary processes. A slow, approximately linear decay of autocorrelations and r_1 close to one are (heuristic) indicators of an integrated series. However, there are integrated processes where the autocorrelations of first differences decay slowly, but the decay starts at $r_1 \approx 0.5$ rather than 1.0. The ARIMA(0,1,1) process $\Delta Y_t=(1-0.8B)\epsilon_t$ is an example of this case (see Box and Jenkins, 1976, p.200).

The **Dickey-Fuller (DF)** unit-root test is based on the equation

$$\Delta Y_t = \nu + (\phi_1 - 1)Y_{t-1} + \epsilon_t = \nu + \gamma Y_{t-1} + \epsilon_t \qquad \gamma = \phi_1 - 1$$

$$H_0 : \gamma = 0 \quad (\phi_1 = 1) \quad \text{(unit-root)} \qquad H_a : \gamma < 0 \quad |\phi_1| \quad \text{(stationary)}.$$

$\gamma=0$ if Y_t is integrated, and the estimate $\hat{\gamma}$ should be close to zero. When $\hat{\gamma}$ is significantly *less*[98] than zero, the null hypothesis is rejected, and Y_t is assumed to be stationary. However, it is not straightforward to test $\gamma=0$ (or $\phi_1=1$) based on the null hypothesis of a unit-root and the estimated equation

$$\Delta y_t = c + (f_1 - 1)y_{t-1} + e_t = c + \hat{\gamma} y_{t-1} + e_t.$$

According to Fuller (1976) the t-statistic $(f_1-1)/\text{se}[f_1]$ is *not* t-distributed under the null hypothesis (irrespective of n). He shows that $n(f_1-1)$ has a non-degenerate distribution with two main characteristics: $\hat{\gamma}$ is downward biased if $\phi=1$ (a fact also indicated by the

[98] The unit-root test is a *one-sided* test since the coefficient γ is *negative* under H_a.

Table 4: Critical values for the ADF test.

	without trend			with trend		
α	0.01	0.05	0.10	0.01	0.05	0.10
n=50	–3.58	–2.93	–2.60	–4.15	–3.50	–3.18
n=100	–3.51	–2.89	–2.58	–4.04	–3.45	–3.15
n=250	–3.46	–2.88	–2.57	–3.99	–3.43	–3.13
n=500	–3.44	–2.87	–2.57	–3.98	–3.42	–3.13
n=∞	–3.43	–2.86	–2.57	–3.96	–3.41	–3.12

results in Table 2), and the variance of $\hat{\gamma}$ under the null hypothesis is of order $1/n^2$ (rather than the usual order $1/n$). Critical values have to be derived from simulations since no analytical expression is available for that distribution. H_0 is rejected if the t-statistic of $\hat{\gamma}$ is less than the corresponding critical value in Table 4[99].

The critical values in Table 4 are valid, even if ϵ_t is heteroskedastic. However, ϵ_t must be white-noise. If ϵ_t is not white-noise, the DF test equation has to be extended (**augmented Dickey-Fuller** (ADF) test)[100]:

$$\Delta Y_t = \nu + \gamma Y_{t-1} + \sum_{i=1}^{p} c_i \Delta Y_{t-i} + \epsilon_t. \tag{46}$$

It is recommended to choose p=$n^{1/3}$, but AIC or SC can be used as well to choose p. Insignificant coefficients c_i should be eliminated, but if in doubt, too large values of p are not very harmful. The **Phillips-Perron test** does not account for the (possible) autocorrelations in the residuals by adding lags to the DF equation. Instead, the test statistic $(f_1-1)/\text{se}[f_1]$ is adjusted for autocorrelation like in the computation of Newey-West standard errors. The critical values are the same as in the ADF test.

If a series shows a more or less monotonic trend it can be either trend-stationary or an integrated series (e.g. a random-walk) with drift. Consider the integrated process

$$Y_t = \nu + Y_{t-1} + W_t$$

where W_t is stationary (hence, Y_t is called **difference-stationary**). This process can be written as

$$Y_t = Y_0 + \nu t + \sum_{i=0}^{t} W_i,$$

[99] Source: Fuller (1976), p.373.

[100] To derive this specification on the basis of an AR(p+1) model we set ρ=ϕ_1+$\cdots$+ϕ_{p+1}, c_s=$-(\phi_{s+1}+\cdots+\phi_{p+1})$ (s=1,...,p), reformulate the AR polynomial of order p+1 as

$$(1-\rho B) - (c_1 B + \cdots + c_p B^p)(1-B),$$

and write the AR(p+1) model as

$$Y_t = \nu + \rho Y_{t-1} + c_1 \Delta Y_{t-1} + \cdots + c_p \Delta Y_{t-p} + \epsilon_t.$$

The ADF test equation is obtained by subtracting Y_{t-1} from both sides and setting γ=$\rho-1$.

where the sum of (stationary) disturbances makes this process evolve around a linear trend in an integrated (non-stationary) way. If W_t is white-noise, Y_t is a random-walk with drift. A natural alternative to this process is the **trend-stationary** process

$$Y_t = \nu_0 + \nu t + U_t$$

with stationary disturbances U_t. Although both W_t and U_t are stationary in these specifications, their properties have to be quite different to make the resulting series appear similar. For example, if W_t is white-noise with σ_W=1, and U_t is an AR(1) with ϕ_1=0.9 and σ_Y=5, some similarity between sample paths of those processes can be obtained (see file `nonstationary.xls`).

A unit-root test to distinguish among these alternatives is based on estimating the equation

$$\Delta y_t = \hat{\gamma} y_{t-1} + c_0 + ct + \sum_{i=1}^{p} c_i \Delta y_{t-i} + e_t,$$

and the critical values from Table 4 (column 'with trend'). If H_0 is not rejected, y_t is concluded to be integrated with a drift corresponding to $-c/\hat{\gamma}$ (assuming that $\hat{\gamma}$<0 in any finite sample). If H_0 is rejected, y_t is assumed to be trend-stationary with slope $\approx -c/\hat{\gamma}$.

If a series shows no clear trends, a unit-root test can be used to decide whether the series is stationary or integrated without a drift. The integrated process

$$Y_t = Y_{t-1} + W_t \qquad W_t \ldots \text{stationary}$$

can be written as

$$Y_t = Y_0 + \sum_{i=0}^{t} W_i,$$

where the sum introduces non-stationarity. If W_t is white-noise Y_t is a random-walk without drift. A natural alternative to this process is the stationary process

$$Y_t = \nu_0 + U_t$$

with stationary disturbances. In this case we estimate the test equation

$$\Delta y_t = \hat{\gamma} y_{t-1} + c_0 + \sum_{i=1}^{p} c_i \Delta y_{t-i} + e_t.$$

If H_0 is rejected, y_t is assumed to be stationary. If H_0 is not rejected, y_t is assumed to be integrated. In both cases, c_0 is proportional to the mean of y_t with factor $-1/\hat{\gamma}$.

In general, unit root tests should be interpreted with caution. The power of unit-root tests is *low*, which means that stationary processes are *too frequently* assumed to be integrated (in particular if ϕ_1 is close to one). Including irrelevant, deterministic regressors (e.g. constant or trend) in the test equation reduces the power of the test even further (since critical values become more negative). On the other hand, if constant or trend terms are omitted although they belong to the true data generating process, the power can

go to zero. Choosing the correct specification is difficult, because two important issues are interrelated: Unit root tests depend on the presence of deterministic regressors, and conversely, tests for the significance of such regressors depend on the presence of a unit root. A standard recommendation is to choose a specification of the test equation that is plausible under the null *and* the alternative hypotheses (see Hamilton (1994), p.501, or the guidelines in Enders (2004), p.207.)

Kwiatkowski et al. (1992) have proposed a test based on the null hypothesis of (trend) stationarity. The **KPSS test** runs a regression of y_t on a constant (if H_0 is stationarity), or a constant and a time trend t (if H_0 is trend stationarity). The residuals are used to compute the test statistic

$$\sum_{t=1}^{n} \frac{S_t^2}{\hat{\sigma}_e^2} \qquad S_t = \sum_{i=1}^{t} e_i.$$

$\hat{\sigma}_e^2$ is an estimate of the residual variance that accounts for autocorrelation as in the Newey-West estimator. The asymptotic critical values of the KPSS statistic are tabulated in Kwiatkowski et al. (1992, p.166). For α=0.05 the critical value under the null of stationarity is 0.463, and 0.146 for trend stationarity.

Example 40: A unit-root test of the spread between long- and short-term interest rates in the UK[101] leads to ambiguous conclusions. The estimated value of γ is -0.143679 and gives the impression that ϕ_1=1+γ is sufficiently far away from one. The t-statistic of $\hat{\gamma}$ is -3.174306. Although this is *below* the critical value at a 5% significance level (-2.881), it is *above* the critical value for α=0.01 (-3.4758). Therefore the unit-root hypothesis can only be rejected for high significance levels and it remains unclear, whether the spread can be considered stationary or not. However, given the low power of unit-root tests it may be appropriate to conclude that the spread is stationary. The KPSS test confirms this conclusion since the test statistic is far below the critical values. Details can be found in the files `spread.wf1` or `spread.R`.

Example 41: We consider a unit-root test of the AMEX index (see files `amex.wf1` or `amex.R`). Since the index does not follow a clear trend, we do not include a trend term in the test equation. We use p=1 since the coefficients $\hat{c}_i$ (i>1) are insignificant (initially p=6 ($\approx 209^{1/3}$) was chosen). The estimate for γ is -0.0173 and has the expected negative sign. The t-statistic of $\hat{\gamma}$ is -1.6887, and is clearly *above* all critical values in Table 4: It is also above the critical values provided by EViews. Therefore the unit-root hypothesis *cannot* be rejected and the AMEX index is assumed to be integrated (of order one). This is partially confirmed by the KPSS test. The test statistic 0.413 exceeds the critical value only at the 10% level, but stationarity cannot be rejected for lower levels of α. To derive the implied mean of y_t from the estimated equation $\Delta\hat{y}_t$=$-0.0173y_{t-1}$+7.998+$0.331\Delta y_{t-1}$ we reformulate the equation as $\hat{y}_t$=(1$-$0.0173+0.331)y_{t-1}+7.998$-$$0.331y_{t-2}$, and the implied mean is given by 7.998/0.0173$\approx$462.

Example 42: We consider a unit-root test of the log of the FTSE index (see file `ftse.wf1`). We use only data from 1978 to 1986 since during this period it is not clear whether the series has a drift or is stationary around a linear trend. This situation requires to include a trend term in the test equation. The estimated equation is

[101] Source: http://www.lboro.ac.uk/departments/ec/cup/data.html; 'Yield on 20 Year UK Gilts' (long; file `R20Q.txt`) and '91 day UK treasury bill rate' (short; file `RSQ.htm`); the spread is the difference between long and short; quarterly data from 1952 to 1988; 148 observations.

$\Delta\hat{y}_t = -0.163y_{t-1} + 0.868 + 0.0023t$. The t-statistic of $\hat{\gamma}$ is -3.19. This is *above* the 1% and 5% critical values in Table 4 and slightly below the 10% level. Therefore the unit-root hypothesis *cannot* be rejected, and the log of the FTSE index can be assumed to be integrated (of order one). This is confirmed by the KPSS test where the test statistic exceeds the critical value, and stationarity can be rejected. Since augmented terms are not necessary, the log of the index can be viewed as a random walk with drift approximately given by 0.0023/0.163=0.014.

Exercise 22: Consider the ADF test equation (46) and p=1. Show that the implied sample mean of y_t is given by $-\hat{\nu}/\hat{\gamma}$.

Exercise 23: Use annual data on the real price-earnings ratio from the file `pe.wf1` (source: `http://www.econ.yale.edu/~shiller/data/chapt26.xls`). Test the series for a unit-root. Irrespective of the test results, fit stationary *and* non-stationary models to the series using data until 1995. Compute out-of-sample forecasts for the series using both types of models.

2.4 Diffusion models in discrete time

Several areas of finance make extensive use of stochastic processes in continuous time. However, data is only available in discrete time, and the empirical analysis has to be done in discrete time, too. In this section we focus on the relation between continuous and discrete time models.

Review 11:[102] A **geometric Brownian motion (GBM)** is defined as

$$dP_t = \mu P_t\, dt + \sigma P_t\, dW_t \qquad \frac{dP_t}{P_t} = \mu\, dt + \sigma\, dW_t,$$

where W_t is a Wiener process with the following properties:

1. $\Delta W_t{=}Z_t\sqrt{\Delta t}$ where $Z_t{\sim}\mathrm{N}(0,1)$ (standard normal)and $\Delta W_t{\sim}\mathrm{N}(0,\Delta t)$.
2. The changes over distinct (non-overlapping) intervals are *independent*[103].
3. $W_t{\sim}\mathrm{N}(0,t)$ if $W_0{=}0$.
4. W_t evolves in continuous time and has no jumps (no discontinuities). However, its sample paths are not smooth but rather erratic.
5. The *increments* of W_t can be viewed as the counterpart of a discrete time white-noise process (with mean zero and unit variance if $\Delta t{=}1$), and W_t corresponds to a discrete time random-walk.

A GBM is frequently used to describe stock prices and implies non-negativity of the price P_t. μ and σ can be viewed as mean and standard deviation of the *simple* return $R_t{=}dP_t/P_t$. This return is measured over an infinitely small time interval dt and is therefore called **instantaneous** return. The (instantaneous) expected return is given by

$$\mathrm{E}[dP_t/P_t] = \mathrm{E}[\mu\, dt + \sigma\, dW_t] = \mu\, dt.$$

The (instantaneous) variance is given by

$$\mathrm{V}[dP_t/P_t] = \mathrm{V}[\mu\, dt + \sigma\, dW_t] = \sigma^2\mathrm{V}[dW_t] = \sigma^2 dt.$$

Both mean and standard deviation are constant over time. μ and σ are usually measured in *annual terms*.

The **standard** or **arithmetic Brownian motion** defined as

$$dX_t = \mu\, dt + \sigma\, dW_t \qquad (X_{t+\Delta t}{-}X_t){\sim}\mathrm{N}(\mu\Delta t, \sigma^2\Delta t)$$

is not suitable to describe stock prices since X_t can become negative.

A process that is frequently used to model interest rates is the **Ornstein-Uhlenbeck process**

$$dX_t = \kappa(\mu - X_t)dt + \sigma\, dW_t.$$

This is an example of a mean reverting process. When X_t is above (below) μ it tends back to μ at a speed determined by the mean-reversion parameter $\kappa{>}0$. The **square root process**

$$dX_t = \kappa(\mu - X_t)dt + \sigma\sqrt{X_t}\, dW_t$$

[102] Campbell et al. (1997), p.341 or Baxter and Rennie (1996), p.44.

[103] Because of the normality assumption it is sufficient to require that changes are uncorrelated.

is also used to model interest rates. It has the advantage that X_t cannot become negative.

A very general process is the **Ito process**

$$dX_t = \mu(X_t, t)\, dt + \sigma(X_t, t)\, dW_t,$$

where mean and variance can be functions of X_t and t.

Review 12:[104] If X_t is an Ito process then **Ito's lemma** states that a function $G_t{=}f(X_t, t)$ can be described by the stochastic differential equation (SDE)

$$dG_t = \left(\mu(\cdot) f'_X + f'_t + \frac{1}{2}\sigma^2(\cdot) f''_X\right) dt + \sigma(\cdot) f'_X\, dW_t,$$

where

$$f'_X = \frac{\partial G_t}{\partial X_t} \quad f'_t = \frac{\partial G_t}{\partial t} \quad f''_X = \frac{\partial^2 G_t}{\partial X_t^2}.$$

Example 43: Suppose the stock price P_t follows a GBM. We are interested in the process for the *logarithm* of the stock price. We have

$$G_t = \ln P_t \quad \mu(\cdot) = \mu P_t \quad \sigma(\cdot) = \sigma P_t \quad \frac{\partial G_t}{\partial P_t} = \frac{1}{P_t} \quad \frac{\partial G_t}{\partial t} = 0 \quad \frac{\partial^2 G_t}{\partial P_t^2} = -\frac{1}{P_t^2}.$$

Applying Ito's lemma we obtain

$$d \ln P_t = \left(\mu P_t \frac{1}{P_t} - 0.5\sigma^2 P_t^2 \frac{1}{P_t^2}\right) dt + \sigma P_t \frac{1}{P_t}\, dW_t,$$

$$d \ln P_t = (\mu - 0.5\sigma^2) dt + \sigma\, dW_t.$$

Thus, the log stock price $\ln P_t$ is an arithmetic Brownian motion with drift $\mu{-}0.5\sigma^2$, if P_t is a GBM with drift μ.

[104] Tsay (2002), p.226.

2.4.1 Discrete time approximation

We first consider the discrete time approximation of a continuous time stochastic process in the interval $[t_0, T]$. We choose n equidistant time points $t_i{=}t_0{+}i\Delta t$ $(i{=}1,\ldots,n)$, where $\Delta t{=}(T{-}t_0)/n{=}t_{i+1}{-}t_i$. The so-called **Euler approximation** of an Ito process is given by

$$X_{i+1} = X_i + \mu(\cdot)(t_{i+1} - t_i) + \sigma(\cdot)(W_{i+1} - W_i),$$

where X_i is the discrete time approximation of X_t at $t{=}t_i$. Equivalently, this could be written as $X_{i+1}{=}X_i{+}\Delta X_i$ where

$$\Delta X_i = X_{t_i+\Delta t} - X_{t_i} = \mu(\cdot)\Delta t + \sigma(\cdot)\Delta W_{t_i}.$$

Example 44: The Euler approximation of a GBM is given by[105]

$$\Delta P_i = \mu P_i \Delta t + \sigma P_i \Delta W_i \qquad \frac{\Delta P_i}{P_i} = R_i(\Delta t) = \mu\Delta t + \sigma\Delta W_i.$$

2.4.2 Estimating parameters

We now assume that the stock price follows a GBM and consider the SDE of the logarithm of the stock price. From example 43 we know that

$$d\ln P_t = (\mu - 0.5\sigma^2)dt + \sigma\, dW_t.$$

For a discrete time interval Δt the corresponding log return process in discrete time is given by (see Gourieroux and Jasiak, 2001, p.287)[106]

$$\ln P_{t+\Delta t} - \ln P_t = Y_t(\Delta t) = (\mu - 0.5\sigma^2)\Delta t + \sigma Z_t\sqrt{\Delta t}. \qquad Z_t \sim \mathrm{N}(0,1). \tag{47}$$

Suppose we have $n{+}1$ observations of a stock price p_t $(t{=}0,\ldots,n)$ sampled in discrete time. We want to use this sample to estimate the parameters μ and σ^2 of the underlying GBM in *annual terms*. We further suppose that log returns $y_t{=}\ln(p_t/p_{t-1})$ are i.i.d. normal.

Several things should be taken into account when comparing continuous and discrete time models:

1. In section 2.1 we have used the symbol μ to denote the mean of *log* returns (Y_t). However, to follow the notation typically used in diffusion models, in the present section μ is the mean of the corresponding *simple* return R_t.

2. Time series analysis in discrete time usually does not explicitly specify Δt. The corresponding discrete time model would use $\Delta t{=}1$ (i.e. use intervals of *one* day, *one* week, ...).

[105] Simulated sample paths of a GBM can be found in the file `gbm.xls`.
[106] It is understood that t is a discrete point in time t_i but we suppress the index i.

3. A discrete time series model for i.i.d. *log* returns would be formulated as

$$y_t = \bar{y} + e_t \qquad e_t \sim \mathrm{N}(0, s_e^2),$$

where $\bar{y}$ corresponds to $(\mu - 0.5\sigma^2)\Delta t$, and s_e^2 (or s_y^2) to $\sigma^2 \Delta t$. To estimate the GBM parameters μ and σ (which are usually given in *annual terms*) the observation frequency of y_t (which corresponds to Δt) has to be taken into account. We suppose that the time interval between t and $t-1$ is Δt and is measured in years (e.g. $\Delta t = 1/52$ for weekly data).

4. $d \ln P_t$ can be interpreted as the instantaneous log return of P_t. The (instantaneous) mean of the log return $d \ln P_t$ is $\mu - 0.5\sigma^2$. However, when we compare equations (41), p.91 and (47) we find a discrepancy. The mean of log returns Y_t in section 2.1.2 is given by $\ln(1+m) - 0.5\sigma_Y^2$ whereas the mean of $Y_t(\Delta t)$ is given by $(\mu - 0.5\sigma^2)\Delta t$. This can be explained by the fact that $\ln(1 + m\Delta t) \to m\, dt$ as $\Delta t \to dt$.

The sample estimates from log returns ($\bar{y}$ and s^2) correspond to $(\mu - 0.5\sigma^2)\Delta t$ and $\sigma^2 \Delta t$, respectively. Thus estimates of μ and σ^2 are given by

$$\hat{\sigma}^2 = s^2/\Delta t \qquad \hat{\mu} = \frac{\bar{y}}{\Delta t} + 0.5\hat{\sigma}^2 = \frac{\bar{y}}{\Delta t} + 0.5\frac{s^2}{\Delta t}.$$

Gourieroux and Jasiak (2001, p.289) show that the asymptotic variance of $\hat{\sigma}^2$ and $\hat{\mu}$ is given by

$$\mathrm{aV}[\hat{\sigma}^2] = \frac{2\sigma^4}{n} \qquad \mathrm{aV}[\hat{\mu}] = \frac{\sigma^2}{n\Delta t} + \frac{\sigma^4}{2n}.$$

By increasing the sampling frequency more observations become available (n increases), but Δt becomes accordingly smaller. The net effect is that $n\Delta t$ stays constant, the first term in the definition of $\mathrm{aV}[\hat{\mu}]$ does not become smaller as n increases, and the drift cannot be consistently estimated.

Example 45: In example 31 the mean FTSE log return $\bar{y}$ estimated from monthly data was 0.00765 and the standard deviation s was 0.065256. The estimated mean and variance of the underlying GBM in annual terms are given by

$$\hat{\sigma}^2 = 0.065256^2 \cdot 12 = 0.0511 \qquad \hat{\mu} = 0.00765 \cdot 12 + 0.5 \cdot 0.0511 = 0.117346.$$

We now consider estimating the parameters of the Ornstein-Uhlenbeck process using a discrete time series. A *simplified* discrete time version of the process can be written as

$$X_t - X_{t-\Delta t} = \kappa\mu\Delta t - \kappa\Delta t X_{t-\Delta t} + \sigma Z_t \sqrt{\Delta t}$$

$$X_t = \kappa\mu\Delta t + (1 - \kappa\Delta t) X_{t-\Delta t} + \sigma Z_t \sqrt{\Delta t}.$$

This is equivalent to an AR(1) model (using the notation from section 2.2)

$$X_t = \nu + \phi_1 X_{t-1} + \epsilon_t,$$

where ν corresponds to $\kappa\mu\Delta t$, ϕ_1 to $(1-\kappa\Delta t)$, and σ_ϵ^2 to $\sigma^2\Delta t$. The Ornstein-Uhlenbeck process is only mean reverting (or stationary) if $\kappa>0$. This corresponds to the condition $|\phi_1|<1$ for AR(1) models. Thus it is useful to carry out a unit-root test before the parameters κ and μ are estimated.

Given an observed series x_t we can fit the AR(1) model

$$x_t = c + f_1 x_{t-1} + e_t$$

and use the estimates c, f_1 and s_e to estimate κ, μ and σ (*in annual terms*):

$$\hat{\kappa} = \frac{1-f_1}{\Delta t} \qquad \hat{\mu} = \frac{c}{\hat{\kappa}\Delta t} = \frac{c}{1-f_1} \qquad \hat{\sigma} = \frac{s_e}{\sqrt{\Delta t}}.$$

Since estimated AR coefficients are biased[107] downwards in small samples, $\hat{\kappa}$ will be biased upwards.

A *precise* discrete time formulation is given by (see Gourieroux and Jasiak, 2001, p.289)

$$X_t = \mu(1-\exp\{-\kappa\Delta t\}) + \exp\{-\kappa\Delta t\}X_{t-\Delta t} + \sigma\eta Z_t\sqrt{\Delta t},$$

where

$$\eta = \left[\frac{1-\exp\{-2\kappa\Delta t\}}{2\kappa\Delta t}\right]^{1/2}.$$

Using this formulation the parameters are estimated by

$$\hat{\kappa} = \frac{-\ln f_1}{\Delta t} \qquad \hat{\mu} = \frac{c\exp\{\hat{\kappa}\Delta t\}}{\exp\{\hat{\kappa}\Delta t\}-1} \qquad \hat{\sigma} = \frac{s_e}{\sqrt{\Delta t}}\left[\frac{1-\exp\{-2\hat{\kappa}\Delta t\}}{2\hat{\kappa}\Delta t}\right]^{1/2}.$$

Note that f_1 has to be positive in this case.

Example 46: In example 40 we have found that the spread between long- and short-term interest rates in the UK is stationary (or mean reverting). We assume that the spread follows a Ornstein-Uhlenbeck process. The estimated AR(1) model using quarterly data is[108]

$$x_t = 0.1764 + 0.8563x_{t-1} + e_t \qquad s_e = 0.8696$$

which yields the following estimates in annual terms (Δt=1/4):

$$\hat{\kappa} = \frac{1-0.8563}{\Delta t} = 0.575 \quad \hat{\mu} = \frac{0.1764}{0.575\Delta t} = 1.227 \quad \hat{\sigma} = \frac{0.8696}{\sqrt{\Delta t}} = 1.74.$$

Using the precise formulation we obtain

$$\hat{\kappa} = \frac{-\ln 0.8563}{\Delta t} = 0.62 \qquad \hat{\mu} = \frac{c\exp\{0.62\Delta t\}}{\exp\{0.62\Delta t\}-1} = 1.227$$

$$\hat{\sigma} = \frac{s_e}{\sqrt{\Delta t}}\left[\frac{1-\exp\{-2\cdot 0.62\Delta t\}}{2\cdot 0.62\Delta t}\right]^{1/2} = 1.613.$$

[107] The bias increases as the AR parameter ϕ approaches one, or as the mean reversion parameter κ approaches zero.

[108] Details can be found in the file `ornstein-uhlenbeck.xls`.

2.4.3 Probability statements about future prices

We now focus on longer time intervals and consider price changes over T periods (e.g. 30 days). The T-period log return is the change in log prices between t and $t+T$. Thus the log return is normally distributed[109] with mean and variance

$$\mathrm{E}[\ln P_{t+T}] - \ln P_t = \mathrm{E}[Y_t(T)] = (\mu - 0.5\sigma^2)T \qquad \mathrm{V}[Y_t(T)] = \sigma^2 T.$$

Equivalently, P_{t+T} is lognormal and $\ln P_{t+T}$ is normally distributed:

$$\ln P_{t+T} \sim \mathrm{N}(\ln P_t + (\mu - 0.5\sigma^2)T, \sigma^2 T).$$

Conditional on P_t the expected value of P_{t+T} is

$$\mathrm{E}[P_{t+T}|P_t] = P_t \exp\{\mu T\}.$$

The discrepancy between this formula and equation (45), p.118 used to forecast prices in section 2.3.2 can be reconciled by noting that here μ is the mean of *simple* returns. The corresponding discrete time series model for log returns y_t is

$$y_t = \bar{y} + e_t \qquad e_t \sim \mathrm{N}(0, s^2)$$

and the conditional expectation of p_{t+T} is

$$\mathrm{E}[p_{t+T}|p_t] = p_t \exp\{\bar{y}T + 0.5s^2 T\}.$$

A $(1-\alpha)$ confidence interval for the price in $t+T$ can be computed from the properties of T-period log returns. The boundaries of the interval for log returns are given by

$$(\mu - 0.5\sigma^2)T \pm |z_{\alpha/2}|\sigma\sqrt{T},$$

and the boundaries for the price P_{t+T} are given by[110]

$$P_t \exp\left\{(\mu - 0.5\sigma^2)T \pm |z_{\alpha/2}|\sigma\sqrt{T}\right\}.$$

Example 47: December 28, 1990 the value of the FTSE was 2160.4 (according to `finance.yahoo.com`). We use the estimated mean and variance from example 45 to compute a 95% confidence interval for the index in nine months (end of September 1991) and ten years (December 2000).[111]
Using $\hat{\sigma}^2$=0.05 and $\hat{\mu}$=0.117 the interval for T=0.75 is given by[112]

$$2160.4 \cdot \exp\left\{(0.117 - 0.5 \cdot 0.05)0.75 \pm 1.96\sqrt{0.05 \cdot 0.75}\right\} = [1584, 3383]$$

and for T=10

$$2160.4 \cdot \exp\left\{(0.117 - 0.5 \cdot 0.05)10 \pm 1.96\sqrt{0.05 \cdot 10}\right\} = [1356, 21676].$$

Note: the actual values of the FTSE were 2621.7 (September 30, 1991) and 6222.5 (December 29, 2000).

[109] The normal assumption for log returns cannot be justified empirically unless the observation frequency is low.
[110] Note that the bounds are *not* given by $\mathrm{E}[P_{t+T}] \pm |z_{\alpha/2}|\sqrt{\mathrm{V}[P_{t+T}]}$.
[111] Details can be found in the file `probability statements.xls`.
[112] We use rounded values of the estimates $\hat{\mu}$ and $\hat{\sigma}$.

We now consider probabilities like $\mathrm{P}[P_{t+T} \leq K]$, where K is a pre-specified, non-stochastic value (e.g. the strike price in option pricing).

Given that log returns over T periods are normally distributed $Y_t(T) \sim \mathrm{N}((\mu - 0.5\sigma^2)T, \sigma^2 T)$, probability statements about P_{t+T} can be based on the properties of $Y_t(T)$:

$$\mathrm{P}[P_t \exp\{Y_t(T)\} \leq K] = \mathrm{P}[P_{t+T} \leq K] = \mathrm{P}[Y_t(T) \leq \ln(K/P_t)].$$

For instance, the probability that the price in $t+T$ is less than K is given by

$$\mathrm{P}[Y_t(T) \leq \ln(K/P_t)] = \Phi\left(\frac{\ln(K/P_t) - (\mu - 0.5\sigma^2)T}{\sigma\sqrt{T}}\right).$$

Similar probabilities are used in the Black-Scholes option pricing formula, and can be used in a *heuristic* derivation of that formula[113].

Example 48: We use the information from example 47 to compute the probability that the FTSE will be below K=2000 in September 1991.

$$\mathrm{P}[P_{t+T} \leq K] = \Phi\left(\frac{\ln(2000/2160.4) - (0.117 - 0.5 \cdot 0.05)0.75}{\sqrt{0.05 \cdot 0.75}}\right) = 0.225.$$

Exercise 24:

1. Use a time series from exercise 17 (stock price, index or exchange rate). Assume that this series follows a GBM and estimate the parameters μ and σ (in annual terms).
2. Select a time series that appears to be mean-reverting. Verify this assumption by a unit-root test. Assume that this series follows a Ornstein-Uhlenbeck process and estimate the parameters κ, μ and σ.

[113] For details see Jarrow and Rudd (1983), p.90.

2.5 GARCH models

For many problems in finance the variance or volatility of returns is a parameter of central importance. It can serve as a risk measure, it is necessary for portfolio selection, it is required in option pricing, and in the context of value-at-risk.

The time series models from section 2.2 can be used to replace the unconditional mean by a conditional mean (i.e. the sample mean $\bar{y}$ is replaced by $\hat{y}_t$). Similarly, the purpose of modelling the variance is to replace the unconditional sample estimate s^2 by a conditional estimate s_t^2. Given that the volatility of returns is typically not constant over time, the conditional variance s_t^2 should be a better variance estimate or forecast than s^2.

The variance of a GARCH process is not constant over time (heteroscedastic), and its conditional variance follows a generalized AR model (see below). The acronym GARCH stands for 'generalized autoregressive conditional heteroscedasticity'. A GARCH model always consists of two equations:

1. The equation for the conditional mean has the following general form:

$$Y_t = \hat{Y}_{t-1,1} + \epsilon_t = \mathrm{E}[Y_t|I_{t-1}] + \epsilon_t \qquad \epsilon_t \ldots \text{white-noise.}$$

   $\hat{Y}_{t-1,1}$ is the conditional expectation (or the one-step ahead forecast) of Y_t derived from a time series or regression model. I_{t-1} is the information set available at time $t-1$. If Y_t is white-noise $\hat{Y}_{t-1,1}=\mu$.

   In a GARCH model the variance of the disturbance term ϵ_t is not constant but the *conditional* variance is time-varying:

$$\mathrm{E}[(Y_t - \hat{Y}_{t-1,1})^2|I_{t-1}] = \sigma_t^2.$$

   What we need is a model that determines how σ_t^2 evolves over time.

2. In a GARCH(1,1) model the time variation of the conditional variance is given by

$$\begin{aligned}\sigma_t^2 &= \omega_0 + \omega_1(Y_{t-1} - \hat{Y}_{t-2,1})^2 + \lambda_1\sigma_{t-1}^2 \\ &= \omega_0 + \omega_1\epsilon_{t-1}^2 + \lambda_1\sigma_{t-1}^2 \qquad \omega_0, \omega_1, \lambda_1 \geq 0,\ (\omega_1 + \lambda_1) < 1.\end{aligned}$$

   It is frequently assumed but not necessary that the conditional distribution of ϵ_t is normal: $\epsilon_t|I_{t-1} \sim \mathrm{N}(0, \sigma_t^2)$.

The conditional variance in t is based on 'news' or 'shocks' (i.e. forecast errors ϵ_t) introduced by the term $\omega_1\epsilon_{t-1}^2$. In addition, the variance in t is based on the conditional variance of the previous period weighted by λ_1. ω_1 determines the immediate (but lagged) response to shocks and λ_1 determines the duration of the effect. If λ_1 is much greater than ω_1, σ_t^2 decays very slowly after extraordinary events (large ϵ_t).

The coefficients ω_0, ω_1 and λ_1 also determine the *average* level of σ_t^2 which is identical to the *unconditional* variance of ϵ_t:

$$\sigma_\epsilon^2 = \frac{\omega_0}{1 - \omega_1 - \lambda_1}. \tag{48}$$

Note that conditional (and unconditional) variance of Y_t and ϵ_t are only identical if the conditional mean of Y_t is constant ($\hat{Y}_{t-1,1}=\mu$).

GARCH models account for two well documented features of financial returns:

1. *Volatility clustering (heteroscedasticity):* Suppose a (relatively) large value of ϵ_t occurs. This leads to an increase in σ_t^2 in the following period. Thus, the conditional distribution of returns in the subsequent period(s) has a higher variance. This makes further large disturbances more likely. As a result, a phase with approximately the same level of volatility – a volatility cluster – is formed. If ω_1 is greater than λ_1, the conditional variance returns very quickly to a lower level and the degree of the volatility clustering is small.

2. *Non-normality:* The kurtosis of a GARCH(1,1) model is given by[114]

$$\frac{\mathrm{E}[\epsilon_t^4]}{\mathrm{V}[\epsilon_t]^2} = \frac{3[1-(\omega_1+\lambda_1)^2]}{1-(\omega_1+\lambda_1)^2-2\omega_1^2} > 3.$$

   Thus, GARCH models can account for fat tails. Although the unconditional moments implied by a GARCH model can be determined, the unconditional GARCH distribution is not known analytically, even when the conditional distribution is normal.

If a time series or regression model has heteroscedastic or non-normal residuals the standard errors of estimated parameters (and p-values) are biased (see section 1.7). Since GARCH models can account for both problems, adding a GARCH equation to a model for the conditional mean may lead to choose a different ARMA model or different explanatory variables in a regression model.

The GARCH(p,q) model is a generalization of the GARCH(1,1) model where q past values of ϵ_t^2 and p past values of σ_t^2 are used:

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{q} \omega_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \lambda_i \sigma_{t-i}^2.$$

Many empirical investigations found that GARCH(1,1) models are sufficient (see e.g. Bollerslev et al., 1992).

[114] For details see Tsay (2002) p.118.

2.5.1 Estimating and diagnostic checking of GARCH models

GARCH models cannot be estimated with least squares because the variance cannot be observed directly. Thus, the difference between 'actual' and fitted variance cannot be computed. GARCH models can be estimated by maximum-likelihood.[115] To estimate a GARCH model we need (a) a model for the conditional mean $\hat{y}_t$ to determine the residuals ϵ_t=y_t−$\hat{y}_t$; (b) a conditional distribution for the residuals, and (c) a model for the conditional variance of ϵ_t. There exists no well established strategy for selecting the order of a GARCH model (similar to the ARMA model building strategy). The choice cannot be based on (partial) autocorrelations of squared returns or residuals. A simple model building strategy starts with a GARCH(1,1) model, adds further lags to the variance equation, and uses AIC or SC to select a final model.

If we assume a conditional normal distribution for the residuals $\epsilon_t|I_{t-1}\sim$N(0,σ_t^2), the log-likelihood function is given by

$$\ell = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{n}\ln\sigma_t^2 - \frac{1}{2}\sum_{t=1}^{n}\frac{\epsilon_t^2}{\sigma_t^2},$$

where σ_t^2 can be defined in terms of a GARCH(p,q) model. The log-likelihood is a straightforward extension of equation (15) in section 1.4. It is obtained by replacing the constant variance σ^2 by the conditional variance σ_t^2 from the GARCH equation.

Diagnostic checking of a GARCH model is based on the standardized residuals $\tilde{e}_t$=e_t/s_t. The GARCH model is adequate if $\tilde{e}_t$, $\tilde{e}_t^2$ and $|\tilde{e}_t|$ are white-noise, and $\tilde{e}_t$ is normal.

2.5.2 Example 49: ARMA-GARCH models for IBM and FTSE returns

In example 33 we found that IBM log returns are white-noise and example 34 indicated heteroscedasticity of returns. Therefore we estimate a GARCH(1,1) model with constant mean:

$$y_t = \underset{(0.75)}{0.0002} + e_t \qquad s_t^2 = \underset{(0.002)}{9.6\cdot 10^{-6}} + \underset{(0.0)}{0.27}\, e_{t-1}^2 + \underset{(0.0)}{0.72}\, s_{t-1}^2.$$

However, $\tilde{e}_t$ is not white-noise (r_1=0.138 and Q_1=0.008). Therefore we add AR(1) and MA(1) terms to the conditional mean equation. AIC and SC select the following model:

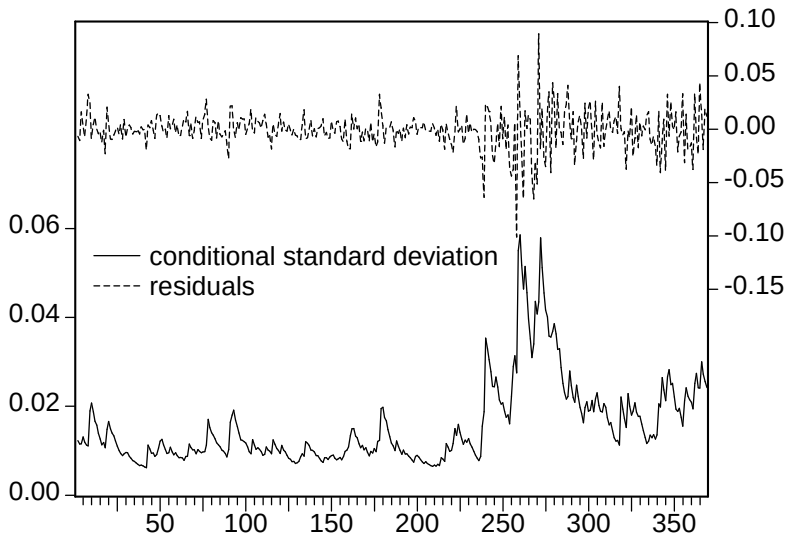
$$y_t = \underset{(0.67)}{0.0003} + \underset{(0.078)}{0.1}\; e_{t-1} + e_t \qquad s_t^2 = \underset{(0.003)}{7.8\cdot 10^{-6}} + \underset{(0.0)}{0.24}\, e_{t-1}^2 + \underset{(0.0)}{0.75}\, s_{t-1}^2.$$

Adding the term e_{t-2}^2 to the variance equation is supported by AIC but not by SC. The standardized residuals and their squares are white-noise (the p-values of Q_5 are 0.75 and 0.355, respectively). The JB-test rejects normality but the kurtosis is only 4.06 (the skewness is –0.19), whereas skewness and kurtosis of observed returns are –0.6 and 8.2. We conclude that the GARCH model explains a lot of the non-normality of IBM log returns.

The conditional standard deviation s_t from this model captures the changes in the volatility of residuals e_t very well (see Figure 9). The conditional mean is very close to the

[115] An example can be found in the file `AR-GARCH ML estimation.xls`.

Figure 9: Conditional standard deviation and residuals from a MA(1)-GARCH(1,1) model for IBM log returns.



unconditional mean, and thus residuals and returns are almost equal (compare the returns in Figure 2 to the residuals in Figure 9).

We extend the MA(2) model for FTSE log returns from example 2.2.6 and fit the MA(2)-GARCH model

$$\underset{}{y_t} = \underset{(0.021)}{0.0074} + \underset{(0.4)}{0.06}\, e_{t-1} - \underset{(0.01)}{0.15}\, e_{t-2} + e_t$$

$$\underset{}{s_t^2} = \underset{(0.11)}{0.0003} + \underset{(0.016)}{0.099}\, e_{t-1}^2 + \underset{(0.0)}{0.82}\, s_{t-1}^2.$$

The p-values of the MA coefficients have changed compared to example 2.2.6. The first MA parameter h_1 is clearly insignificant and could be removed from the mean equation. In example 2.2.6 we found that MA residuals were not normal and not homoscedastic. Since p-values are biased in this case, we expect that adding a GARCH equation which accounts for non-normality and heteroscedasticity should affect the p-values.

The standardized residuals of the MA-GARCH model are white-noise and homoscedastic but not normal. If the conditional normal assumption does not turn out to be adequate a different conditional distribution has to be used (e.g. a t-distribution).

Exercise 25: Use the ARMA models from exercise 20, estimate ARMA-GARCH models, and carry out diagnostic checking.

2.5.3 Forecasting with GARCH models

GARCH models can be used to determine static and dynamic variance forecasts of a time series. The GARCH(1,1) forecasting equation for future dates $t+\tau$ is

$$\begin{aligned}\sigma^2_{t,1} &= \omega_0 + \omega_1\epsilon^2_t + \lambda_1\sigma^2_t \\ \sigma^2_{t,2} &= \omega_0 + \omega_1\epsilon^2_{t+1} + \lambda_1\sigma^2_{t,1} \\ &= \omega_0 + \omega_1\epsilon^2_{t+1} + \lambda_1(\omega_0 + \omega_1\epsilon^2_t + \lambda_1\sigma^2_t).\end{aligned}$$

The unknown future value ϵ^2_{t+1} in this equation is replaced by the conditional expectation $\mathrm{E}[\epsilon^2_{t+1}|I_t]=\sigma^2_{t,1}$:

$$\sigma^2_{t,2} = \omega_0 + \omega_1\sigma^2_{t,1} + \lambda_1(\omega_0 + \omega_1\epsilon^2_t + \lambda_1\sigma^2_t) = \omega_0 + (\omega_1 + \lambda_1)\sigma^2_{t,1}.$$

Thus, the variance for $t+2$ can be determined on the basis of ϵ_t and σ^2_t. The same procedure can be applied recursively to obtain forecasts for any τ

$$\sigma^2_{t+\tau} = \omega_0 + (\omega_1 + \lambda_1)\sigma^2_{t+\tau-1}.$$

For increasing τ the forecasts $\sigma^2_{t,\tau}$ converge to the unconditional variance σ^2 from equation (48), provided $(\omega_1+\lambda_1)<1$. The time until the level of the unconditional variance is reached depends on the GARCH parameters, the value of the last residual in the sample, and the difference between the unconditional variance and the conditional variance in t (when the forecast is made).

We finally note that, in general, the variance of h-period returns $y_t(h)$ estimated from a GARCH model will differ from the (frequently used) unconditional estimate $h\sigma^2$ which is based on homoscedastic returns. The h-period variance is given by the sum

$$\sigma^2(h) = \sum_{\tau=1}^{h} \sigma^2_{t,\tau},$$

which also depends on the current level σ^2_t.

2.5.4 Special GARCH models

In empirical studies it is usually found that $\omega_1+\lambda_1$ is close to one. For instance, in the models estimated in example 49 we found $\omega_1+\lambda_1$=0.99 and 0.92. The sum of the GARCH parameters ($\omega_1+\omega_2+\cdots+\lambda_1+\lambda_2+\cdots$) can be used as a measure of **persistence** in variance. Persistence implies that the conditional variance tends to remain at a particular (high or low) level. This tendency increases with the level of persistence. If persistence is high this leads to volatility clustering.

The **integrated I-GARCH** model is a special case of a GARCH model with the constraint that $\omega_1+\lambda_1$=1. This saves one parameter to be estimated. The forecasts from a I-GARCH model are given by

$$\sigma^2_{t+\tau} = \tau\omega_0 + \sigma^2_t.$$

A further special case is the **exponentially weighted moving average (EWMA)** where ω_0=0 and $\omega_1+\lambda_1$=1, and only one parameter λ is required:

$$\sigma^2_t = (1-\lambda)\epsilon^2_{t-1} + \lambda\sigma^2_{t-1}.$$

The EWMA model is used by RiskMetrics for value-at-risk[116] calculations. In this context the parameter λ is not estimated. RiskMetrics recommends to use values around 0.95.

> **Example 50:** Figure 10 shows the estimated in-sample (until end of 1987) and out of-sample (starting 1988) variance of FTSE log returns from the GARCH(1,1) model
>
> $$\underset{}{s_t} = \underset{(0.13)}{0.0003} + \underset{(0.01)}{0.115}\,(y_{t-1} - \underset{(0.05)}{0.0079})^2 + \underset{(0.0)}{0.83}\,s^2_{t-1}$$
>
> with constant mean return. The EWMA variance[117] using λ=0.95 is shown for comparison. The dynamic forecasts converge to the unconditional variance based on the estimated parameters ($\omega_0/(1-\omega_1-\lambda_1)$=0.0003/(1−0.115−0.83)=0.0055). During the in-sample period EWMA and GARCH variance behave similarly. Differences in the decay after large shocks are due to the difference between λ_1=0.83 and λ=0.95.

GARCH models can be extended in various ways, and numerous formulations of the variance equation exist. In the **threshold ARCH (TARCH)** model, for instance, asymmetric effects of news on the variance can be taken into account. In this case the variance equation has the following form:

$$\sigma^2_t = \omega_0 + \omega_1\epsilon^2_{t-1} + \gamma_1\epsilon^2_{t-1}d_{t-1} + \lambda_1\sigma^2_{t-1},$$

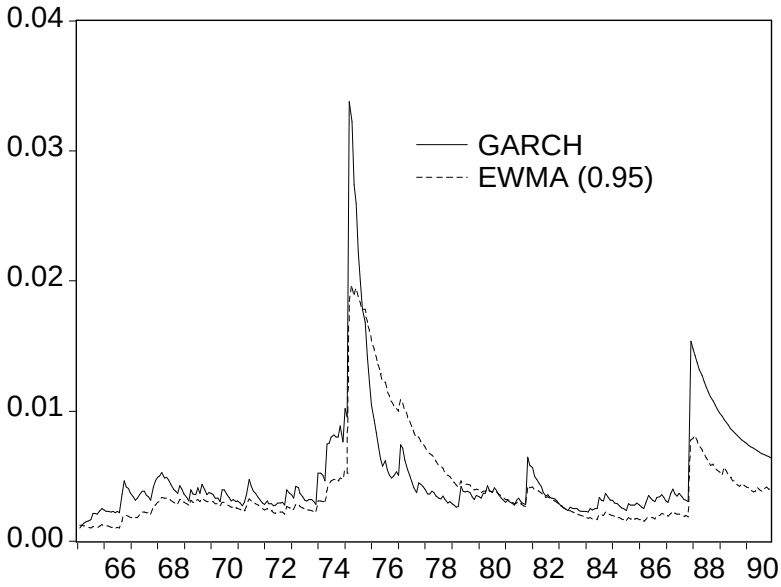
where, d_t=1 (d_t=0) if ϵ_t<0 ($\epsilon_t\geq$0). If γ_1>0 negative disturbances have a stronger effect on the variance than positive ones. The **exponential GARCH (EGARCH)** model also allows for modelling asymmetric effects. It is formulated in terms of the logarithm of σ^2_t:

$$\ln\sigma^2_t = \omega_0 + \gamma_1\frac{\epsilon_{t-1}}{\sigma_{t-1}} + \delta_1\frac{|\epsilon_{t-1}|}{\sigma_{t-1}} + \lambda_1\ln\sigma^2_{t-1}.$$

[116] http://www.riskmetrics.com/mrdocs.html.

[117] The EWMA variance during the out-of-sample period is based on observed returns, while the dynamic GARCH variance forecasts do not use any data at all from that period.

Figure 10: GARCH and EWMA estimates and forecasts of the variance of FTSE log returns.



If $\epsilon_{t-1}<0$ ($\epsilon_{t-1}>0$) the total impact of $\epsilon_{t-1}/\sigma_{t-1}$ on the conditional (log) variance is given by $\gamma_1-\delta_1$ ($\gamma_1+\delta_1$). If bad news have a stronger effect on volatility the expected signs are $\gamma_1+\delta_1>0$ and $\gamma_1<0$.

As a further extension, explanatory variables can be included in the variance equation. Some empirical investigations show that the number or the volume of trades have a significant effect on the conditional variance (see Lamoureux and Lastrapes, 1990). After including such explanatory variables the GARCH parameters frequently become smaller or insignificant.

In the **GARCH-in-the-mean (GARCH-M)** model the conditional variance or standard deviation is used as an explanatory variable in the equation for the conditional mean:

$$Y_t = \nu + \delta\sigma_t^2 + \epsilon_t,$$

where any GARCH model can be specified for σ_t^2. A significant parameter δ would support the hypothesis that expected returns of an asset contain a risk premium that is proportional to the variance (or standard deviation) of *that* asset's returns. However, according to financial theory (e.g. the CAPM) the risk premium of an asset has to be determined in the context of a portfolio of many assets.

Exercise 26: Use the log returns defined in exercise 17 and estimate a TARCH model to test for asymmetry in the conditional variance.

Obtain a daily financial time series from finance.yahoo.com and retrieve the trading volume, too. Add volume as explanatory variable to the GARCH equation. Hint: Rescale the volume series (e.g. divide by 10^6 or a greater number), and/or divide by the price or index to convert volume into number of trades.

Use the log returns defined in exercise 17 and estimate a GARCH-M model.

3 Vector time series models

3.1 Vector-autoregressive models

3.1.1 Formulation of VAR models

Multivariate time series analysis deals with more than one series and accounts for feedback among the series. The models can be viewed as extensions or generalizations of univariate ARMA models. A basic model of multivariate analysis is the **vector-autoregressive (VAR)** model.[118]

VAR models have their origin mainly in macroeconomic modeling, where simultaneous (structural) equation models developed in the fifties and sixties turned out to have inferior forecasting performance. There were also concerns about the validity of the theories underlying the structural models. Simple, small-scale VAR models were found to provide suitable tools for analyzing the impacts of policy changes or external shocks. VAR models are mainly applied in the context of Granger causality tests and impulse-response analyses (see Greene, 2003, p.592). In addition, they are the basis for vector error correction models (see section 3.2).

The **standard form** or **reduced form** of a first order VAR model – VAR(1) – for two processes Y_t and X_t is given by

$$Y_t = \nu_y + \phi_{yy} Y_{t-1} + \phi_{yx} X_{t-1} + \epsilon_t^y$$

$$X_t = \nu_x + \phi_{xy} Y_{t-1} + \phi_{xx} X_{t-1} + \epsilon_t^x,$$

where ϵ_t^y and ϵ_t^x are white-noise disturbances which may be correlated. A VAR(1) process can be written in matrix form as

$$\mathbf{Y}_t = \boldsymbol{V} + \boldsymbol{\Phi}_1 \mathbf{Y}_{t-1} + \boldsymbol{\epsilon}_t \qquad \epsilon_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon),$$

where $\mathbf{Y}_t$ is a column vector which contains all k series in the model. $\boldsymbol{V}$ is a vector of constants. $\boldsymbol{\Phi}_1$ is a $k \times k$ matrix containing the autoregressive coefficients for lag 1. $\boldsymbol{\epsilon}_t$ is a column vector of disturbance terms assumed to be normally distributed with covariance $\boldsymbol{\Sigma}_\epsilon$. In the two-variable VAR(1) model formulated above $\mathbf{Y}_t$, $\boldsymbol{V}$, $\boldsymbol{\Phi}_1$ and $\boldsymbol{\epsilon}_t$ are given by

$$\mathbf{Y}_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} \nu_y \\ \nu_x \end{bmatrix} \qquad \boldsymbol{\Phi}_1 = \begin{bmatrix} \phi_{yy} & \phi_{yx} \\ \phi_{xy} & \phi_{xx} \end{bmatrix} \qquad \boldsymbol{\epsilon}_t = \begin{bmatrix} \epsilon_t^y \\ \epsilon_t^x \end{bmatrix}.$$

$\boldsymbol{\Sigma}_\epsilon$ is related to the correlation matrix of disturbances $\mathbf{C}_\epsilon$ and the vector of standard errors $\boldsymbol{\sigma}_\epsilon$ by $\boldsymbol{\Sigma}_\epsilon = \mathbf{C}_\epsilon \cdot (\boldsymbol{\sigma}_\epsilon \boldsymbol{\sigma}_\epsilon')$.

The moving average (MA) representation of a VAR(1) model exists, if the VAR process is stationary. This requires that all eigenvalues of $\boldsymbol{\Phi}_1$ have modulus less than one (see Lütkepohl, 1993, p.10). In this case

$$\mathbf{Y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Phi}_1^i \boldsymbol{\epsilon}_{t-i} = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Theta}_i \boldsymbol{\epsilon}_{t-i},$$

[118] The general case of vector ARMA models will not be presented in this text; see Tsay (2002), p.322 for details.

where $\mathbf{\Phi}_1^i$ denotes the matrix power of order i, $\mathbf{\Theta}_i$ is the MA coefficient matrix for lag i, and $\boldsymbol{\mu}=(\mathbf{I}-\mathbf{\Phi}_1)^{-1}\boldsymbol{V}$. The autocovariance of $\mathbf{Y}_t$ for lag ℓ is given by

$$\sum_{i=0}^{\infty} \mathbf{\Phi}_1^{\ell+i} \mathbf{\Sigma}_\epsilon (\mathbf{\Phi}_1^i)' = \sum_{i=0}^{\infty} \mathbf{\Theta}_{\ell+i} \mathbf{\Sigma}_\epsilon \mathbf{\Theta}_i'.$$

Extensions to higher order VAR models are possible (see Lütkepohl, 1993, p.11).

The VAR model in standard form only contains lagged variables on the right hand side. This raises the question whether and how *contemporaneous* dependencies between Y_t and X_t are taken into account. To answer this question we consider the following example:

$$Y_t = \omega_0 X_t + \omega_1 X_{t-1} + \delta_1 Y_{t-1} + U_t$$

$$X_t = \phi_1 X_{t-1} + W_t.$$

These equations can be formulated as a VAR(1) model in **structural form**[119]:

$$\begin{bmatrix} 1 & -\omega_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \delta_1 & \omega_1 \\ 0 & \phi_1 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} U_t \\ W_t \end{bmatrix}.$$

The structural form may include contemporaneous relations represented by the coefficient matrix on the left side of the equation. Substituting X_t from the second equation into the first equation yields

$$Y_t = (\omega_0\phi_1 + \omega_1)X_{t-1} + \delta_1 Y_{t-1} + \omega_0 W_t + U_t,$$

or in matrix form:

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \delta_1 & \omega_0\phi_1 + \omega_1 \\ 0 & \phi_1 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \omega_0 W_t + U_t \\ W_t \end{bmatrix}.$$

Formulating this VAR(1) model in reduced form

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \phi_{yy} & \phi_{yx} \\ \phi_{xy} & \phi_{xx} \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t^y \\ \epsilon_t^x \end{bmatrix}$$

yields the following identities:

$$\phi_{yy} = \delta_1 \quad \phi_{yx} = (\omega_0\phi_1 + \omega_1) \quad \phi_{xy} = 0 \quad \phi_{xx} = \phi_1$$

$$\sigma^2_{\epsilon^y} = \omega_0^2\sigma_W^2 + \sigma_U^2 \quad \sigma^2_{\epsilon^x} = \sigma_W^2 \quad \operatorname{cov}[\epsilon_t^y \epsilon_t^x] = \omega_0\sigma_W^2.$$

Thus, if Y_t and X_t are contemporaneously related the disturbance terms ϵ_t^y and ϵ_t^x of the reduced form are correlated. This correlation depends on the parameter ω_0 in the structural equation. In example 51 the correlation between the residuals is 0.41, which can be used to estimate the parameter ω_0. In general, it is not possible to uniquely determine the parameters of the structural from the (estimated) parameters of a VAR model in reduced form. For this purpose, suitable assumptions about the dependencies in the structural form must be made.

[119] Note that appropriate estimation of structural forms depends on the specific formulation. For example, if Y_t also appeared as a regressor in the equation for X_t, separately estimating each equation would lead to inconsistent estimates because of the associated endogeneity (simultaneous equation bias). The same applies in the present formulation if U_t and W_t are correlated.

3.1.2 Estimating and forecasting VAR models

The joint estimation of two or more regression equations (system of equations) is beyond the scope of this text. In general, possible dependencies across equations need to be taken into account using GLS or ML. As a major advantage, VAR models in reduced form can be estimated by applying least-squares separately to each equation of the model. OLS yields consistent and asymptotically efficient estimates. None of the series in a VAR models is exogenous as defined in a regression context. A necessary condition is that the series are stationary (i.e. $\mathbf{AR}_t$ has to hold), and the residuals in each equation are white-noise. If the residuals are autocorrelated, additional lags are added to the model. The number of lags can also be selected on the basis of information criteria like AIC or SC. No precautions are necessary if the residuals are correlated across equations. Since a VAR model can be viewed as a seemingly unrelated regression (SUR) with identical regressors, OLS has the same properties as GLS (see Greene, 2003, p.343).

The VAR model should only include variables with the *same* order of integration. If the series are integrated the VAR model is fitted to (first) differences.[120] In section 3.2 we will present a test for integration of several series that can be interpreted as a multivariate version of the DF test.

Lags with insignificant coefficients are usually *not* eliminated from the VAR model. This may have a negative effect on the forecasts from VAR models since (in most cases) too many parameters are estimated. This inefficiency leads to an unnecessary increase in the variance of forecasts. However, if some coefficients are restricted to zero, least-square estimates are not efficient any more. In this case, the VAR model can be estimated by (constrained) maximum likelihood[121].

Figure 11: VAR(2) model for one-month (Y_1M) and five-year interest rates (Y_5Y).

Sample(adjusted): 1964:04 1993:12
Included observations: 357
t-statistics in parentheses

	D(Y_1M)	D(Y_5Y)
D(Y_1M(-1))	-0.198355	0.014609
	(-3.41811)	(0.46413)
D(Y_1M(-2))	0.010262	0.048413
	(0.18134)	(1.57724)
D(Y_5Y(-1))	0.624043	0.063369
	(5.81580)	(1.08884)
D(Y_5Y(-2))	-0.275744	-0.129845
	(-2.45457)	(-2.13101)
C	-0.003692	0.003407
	(-0.08909)	(0.15157)
R-squared	0.116266	0.018756
Adj. R-squared	0.106224	0.007605
S.E. equation	0.783066	0.424725
S.D. dependent	0.828293	0.426349
Akaike Information Criteria		3.317573
Schwarz Criteria		3.426193

[120] For a discussion of various alternatives see Hamilton (1994), p.651.
[121] For details see Hamilton (1994), p.315.

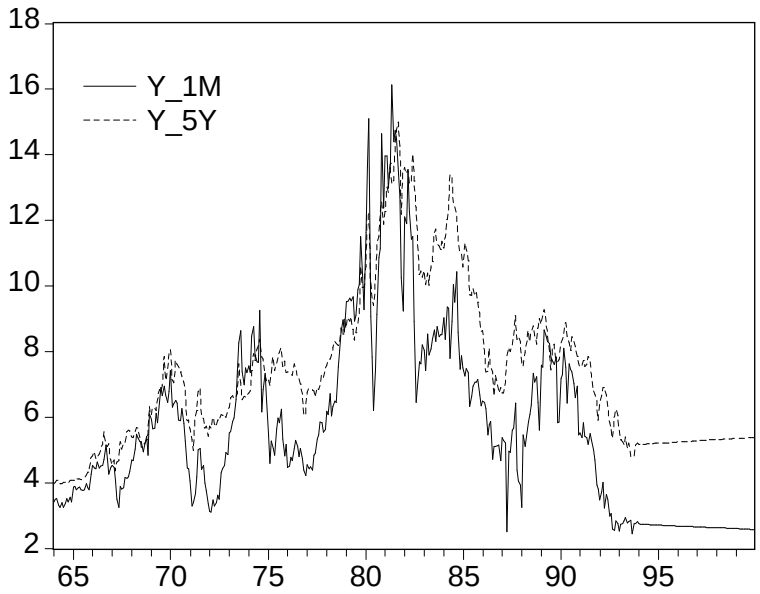
Forecasts of VAR(1) models have the same structure as forecasts of AR(1) models. The τ-step ahead forecast is given by

$$\hat{\mathbf{Y}}_{t,\tau} = (\mathbf{I} + \mathbf{\Phi}_1 + \cdots + \mathbf{\Phi}_1^{\tau-1})\boldsymbol{V} + \mathbf{\Phi}_1^{\tau}\mathbf{Y}_t.$$

These forecasts are unbiased (i.e. $\mathrm{E}[\mathbf{Y}_{t+\tau} - \hat{\mathbf{Y}}_{t,\tau}] = \mathbf{0}$). The mean squared errors of the forecasts are minimal if the disturbances are independent white-noise (see Lütkepohl, 1993, p.29). The covariance of τ-step ahead forecast errors is given by (see Lütkepohl, 1993, p.31)

$$\sum_{i=0}^{\tau-1} \mathbf{\Phi}_1^i \mathbf{\Sigma}_\epsilon (\mathbf{\Phi}_1^i)'.$$

Figure 12: Out-of-sample forecasts of one-month (Y_1M) and five-year (Y_5Y) interest rates using the VAR model in Figure 11 (estimation until 12/93).



Example 51: We consider the monthly interest rates of US treasury bills for maturities of one month ($y_t^{1\mathrm{M}}$, Y_1M) and five years ($y_t^{5\mathrm{Y}}$, Y_5Y)[122]. Both series are integrated and we fit a VAR(2) models to the first differences. The VAR(2) model was selected by AIC. The significance of estimated parameters can be used to draw conclusions about the dependence structure among the series. The estimation results in Figure 11 show a feedback relationship. The one-month rate depends on the five-year rate *and* the five-year rate depends on the one-month rate (with a lag of two periods). However, the dependence of the one-month rate on the five-year rate is much stronger (as can be seen from R^2).

Figure 12 shows dynamic out-of-sample forecasts (starting in January 1994) of the two interest rates. The forecasts converge rapidly to weakly ascending and descending linear trend lines. Their slope is determined by the (insignificant) constant terms.

[122]Source: CRSP database, Government Bond file; see file `us-tbill.wf1`; monthly data from January 1964 to December 1993; 360 observations.

Exercise 27: Use the data in the file `ccva.wf1` which is taken from Campbell et al. (2003). 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.

3.2 Cointegration and error correction models

Time series models for integrated series are usually based on applying ARMA or VAR models to (first) differences. However, it was frequently argued that differencing may eliminate valuable information about the relationship among integrated series. We now consider the case that two or more integrated series are related in terms of differences *and* levels.

3.2.1 Cointegration

Two[123] processes Y_t and X_t are **cointegrated** of first order if

1. each process is integrated of order one[124] and
2. $Z_t = Y_t - \nu - \beta X_t$ is stationary: $Z_t \sim I(0)$.

$$Y_t = \nu + \beta X_t + Z_t \tag{49}$$

is the **cointegration regression** or **cointegrating equation**.

Suppose there is an equilibrium relation between Y_t and X_t. Then Z_t represents the extent of *dis*equilibrium in the system. If Z_t is not stationary, it can move 'far away' from zero 'for a long time'. If Z_t is stationary, Z_t will 'stay close' to zero or frequently return to zero (i.e. it is mean-reverting). This is consistent with the view that both processes are controlled by a common (unobserved) stationary process. This process prevents Y_t and X_t from moving 'too far away' from each other.

3.2.2 Error correction model

If Y_t and X_t are cointegrated a **vector error correction model** (VEC) can be formulated:

$$\begin{aligned}\Delta Y_t &= \alpha_y Z_{t-1} + \nu_y + \omega_1^y \Delta X_{t-1} + \cdots + \delta_1^y \Delta Y_{t-1} + \cdots + \epsilon_t^y \\ \Delta X_t &= \alpha_x Z_{t-1} + \nu_x + \omega_1^x \Delta X_{t-1} + \cdots + \delta_1^x \Delta Y_{t-1} + \cdots + \epsilon_t^x .\end{aligned} \tag{50}$$

At least one of the coefficients α_y or α_x must be different from zero. The number of lagged differences in the VEC model can be determined by AIC or SC. If cointegration holds, models which do not include Z_{t-1} are *misspecified*.
Substituting Z_{t-1} in (50) by using the cointegrating equation (49) yields

$$\Delta Y_t = \alpha_y (Y_{t-1} - \nu - \beta X_{t-1}) + \nu_y + \omega_1^y \Delta X_{t-1} + \cdots + \delta^y \Delta Y_{t-1} + \cdots + \epsilon_t^y$$

$$\Delta X_t = \alpha_x (Y_{t-1} - \nu - \beta X_{t-1}) + \nu_x + \omega_1^x \Delta X_{t-1} + \cdots + \delta^x \Delta Y_{t-1} + \cdots + \epsilon_t^x .$$

[123] The concept of cointegration is also defined for $k>2$ series. We start to introduce the topic by considering only two time series and will gradually broaden the scope to more than two series.
[124] For simplicity, Y_t and X_t are assumed to be $I(1)$ processes. This is very often the case. However, cointegration can also be defined in terms of $I(d)$ processes.

Therefore, the structure of a VEC model corresponds to a VAR model in differences that accounts for the levels of the series using a special (linear) constraint.

A VEC model can be interpreted as follows: deviations from equilibrium, represented by Z_t, affect ΔY_t and ΔX_t such that Y_t and X_t approach each other. This mechanism 'corrects' errors (or imbalances) in the system. Therefore αZ_{t-1} is also called **error correction term**. The degree of correction depends on the so-called **speed-of-adjustment** parameters α_y and α_x.

Consider the simple case

$$\Delta Y_t = \alpha_y(Y_{t-1} - \beta X_{t-1}) + \epsilon_t^y \qquad \alpha_y < 0$$

$$\Delta X_t = \epsilon_t^x,$$

which implies $\mathrm{E}[\Delta Y_t|I_{t-1}]=\alpha_y(Y_{t-1}-\beta X_{t-1})$ and $\mathrm{E}[\Delta X_t|I_{t-1}]=0$. Three cases can be distinguished:

1. If $Y_{t-1}=\beta X_{t-1}$ ($Z_t=0$) the system is in long-run equilibrium. There is no need for adjustments and $\mathrm{E}[\Delta Y_t|I_{t-1}]=0$.
2. If $Y_{t-1}>\beta X_{t-1}$ ($Z_t>0$) the system is not in long-run equilibrium. There is a need for a downward adjustment of Y_t affected by $\mathrm{E}[\Delta Y_t|I_{t-1}]<0$.
3. If $Y_{t-1}<\beta X_{t-1}$ ($Z_t<0$) there is an upward adjustment of Y_t since $\mathrm{E}[\Delta Y_t|I_{t-1}]>0$.

Example 52: [125] Consider the relation between the spot price S_t of a stock and its corresponding futures[126] price F_t. The *cost of carry model* states that the equilibrium relation between S_t and F_t is given by

$$F_t = S_t e^{(r-d)\tau},$$

where r is the risk-free interest rate, and d is the dividend yield derived from holding the stock until the future matures in $t+\tau$. Taking logarithms yields

$$\ln F_t = \ln S_t + (r-d)\tau.$$

In practice, this relation will not hold exactly. But the difference between the left and right hand side can be expected to be stationary (or even white-noise). This suggests that the logs of spot and futures prices can described by a cointegration regression with $\nu\approx(r-d)\tau$ and $\beta\approx 1$.

[125] For details and an empirical example see Brooks et al. (2001).

[126] Futures are standardized, transferable, exchange-traded contracts that require delivery of a commodity, bond, currency, or stock index, at a specified price, on a specified future date. Unlike options, futures convey an obligation to buy.

3.2.3 Example 53: The expectation hypothesis of the term structure

The (unbiased) **expectation hypothesis** of the term structure of interest rates (EHT) states that investors are risk neutral, and bonds with different maturities are perfect substitutes. Accordingly, interest rate differentials cannot become too large since otherwise arbitrage opportunities would exist. In efficient markets such possibilities are quickly recognized and lead to a corresponding reduction of interest rate differentials. This is true even if the assumption of risk neutrality is dropped and liquidity premia are taken into account.[127] According to the EHT, a long-term interest rate can be expressed as a weighted average of current and expected short-term interest rates. Let $R_t(\tau)$ be the spot rate of a zero bond with maturity $\tau>1$ and $S_t=R_t(1)$ a short-term rate (e.g. the one-month rate). The EHT states that

$$R_t(\tau) = \frac{1}{\tau}\sum_{j=0}^{\tau-1} \mathrm{E}[S_{t+j}|I_t] + \pi(\tau),$$

where $\pi(\tau)$ is a time-invariant but maturity dependent term premium. For instance, the relation between three- and one-month interest rates is given by

$$R_t(3) = \frac{1}{3}(S_t + \mathrm{E}_t[S_{t+1}] + \mathrm{E}_t[S_{t+2}]) + \pi(3).$$

If we consider the spread between the long and the short rate we find

$$R_t(3) - S_t = \frac{1}{3}(\mathrm{E}_t[S_{t+1} - S_t] + \mathrm{E}_t[S_{t+2} - S_t]) + \pi(3).$$

Usually, interest rates are considered to be integrated processes. Thus, the terms on the right hand side are (first and higher order) differences of integrated processes and should therefore be stationary. This implies that the spread $R_t(3)-S_t$ is also stationary since both sides of the equation must have the same order of integration.

More generally, we now consider the linear combination $\beta_1 R_t(3)+\beta_2 S_t$ which can be written as (ignoring the term premium)

$$\beta_1 R_t(3) + \beta_2 S_t = (\beta_1 + \beta_2)S_t + \frac{\beta_1}{3}\left(\mathrm{E}_t[S_{t+1} - S_t] + \mathrm{E}_t[S_{t+2} - S_t]\right).$$

The linear combination $\beta_1 R_t(3)+\beta_2 S_t$ will only be stationary if the non-stationary series $(\beta_1+\beta_2)S_t$ drops from the right-hand side. Thus, the right hand side will be stationary if $\beta_1+\beta_2=0$, e.g. if $\beta_1=1$ and $\beta_2=-1$. Empirically, the EHT implies that the residuals from the cointegration regression between $R_t(3)$ and S_t should be stationary and $Z_t \approx R_t(3)-S_t-\nu$.

[127] For theoretical details see Ingersoll (1987), p.389; for an empirical example see Engsted and Tanggaard (1994).

3.2.4 The Engle-Granger procedure

Engle and Granger (1987) have developed an approach to specify and estimate error correction models which is only based on least-square regressions. The procedure consists of the following steps:

1. Test whether each series is integrated of the *same* order.
2. Estimate the cointegration regression (49) and compute $z_t = y_t - c - bx_t$. In general, fitting a regression model to the levels of integrated series may lead to the so-called **spurious regression** problem[128]. However, if cointegration holds, the parameter estimate b converges (with increasing sample size) faster to β than usual (this is also called **super-consistency**). If a VAR model is fitted to the levels of integrated series a sufficient number of lags should be included, such that the residuals are white-noise. This should avoid the spurious regression problem.
3. Test whether z_t is stationary. For that purpose use a ADF test *without* intercept since z_t has zero mean:

$$\Delta z_t = g z_{t-1} + \sum_{j=1}^{p} c_j \Delta z_{t-j} + e_t.$$

The t-statistic of g *must not* be compared to the usual critical values (e.g. those in Table 4 or those supplied by EViews). Since z_t is an estimated rather than observed time series, the critical values in Table 5[129] must be used. These critical values also depend on k (the number of series which are tested for cointegration).

If z_t is stationary we conclude that y_t and x_t are cointegrated, and a VEC model for the cointegrated time series is estimated. If z_t is integrated a VAR model using differences of y_t and x_t is appropriate.

Example 54: We illustrate the Engle-Granger procedure by using the two interest series $y_t = y_t^{1M}$ and $x_t = y_t^{5Y}$ from example 51. The assignment of the symbols y_t and x_t to the two time series is only used to clarify the exposition. It implies no assumptions about the direction of dependence, and usually[130] has no effect on the results. Details can be found in the file `us-tbill.wf1`.

Both interest rate series are assumed to be integrated, although the ADF test statistic for y_t^{1M} is -2.98, which is less than the critical value -2.87 at the 5% level. The OLS estimate of the cointegration regression is given by $y_t = -0.845 + 0.92 x_t + z_t$. The t-statistic of g in a unit-root test of z_t (using $p=1$) is -4.48. No direct comparison with the values in Table 5 is possible ($n=360$, $k=2$, $p=1$). However, -4.48 is far less than the critical values in case of $n=200$ and $\alpha=0.01$, so that the unit-root hypothesis for z_t can be rejected at the 1% level. We conclude that z_t is stationary and there is cointegration among the two interest series.

The estimated VEC model is presented in Figure 13. The upper panel of the table shows the cointegration equation and defines z_t (CointEq1): $z_t = y_t - 0.932 x_t + 0.926$. This equation is estimated by maximum likelihood, and thus differs slightly from the

[128] For details see Granger and Newbold (1974).
[129] Source: Engle and Yoo (1987), p.157.
[130] For details see Hamilton (1994), p.589.

Table 5: Critical values of the ADF t-statistic for the cointegration test.

		$p=0$			$p=4$		
		α			α		
k	n	0.01	0.05	0.10	0.01	0.05	0.10
	50	–4.32	–3.67	–3.28	–4.12	–3.29	–2.90
2	100	–4.07	–3.37	–3.03	–3.73	–3.17	–2.91
	200	–4.00	–3.37	–3.02	–3.78	–3.25	–2.98
	50	–4.84	–4.11	–3.73	–4.54	–3.75	–3.36
3	100	–4.45	–3.93	–3.59	–4.22	–3.62	–3.32
	200	–4.35	–3.78	–3.47	–4.34	–3.78	–3.51
	50	–4.94	–4.35	–4.02	–4.61	–3.98	–3.67
4	100	–4.75	–4.22	–3.89	–4.61	–4.02	–3.71
	200	–4.70	–4.18	–3.89	–4.72	–4.13	–3.83

p is the number of lags in the ADF regression. k is the number of series. α is the significance level.

OLS estimates mentioned above. The lower panel shows the error correction model. p=2 was based on the results of the VAR model in example 51. Both (changes in) interest rates depend significantly on the error correction term z_{t-1}. Thus, the changes of each time series depend on the interest rate levels, and differences between their levels, respectively. The dependencies on past interest rate changes already known from example 51 are confirmed.

The negative sign of the coefficient -0.1065 of z_{t-1} in the equation for $y_t^{1\mathrm{M}}$ can be interpreted as follows. If the five-year interest rates are much greater than the interest rates for one month, z_{t-1} is negative (according to the cointegration regression $z_t{=}y_t^{1\mathrm{M}}{-}0.932y_t^{5\mathrm{Y}}{+}0.926$). Multiplication of this negative value with the negative coefficient -0.1065 has a positive effect (c.p.) on the expected changes in $y_t^{1\mathrm{M}}$, and therefore leads to increasing short-term interest rates. This implies a tendency to reduce (or correct) large differences in interest rates. These results agree with the EHT. In efficient markets spreads among interest rates cannot become too large. The positive coefficient 0.041 in the equation for $y_t^{5\mathrm{Y}}$ can be interpreted in a similar way. A negative z_{t-1} leads to negative expected changes (c.p.) in $y_t^{5\mathrm{Y}}$, and therefore leads to a decline of the long-term interest rates. In addition, these corrections depend on past changes of both interest rates. Whereas the dependence on lagged changes could be called short-term adjustment, the response to z_{t-1} is a long-term adjustment effect.

Figure 14 shows out-of-sample forecasts (starting January 1994) of the two interest rate series using the VEC model from Figure 13. In contrast to forecasts based on the VAR model (see Figure 12), these forecasts do not diverge. This may be explained by the additional error correction term.

The Engle-Granger procedure has two drawbacks. First, if $k{>}2$ at most $(k{-}1)$ cointegration relations are (theoretically) possible. It is not straightforward how to test for cointegration in this case. Second, even when $k{=}2$ the cointegration regression between y_t and x_t can also be estimated in reverse using

$$x_t = c' + b'y_t + z'_t.$$

In principle, the formulation is arbitrary. However, since z_t and z'_t are *not*[131] identical,

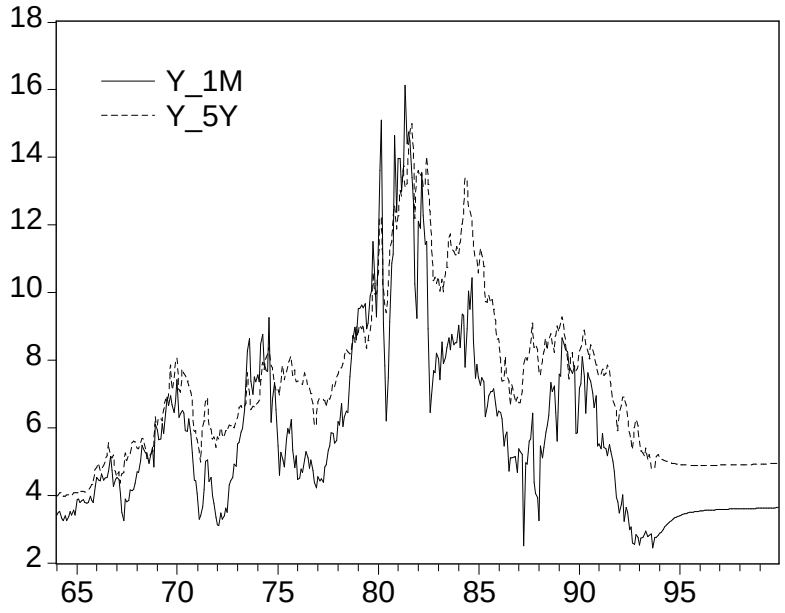
[131] Suppose we estimate the equation $y{=}b_0{+}bx{+}e$. The estimated slope is given by $b{=}s_{yx}/s_x^2$. If we

Figure 13: Cointegration regression and error correction model for one-month (Y_1M) and five-year (Y_5Y) interest rates.

Sample(adjusted): 1964:04 1993:12
Included observations: 357
t-statistics in parentheses

Cointegrating Eq:	CointEq1	
Y_1M(-1)	1.000000	
Y_5Y(-1)	-0.932143	
	(-9.54095)	
C	0.925997	
Error Correction:	**D(Y_1M)**	**D(Y_5Y)**
CointEq1	-0.106479	0.041026
	(-3.15443)	(2.22526)
D(Y_1M(-1))	-0.138652	-0.008395
	(-2.29743)	(-0.25467)
D(Y_1M(-2))	0.054517	0.031362
	(0.94613)	(0.99654)
D(Y_5Y(-1))	0.599125	0.072970
	(5.63847)	(1.25734)
D(Y_5Y(-2))	-0.270073	-0.132030
	(-2.43414)	(-2.17871)
C	-0.003420	0.003302
	(-0.08356)	(0.14772)
R-squared	0.140629	0.032406
Adj. R-squared	0.128387	0.018623
S.E. equation	0.773296	0.422361
S.D. dependent	0.828293	0.426349
Akaike Information Criteria		3.268280
Schwarz Criteria		3.420348

Figure 14: Out-of-sample forecasts of one-month (Y_1M) and five-year (Y_5Y) interest rates using the VEC model in Figure 13 (estimation until 12/93).



unit-root tests can lead to different results[132]. Engle-Granger suggest to test both variants.[133]

> **Exercise 28:** Choose two time series which you expect to be cointegrated. Use the Engle-Granger procedure to test the series for cointegration. Depending on the outcome, fit an appropriate VAR or VEC model to the series, and interpret the results.

estimate $x{=}c_0{+}cy{+}u$ (reverse regression) the estimate c will not be equal to $1/b$. $c{=}s_{yx}/s_y^2$ which is different from $1/b$ except for the special case $s_y^2{=}s_{yx}^2/s_x^2$.

[132]Unit-root tests of z_t and z_t' are equivalent only asymptotically.

[133]For details see Hamilton (1994), p.589.

3.2.5 The Johansen procedure

The Johansen procedure[134] can be used to overcome the drawbacks of the Engle-Granger approach. In addition, it offers the possibility to test whether a VEC model, a VAR model in levels, or a VAR model in (first) differences is appropriate.

The Johansen approach is based on a VAR(p+1) model of k (integrated) variables:

$$\mathbf{Y}_t = \boldsymbol{V} + \boldsymbol{\Phi}_1 \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p+1} \mathbf{Y}_{t-p-1} + \boldsymbol{\epsilon}_t.$$

This model can be reformulated to obtain the following VEC representation:

$$\Delta \mathbf{Y}_t = \boldsymbol{V} + \boldsymbol{\Gamma} \mathbf{Y}_{t-1} + \sum_{i=1}^{p} \mathbf{C}_i \Delta \mathbf{Y}_{t-i} + \boldsymbol{\epsilon}_t, \tag{51}$$

where

$$\boldsymbol{\Gamma} = \sum_{i=1}^{p+1} \boldsymbol{\Phi}_i - \mathbf{I} \qquad \mathbf{C}_i = -\sum_{j=i+1}^{p+1} \boldsymbol{\Phi}_j,$$

and $\mathbf{I}$ is a $k \times k$ unit matrix. Comparing equation(51) to the ADF test regression

$$\Delta Y_t = \nu + \gamma Y_{t-1} + \sum_{i=1}^{p} c_i \Delta Y_{t-i} + \epsilon_t$$

shows that Johansen's approach can be interpreted as a multivariate unit-root test. In the univariate case, differences of y_t are regressed on the level of y_{t-1}. In the multivariate case, differences of the vector of variables are regressed on linear combinations of the vector of past levels (represented by $\boldsymbol{\Gamma}\mathbf{Y}_{t-1}$). In the univariate case, conclusions about a unit-root of Y_t are based on the null hypothesis γ=0. Analogously, in the multivariate case, conclusions are based on the properties of the matrix $\boldsymbol{\Gamma}$ estimated from equation (51) by maximum-likelihood.

> **Review 13:** The rank $r(\mathbf{A})$ of a $m \times n$ matrix $\mathbf{A}$ is the maximum number of linearly independent rows or columns of $\mathbf{A}$ and $r(\mathbf{A}) \leq \min\{m, n\}$.
>
> A scalar λ is called eigenvalue of $\mathbf{A}$ if the equation $(\mathbf{A} - \lambda\mathbf{I})\boldsymbol{\omega} = 0$ can be solved for the non-zero eigenvector $\boldsymbol{\omega}$. The solution will be non-trivial if $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ (this is the characteristic equation). For example, for a 2×2 matrix, the characteristic equation is given by $\lambda^2 - \lambda(a_{11} + a_{22}) + (a_{11}a_{22} - a_{12}a_{21}) = 0$. Alternatively, $\lambda^2 - \lambda\text{tr}(\mathbf{A}) + |\mathbf{A}| = 0$ with solution $0.5[\text{tr}(\mathbf{A}) \pm \sqrt{\text{tr}(\mathbf{A})^2 - 4|\mathbf{A}|}]$.
>
> The maximum number of eigenvalues of a $n \times n$ matrix is n. The rank of $\mathbf{A}$ is the number of non-zero eigenvalues. Special cases are: The eigenvalues of a unit matrix are all equal to 1 (the matrix has full rank). If the unit matrix is multiplied by a all eigenvalues are equal to a. A $n \times n$ matrix with identical elements c has only one non-zero eigenvalue equal to $c \cdot n$ (its rank is one). A null matrix has rank zero.

[134] For details see Enders (2004), p.362.

The formal basis for conclusions derived from the Johansen test is provided by **Granger's representation theorem**: If the rank r of the matrix $\mathbf{\Gamma}$ is less than k, there exist $k \times r$ matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (each with rank r) such that $\mathbf{\Gamma}=\boldsymbol{\alpha}\boldsymbol{\beta}'$ and such that $\mathbf{Z}_t=\boldsymbol{\beta}'\mathbf{Y}_t$ is stationary. The rank r is equal to the number of cointegration relations (the so-called cointegration rank) and every column of $\boldsymbol{\beta}$ is a cointegration vector.

The following cases can be distinguished:

1. If $\mathbf{\Gamma}$ has rank zero there is no cointegration. This corresponds to the case $\gamma=0$ in the univariate ADF test. In this case, all elements of $\mathbf{Y}_t$ are unit-root series and there exists no stationary linear combination of these elements. Therefore, a VAR model in first differences (with no error correction term) is appropriate.

2. If $\mathbf{\Gamma}$ has full rank there is no cointegration either. In this case, all elements of $\mathbf{Y}_t$ are stationary, which corresponds to the situation $\gamma \neq 0$ in the univariate ADF test. Therefore, a VAR model in levels (with no error correction term) is appropriate.

3. Cointegration holds if $\mathbf{\Gamma}$ has rank $0<r<k$. In this case a VEC model is appropriate.

If $k=2$ and $r \leq 1$ the decomposition of $\mathbf{\Gamma}$ can be written as

$$\mathbf{\Gamma} = \boldsymbol{\alpha}\boldsymbol{\beta}' = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} (\beta_1 \ \beta_2) = \begin{pmatrix} \alpha_1\beta_1 & \alpha_1\beta_2 \\ \alpha_2\beta_1 & \alpha_2\beta_2 \end{pmatrix}.$$

This implies the error correction model

$$\Delta Y_{1t} = \alpha_1(\beta_1 Y_{1t-1} + \beta_2 Y_{2t-1}) + \epsilon_{1t}$$

$$\Delta Y_{2t} = \alpha_2(\beta_1 Y_{1t-1} + \beta_2 Y_{2t-1}) + \epsilon_{2t}.$$

This decomposition is not feasible if the rank is full (i.e. $|\mathbf{\Gamma}| \neq 0$). If the rank is not full (i.e. $|\mathbf{\Gamma}|=0$) the eigenvalues are given by $0.5[\text{tr}(\mathbf{\Gamma}) \pm \text{tr}(\mathbf{\Gamma})]$. Thus, one eigenvalue will be zero and the second eigenvalue is the trace of $\mathbf{\Gamma}$. Cointegration obtains if the trace is different from zero. For the rank to be zero (i.e. no cointegration) $\text{tr}(\mathbf{\Gamma})$ must be zero. This holds if $\alpha_1/\alpha_2=-\beta_2/\beta_1$, i.e. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal.

Note that the matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are *not unique*. If $\mathbf{Z}_t=\boldsymbol{\beta}'\mathbf{Y}_t$ is stationary, then $c\boldsymbol{\beta}'\mathbf{Y}_t$ will also be stationary for any nonzero scalar c. In general, any linear combination of the cointegrating relations is also a cointegrating relation. This non-uniqueness typically leads to normalizations of $\boldsymbol{\beta}$ that make the interpretation of $\mathbf{Z}_t$ easier (see example 58 below).

Example 55: We use the results from example 54 (see Figure 13) and omit the constants for simplicity. $\mathbf{\Gamma}$ can be written as

$$\mathbf{\Gamma} = \begin{pmatrix} -0.107 \\ 0.041 \end{pmatrix} (1 \ -0.932) = \begin{pmatrix} -0.107 & 0.0997 \\ 0.041 & -0.0382 \end{pmatrix}.$$

The non-zero eigenvalue is –0.145. Simulated sample paths of the cointegrated series can be found in the file `vec.xls` on the sheet `rank=1`. The sheet `rank=0` illustrates the case $r=0$ where $\alpha_1=-\alpha_2\beta_2/\beta_1=0.0382$ is used. The simulated paths of both cases are based on the same disturbances. Comparisons clearly show that the paths of the two

random walks typically deviate from each other more strongly and for longer periods of time than the two cointegrated series.

A different normalization can be obtained if $\boldsymbol{\beta}$ is divided by -0.932 and $\boldsymbol{\alpha}$ is multiplied by -0.932. $\boldsymbol{\Gamma}$ remains unchanged by this linear transformation. The new normalization for $\boldsymbol{\beta}=(-1.075 \quad 1)'$ implies that $z_t \approx y_t^{5Y} - y_t^{1M}$, which corresponds to the more frequently used definition of an interest rate spread.

The Johansen test involves estimating[135] the VEC model (51) and to test how many eigenvalues of the estimated matrix $\boldsymbol{\Gamma}$ are significant. Two different types of tests are available. Their critical values are tabulated and depend on p – the number of lags in the VEC model – and on assumptions about constant terms. To determine the order p of the VEC model VAR models with increasing order are fitted to the *levels* of the series. p is chosen such that a VAR(p+1) model fitted to the levels has minimum AIC or SC. Setting p larger than necessary is less harmful than choosing a value of p that is too small. If a level VAR(1) has minimum AIC or SC (i.e. p=0) this may indicate that the series are stationary. In this case the Johansen test can be carried out using p=1 to confirm this (preliminary) evidence.

The following five assumptions about constant terms and trends in the cointegrating equation (49) and in the error correction model (50) can be distinguished:[136]

1. There are no constant terms in the cointegrating equation and the VEC model: $\nu=\nu_y=\nu_x=0$.[137]

2. The cointegrating equation has a constant term $\nu\neq 0$, but the VEC model does not have constant terms: ν_y, $\nu_x=0$.[138]

3. The cointegrating equation and the VEC model have constant terms: ν, ν_y, $\nu_x \neq 0$.[139] ν_y, $\nu_x\neq 0$ is equivalent to assuming a 'linear trend in the data' because a constant term in the VEC model for $\Delta\mathbf{Y}_t$ corresponds to a drift in the levels $\mathbf{Y}_t$.

4. The cointegrating equation has a constant and a linear trend ($Y_t=\nu+\delta t+\beta X_t+Z_t$). This case accounts for the possibility that the imbalance between Y_t and X_t may linearly increase or decrease. Accordingly, the difference in the levels need not necessarily approach zero or ν, but may change in a *deterministic* way. The VEC model has constant terms: ν_y, $\nu_x\neq 0$.[140]

5. The cointegrating equation has a constant term $\nu\neq 0$ and a linear trend. The VEC model has constants (ν_y, $\nu_x\neq 0$) and a linear trend. The presence of a linear trend in addition to the drift corresponds to a quadratic trend in the level of the series.[141]

The conclusions about cointegration will usually depend on the assumptions about constant terms and trends. This choice may be supported by inspecting graphs of the series or by economic reasoning (for instance, a quadratic trend in interest rates may be excluded

[135] For details about the maximum likelihood estimation of VEC models see Hamilton (1994), p.635.
[136] We only consider the simplest case of two series.
[137] EViews: VAR assumes no deterministic trend in data: No intercept or trend in CE or test VAR.
[138] EViews: Assume no deterministic trend in data: intercept (no trend) in CE - no intercept in VAR.
[139] EViews: Allow for linear deterministic trend in data: Intercept (no trend) in CE and test VAR.
[140] EViews: Allow for linear deterministic trend in data: Intercept and trend in CE - no trend in VAR.
[141] EViews: Allow for quadratic deterministic trend in data: Intercept and trend in CE - linear trend in VAR.

apriori). If it is difficult to decide which assumption is most reasonable, the Johansen test can be carried out under all five assumptions. The results can be used to select an assumption that is well supported by the data.

Figure 15: Johansen test for cointegration among y_t^{1M} and y_t^{5Y}.

Sample: 1964:01 1993:12
Included observations: 357
Test assumption: No deterministic trend in the data
Series: Y_1M Y_5Y
Lags interval: 1 to 2

Eigenvalue	Likelihood Ratio	5 Percent Critical Value	1 Percent Critical Value	Hypothesized No. of CE(s)
0.069316	29.35324	19.96	24.60	None **
0.010333	3.708043	9.24	12.97	At most 1

*(**) denotes rejection of the hypothesis at 5%(1%) significance level
L.R. test indicates 1 cointegrating equation(s) at 5% significance level

Figure 16: Summary of the Johansen test for cointegration among y_t^{1M} and y_t^{5Y} under different assumptions.

Sample: 1964:01 1993:12
Included observations: 358
Series: Y_1M Y_5Y
Lags interval: 1 to 1

Data Trend:	None	None	Linear	Linear	Quadratic
Rank or No. of CEs	No Intercept No Trend	Intercept No Trend	Intercept No Trend	Intercept Trend	Intercept Trend
Akaike Information Criteria by Model and Rank					
0	3.306984	3.306984	3.318016	3.318016	3.324710
1	3.253165	3.255658	3.261239	3.256909	3.259051
2	3.274557	3.270444	3.270444	3.271651	3.271651
Schwarz Criteria by Model and Rank					
0	3.350342	3.350342	3.383053	3.383053	3.411426
1	3.339881	3.353214	3.369634	3.376144	3.389125
2	3.404631	3.422196	3.422196	3.445082	3.445082
L.R. Test:	Rank = 1	Rank = 1	Rank = 2	Rank = 1	Rank = 1

Example 56: Fitting VAR models to the levels of y_t^{1M} and y_t^{5Y} indicates that p=1 should be used to estimate the VEC model for the Johansen test. However, we choose p=2 to obtain results that are comparable to example 54. Below we will obtain test results using p=1. Figure 15 shows the results of the test. The assumption No deterministic trend in the data was used because it appears most plausible in economic terms, and is supported by the results obtained in example 54. EViews provides an interpretation of the test results: L.R. test indicates 1 cointegrating equation(s) at 5% significance level. The null hypothesis 'no cointegration' (None) is rejected at the 1% level. The hypothesis of at most one cointegration relation cannot be rejected. This confirms the conclusion drawn in example 54 that cointegration among y_t^{1M} and y_t^{5Y} exists.

Figure 16 contains a summary of the results for various assumptions and p=1. The last line indicates which rank can be concluded on the basis of the likelihood-ratio test for

each assumption, using a 5% level. The conclusion r=1 is drawn for all assumptions, except the third.

In addition, AIC and SC for every possible rank and every assumption are provided. Note that the specified rank in the row L.R. Test is based on the estimated eigenvalues. The rank is not determined on the basis of AIC or SC, and therefore need not correspond to these criteria (e.g., under assumption 2, SC points at r=0).

For a given rank, the values in a row can be compared to find out which assumption about the data is most plausible. Since the alternatives within a line are nested, the precondition for a selection on the basis of AIC and SC is met. If conclusions about the cointegration rank are not unique, and/or no assumption about constant terms and trends is particularly justified, AIC and SC may be used heuristically in order to search for a global minimum across assumptions and ranks. As it turns out both criteria agree in pointing at assumption 1. This corresponds to the result that the intercept terms in the VEC model are not significant (see Figure 13). Therefore, assuming a drift in interest rates is not compatible with the data and could hardly be justified using economic reasoning.

Exercise 29: Choose two time series which you expect to be cointegrated. Use the Johansen procedure to test the series for cointegration. Depending on the outcome of the test, fit an appropriate VAR or VEC model to the series, and interpret the results.

3.2.6 Cointegration among more than two series

Example 57: [142] The purchasing power parity (PPP) states that the currencies of two countries are in equilibrium when their purchasing power is the same in each country. In the long run the exchange rate should equal the ratio of the two countries' price levels. There may be short-term deviations from this relation which should disappear rather quickly. According to the theory, the real exchange rate is given by

$$Q_t = \frac{F_t P_t^f}{P_t^d},$$

where F_t is the nominal exchange rate in domestic currency per unit of foreign currency, P_t^d is the domestic price level, and P_t^f is the foreign price level. Taking logarithms yields the linear relation

$$\ln F_t + \ln P_t^f - \ln P_t^d = \ln Q_t.$$

The PPP holds if the logs of F_t, P_t^d and P_t^f are cointegrated with cointegration vector $\boldsymbol{\beta}=(1\ 1\ {-1})'$, and the log of Q_t is stationary.

Example 58: Applying the EHT to more than two interest rates implies that all spreads between long- and short-term interest rates ($R_t(\tau_1)-S_t$, $R_t(\tau_2)-S_t$, etc.) should be stationary. In a VEC model with k interest rates this implies $k-1$ cointegration relations. For instance, if $k=4$ and $\mathbf{Y}_t=(S_t, R_t(\tau_1), R_t(\tau_2), R_t(\tau_3))'$ the $k\times(k-1)$ cointegration matrix $\boldsymbol{\beta}$ is given by

$$\begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{52}$$

Extending example 51, we add the one year interest rate ($y_t^{1\mathrm{Y}}$, `Y_1Y`) to the one-month and five-year rates. Fitting VAR models to the levels indicates that lagged differences of order one are sufficient. The results from the Johansen test clearly indicate the presence of two cointegration relations (see file `us-tbill.wf1`). The upper panel of Figure 17 shows the so-called **triangular representation** (see Hamilton, 1994, p.576) of the two cointegration vectors used by EViews to identify $\boldsymbol{\beta}$. Since any linear combination of the cointegrating relations is also a cointegrating relation, this representation can be transformed to obtain the structure of $\boldsymbol{\beta}$ in equation 52. For simplicity, we set the coefficients in the row of `Y_5Y(-1)` in Figure 17 equal to -1, and ignore the constants. The representation in Figure 17 implies that the spreads $y_t^{1\mathrm{M}}-y_t^{5\mathrm{Y}}$ and $y_t^{1\mathrm{Y}}-y_t^{5\mathrm{Y}}$ are stationary. Using a suitable transformation matrix we obtain

$$\begin{pmatrix} -0.5 & -0.5 & -0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The transformed matrix $\boldsymbol{\beta}$ now implies that the spreads $y_t^{1\mathrm{Y}}-y_t^{1\mathrm{M}}$ and $y_t^{5\mathrm{Y}}-y_t^{1\mathrm{M}}$ are stationary. The lower panel in Figure 17 shows significant speed-of-adjustment coefficients in all cases. The effects of lagged differences are clearly less important.

[142] For empirical examples see Hamilton (1994), p.582 or Chen (1995).

Figure 17: VEC model for one-month, and one- and five-year interest rates.

Sample(adjusted): 1964:03 1993:12
Included observations: 358
t-statistics in parentheses

Cointegrating Eq:	CointEq1	CointEq2	
Y_1M(-1)	1.000000	0.000000	
Y_1Y(-1)	0.000000	1.000000	
Y_5Y(-1)	-0.868984	-0.965373	
	(-8.57599)	(-12.0787)	
C	0.522425	0.332333	
	(0.63841)	(0.51487)	
Error Correction:	**D(Y_1M)**	**D(Y_1Y)**	**D(Y_5Y)**
CointEq1	-0.314107	0.122230	0.127018
	(-4.45035)	(2.14357)	(3.21582)
CointEq2	0.327518	-0.211224	-0.141137
	(3.30355)	(-2.63713)	(-2.54388)
D(Y_1M(-1))	-0.203101	-0.068047	-0.081964
	(-2.82280)	(-1.17063)	(-2.03563)
D(Y_1Y(-1))	0.375950	0.149988	0.130965
	(2.32363)	(1.14745)	(1.44644)
D(Y_5Y(-1))	0.093839	0.087987	-0.007274
	(0.48405)	(0.56177)	(-0.06704)
R-squared	0.183793	0.033800	0.035264
Adj. R-squared	0.174545	0.022851	0.024332
S.E. equation	0.751489	0.607129	0.420547
S.D. dependent	0.827134	0.614187	0.425759
Akaike Information Criteria		3.226556	
Schwarz Criteria		3.475864	

Exercise 30: Choose three time series which you expect to be cointegrated. Use the Johansen procedure to test the series for cointegration. Depending on the outcome, fit an appropriate VAR or VEC model to the series, and interpret the results.

3.3 State space modeling and the Kalman filter[143]

3.3.1 The state space formulation

The objective of state-space modeling is to estimate (the parameters of) an unobservable vector process $\boldsymbol{\alpha}_t$ ($k\times1$) on the basis of an observable process y_t (which may, in general, be a vector process, too). Two equations are distinguished. For a single observation t the **measurement**, **signal** or **observation equation** is given by

$$y_t = \boldsymbol{c}_t + \boldsymbol{z}_t'\boldsymbol{\alpha}_t + \epsilon_t,$$

and can be viewed as a regression model with (potentially) time-varying coefficients $\boldsymbol{\alpha}_t$ and $\boldsymbol{c}_t$. $\boldsymbol{z}_t$ is the $k\times1$ vector of regressors and ϵ_t is the residual. $\boldsymbol{\alpha}_t$ is assumed to be a first-order (vector) autoregression as defined in the **system** or **transition equation**

$$\boldsymbol{\alpha}_t = \boldsymbol{d}_t + \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t.$$

The disturbances ϵ_t and $\boldsymbol{\eta}_t$ are assumed to be serially independent with mean zero and covariance

$$\mathrm{V}\begin{bmatrix}\epsilon_t\\ \boldsymbol{\eta}_t\end{bmatrix} = \begin{bmatrix}h & \boldsymbol{G}\\ \boldsymbol{G}' & \boldsymbol{Q}\end{bmatrix}.$$

The state space formulation can be used for a variety of models (see Harvey (1989) or Wang (2003)). The main areas of application are regressions with time-varying coefficients and the extraction of unobserved components (or latent, underlying factors) from observed series. Harvey (1984) has proposed so-called **structural models** to extract (or estimate) trend and seasonal components from a time series. One example is a model with (unobservable) level μ_t and trend β_t defined in the system and measurement equations as follows

$$\begin{bmatrix}\mu_t\\ \beta_t\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mu_{t-1}\\ \beta_{t-1}\end{bmatrix} + \begin{bmatrix}u_t\\ v_t\end{bmatrix} \qquad y_t = [1\ 0]\begin{bmatrix}\mu_t\\ \beta_t\end{bmatrix} + \epsilon_t. \tag{53}$$

This model can be viewed as a random walk with time-varying drift β_t. If σ_v^2=0 the drift is constant.

The **stochastic volatility model** is another model that can be formulated in state space form. Volatility is unobservable and is treated as the state variable. We define h_t=$\ln\sigma_t^2$ with transition equation

$$h_t = d + Th_{t-1} + \eta_t.$$

The observed returns are defined as y_t=$\sigma_t\epsilon_t$ where $\epsilon_t\sim$N$(0,1)$. If we define g_t=$\ln y_t^2$ and κ_t=$\ln\epsilon_t^2$ the observation equation can be written as

$$g_t = h_t + \kappa_t.$$

[143] For a more comprehensive treatment of this topic see Harvey (1984, 1989), Hamilton (1994), chapter 13, or Wang (2003), chapter 7.

3.3.2 The Kalman filter

The Kalman filter is a recursive procedure to estimate $\boldsymbol{\alpha}_t$. Assume for the time being that all vectors and matrices except the state vector are known. The recursion proceeds in two steps. In the **prediction step** $\boldsymbol{\alpha}_t$ is estimated using the available information in $t-1$. This estimate $\boldsymbol{a}_{t|t-1}$ is used to obtain the prediction $y_{t|t-1}$ for the observable process y_t. In the **updating step** the actual observation y_t is compared to $y_{t|t-1}$. Based on the prediction error $y_t - y_{t|t-1}$ the original estimate of the state vector is updated to obtain the (final) estimate $\boldsymbol{a}_t$.

The conditional expectation of $\boldsymbol{\alpha}_t$ is given by

$$\boldsymbol{a}_{t|t-1} = \mathrm{E}_{t-1}[\boldsymbol{\alpha}_t] = \boldsymbol{d}_t + \boldsymbol{T}_t \boldsymbol{a}_{t-1},$$

and the covariance of the prediction error is

$$\boldsymbol{P}_{t|t-1} = \mathrm{E}_{t-1}[(\boldsymbol{\alpha}_t - \boldsymbol{a}_t)(\boldsymbol{\alpha}_t - \boldsymbol{a}_t)'] = \boldsymbol{T}_t \boldsymbol{P}_{t-1} \boldsymbol{T}_t' + \boldsymbol{Q}.$$

Given the estimate $\boldsymbol{a}_{t|t-1}$ for $\boldsymbol{\alpha}_t$ we can estimate the conditional mean of y_t from

$$y_{t|t-1} = \mathrm{E}_{t-1}[y_t] = \boldsymbol{c}_t + \boldsymbol{z}_t' \boldsymbol{a}_{t|t-1}.$$

The prediction error $e_t = y_t - y_{t|t-1}$ is used in the updating equations

$$\boldsymbol{a}_t = \boldsymbol{a}_{t|t-1} + \boldsymbol{P}_{t|t-1} \boldsymbol{z}_t \boldsymbol{F}_t^{-1} e_t$$

$$\boldsymbol{P}_t = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \boldsymbol{z}_t \boldsymbol{F}_t^{-1} \boldsymbol{z}_t' \boldsymbol{P}_{t|t-1}.$$

$\boldsymbol{F}_t$ is the **Kalman gain**

$$\boldsymbol{F}_t = \boldsymbol{z}_t' \boldsymbol{P}_{t|t-1} \boldsymbol{z}_t + h,$$

which determines the correction of $\boldsymbol{a}_{t|t-1}$ and $\boldsymbol{P}_{t|t-1}$.

The application of the Kalman filter requires to specify starting values $\boldsymbol{a}_0$ and $\boldsymbol{P}_0$. In addition $\boldsymbol{c}_t$, $\boldsymbol{z}_t$, $\boldsymbol{d}_t$, $\boldsymbol{T}_t$, h, $\boldsymbol{G}$ and $\boldsymbol{Q}$ need to be fixed or estimated from a sample. In general they may depend on further parameters to be estimated. Given a sample of n observations and assuming that ϵ_t and $\boldsymbol{\eta}_t$ are multivariate normal the log-likelihood is given by

$$\log L = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{t=1}^{n} \ln |\boldsymbol{F}_t| - \frac{1}{2} \sum_{t=1}^{n} e_t' \boldsymbol{F}_t^{-1} e_t. \tag{54}$$

The initial state vector $\boldsymbol{\alpha}_0$ can also be estimated or set to 'reasonable' values. The diagonal elements of the initial covariance matrix $\boldsymbol{P}_0$ are usually set to large values (e.g. 10^4), depending on the accuracy of prior information about $\boldsymbol{\alpha}_0$.

The stochastic volatility model cannot be estimated by ML using a normal assumption. Harvey et al. (1994) and Ruiz (1994) have proposed a QML approach for this purpose.

Example 59: Estimating a time-varying beta-factor excluding a constant term is a very simple application of the Kalman filter (see Bos and Newbold (1984) for a more comprehensive study). The system and observation equations are given by

$$\beta_t = \beta_{t-1} + \eta_t \qquad x_t^i = \beta_t x_t^m + \epsilon_t.$$

In other words we assume that the beta-factor evolves like a random walk without drift. Details of the Kalman filter recursion and ML estimation can be found in the file `kalman.xls`. Note that the final, updated estimate of the state vector is equal to the LS estimate using the entire sample.

3.3.3 Example 60: The Cox-Ingersoll-Ross model of the term structure

In the K-factor Cox-Ingersoll-Ross (CIR) term structure model (see Cox et al., 1985) the instantaneous nominal interest rate i_t is assumed to be the sum of K state variables (or factors) $X_{t,j}$:

$$i_t = \sum_{j=1}^{K} X_{t,j}. \tag{55}$$

The factors $X_{t,j}$ are assumed to be independently generated by a square-root process

$$dX_{t,j} = \kappa_j(\theta_j - X_{t,j})dt + \sigma_j\sqrt{X_{t,j}}dZ_{t,j} \quad (j = 1, \ldots, K),$$

where $Z_{t,j}$ are independent Wiener processes, θ_j are the long-term means of $X_{t,j}$, and κ_j are their mean reversion parameters. The volatility parameters σ_j determine the magnitude of changes in $X_{t,j}$.

The price of a pure discount bond with face value 1 maturing at time $t+T$ is

$$P_t(T) = \prod_{j=1}^{K} A_j(T) \exp\left(-\sum_{j=1}^{K} B_j(T)X_{t,j}\right),$$

where

$$A_j(T) = \left(\frac{2\phi_{j,1}\exp(\phi_{j,2}T/2)}{\phi_{j,4}}\right)^{\phi_{j,3}}, \tag{56}$$

$$B_j(T) = \frac{2(\exp(\phi_{j,1}T) - 1)}{\phi_{j,4}}, \tag{57}$$

$$\phi_{j,1} = \sqrt{(\kappa_j + \lambda_j)^2 + 2\sigma_j^2}, \qquad \phi_{j,2} = \kappa_j + \lambda_j + \phi_{j,1}, \qquad \phi_{j,3} = 2\kappa_j\theta_j/\sigma_j^2,$$

$$\phi_{j,4} = 2\phi_{j,1} + \phi_{j,2}(\exp(\phi_{j,1}T) - 1).$$

The parameters λ_j are negatively related to the risk premium.

The yield to maturity at time t of a pure discount bond which matures at time $t+T$ is defined as

$$Y_t(T) = -\frac{\log P_t(T)}{T} = \sum_{j=1}^{K} -\frac{\log A_j(T)}{T} + \frac{B_j(T)X_{t,j}}{T}, \tag{58}$$

which is affine in the state-variables $X_{t,j}$.

To estimate parameters and to extract the unobservable state variables from yields observed at discrete time intervals we use a state-space formulation of the CIR model. We define the state-vector $x_t=(X_{t,1},\ldots,X_{t,K})'$. The exact transition density $\mathrm{P}(x_t|x_{t-1})$ for the CIR-model is the product of K non-central χ^2-densities. A quasi-maximum-likelihood (QML) estimation of the model parameters can be carried out by substituting the exact transition density by a normal density:

$$x_t|x_{t-1} \sim \mathrm{N}(\mu_t, Q_t).$$

μ_t and Q_t are determined in such a way that the first two moments of the approximate normal and the exact transition density are equal. The elements of the K-dimensional vector μ_t are defined as

$$\mu_{t,j} = \theta_j[1-\exp(-\kappa_j\Delta t)] + \exp(-\kappa_j\Delta t)X_{t-1,j},$$

where Δt is a discrete time interval. Q_t is a $K\times K$ diagonal matrix with elements

$$Q_{t,j} = \sigma_j^2\frac{1-\exp(-\kappa_j\Delta t)}{\kappa_j}\left(\frac{\theta_j}{2}[1-\exp(-\kappa_j\Delta t)] + \exp(-\kappa_j\Delta t)X_{t-1,j}\right).$$

Let $y_t=(Y_{t,1},\ldots,Y_{t,m})'$ be the m-dimensional vector of yields observed at time t. The observation density $\mathrm{P}(y_t|x_t)$ is based on the linear relation (58) between observed yields and the state variables. The measurement equation for observed yields is:

$$y_t = a_t + b_t x_t + \epsilon_t \qquad \epsilon_t \sim \mathrm{NID}(0,H) \qquad (t=1,\ldots,n),$$

where n is the number of observations, a_t is a $m\times 1$ vector derived from (56) and b_t is a $m\times K$ matrix derived from (57):

$$a_t = -\sum_{j=1}^{K}\frac{\log A_j(T_{t,i})}{T_{t,i}} \qquad (i=1,\ldots,m),$$

$$b_t = \frac{B_j(T_{t,i})}{T_{t,i}} \qquad (i=1,\ldots,n),(j=1,\ldots,K).$$

T_t is a $m\times 1$ vector of maturities associated with the vector of yields. H is the variance-covariance matrix of ϵ_t with constant dimension $m\times m$. It is assumed to be a diagonal matrix but each diagonal element h_i (i=1,...,m) may be different such that the variance of errors may depend on maturity.

The Kalman filter recursion consists of the following equations:

$$x_{t|t-1} = \theta[1 - \exp(-\kappa)] + \exp(-\kappa)x_{t-1|t-1}$$

$$\hat{y}_t = a_t + b_t x_{t|t-1}.$$

The Kalman filter requires initial values for t=0 for the factors and their variance-covariance matrix. We set the initial values for $X_{t,j}$ and P_t equal to their unconditional moments: $X_{0,j}$=θ_j and diagonal elements of P_0 are $0.5\theta_j\sigma_j^2/\kappa_j$. The initial values for the parameters $\{\kappa_j, \theta_j, \sigma_j, \lambda_j, h_i\}$ can be based on random samples of the parameter vector. Further details and results from an empirical example can be found in Geyer and Pichler (1999).

Bibliography

Albright, S. C., Winston, W., and Zappe, C. J. (2002). *Managerial Statistics*. Duxbury.

Baxter, M. and Rennie, A. (1996). *Financial Calculus*. Cambridge University Press.

Blattberg, R. and Gonedes, N. (1974). A comparison of the stable and Student distributions as statistical models for stock prices. *Journal of Business*, 47:244–280.

Bollerslev, T., Chou, R., and Kroner, K. F. (1992). ARCH modeling in finance. *Journal of Econometrics*, 52:5–59.

Bos, T. and Newbold, P. (1984). An empirical investigation of the possibility of stochastic systematic risk in the market model. *Journal of Business*, 57:35–41.

Box, G. and Jenkins, G. (1976). *Time Series Analysis Forecasting and Control*. Holden-Day, revised edition.

Brooks, C., Rew, A., and Ritson, S. (2001). A trading strategy based on the lead-lag relationship between the spot index and futures contract for the FTSE 100. *International Journal of Forecasting*, 17:31–44.

Campbell, J. Y., Chan, Y. L., and Viceira, L. M. (2003). A multivariate model of strategic asset allocation. *Journal of Financial Economics*, 67:41–80.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princton University Press.

Chan, K., Karoly, G. A., Longstaff, F. A., and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance*, 47:1209–1227.

Chatfield, C. (1989). *The Analysis of Time Series*. Chapman and Hall, 4th edition.

Chen, B. (1995). Long-run purchasing power parity: Evidence from some European monetary system countries. *Applied Economics*, 27:377–383.

Chen, N.-F., Roll, R., and Ross, S. A. (1986). Economic forces and the stock market. *Journal of Business*, 59:383–403.

Cochrane, J. H. (2001). *Asset Pricing*. Princeton University Press.

Coen, P., Gomme, F., and Kendall, M. G. (1969). Lagged relationships in economic forecasting. *Journal of the Royal Statistical Society Series A*, 132:133–152.

Cox, J., Ingersoll, J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53:385–407.

Dhillon, U., Shilling, J., and Sirmans, C. (1987). Choosing between fixed and adjustable rate mortgages. *Journal of Money, Credit and Banking*, 19:260–267.

Enders, W. (2004). *Applied Econometric Time Series*. Wiley, 2nd edition.

Engel, C. (1996). The forward discount anomaly and the risk premium: A survey of recent evidence. *Journal of Empirical Finance*, 3:123–238.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55:251–276.

Engle, R. F. and Yoo, B. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, 35:143–159.

Engsted, T. and Tanggaard, C. (1994). Cointegration and the US term structure. *Journal of Banking and Finance*, 18:167–181.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47:427–465.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81:607–636.

Fielitz, B. and Rozelle, J. (1983). Stable distributions and the mixtures of distributions hypotheses for common stock returns. *Journal of the American Statistical Association*, 78:28–36.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. Wiley.

Geyer, A. and Pichler, S. (1999). A state-space approach to estimate and test multifactor Cox-Ingersoll-Ross models of the term structure. *Journal of Financial Research*, 22:107–130.

Gourieroux, C. and Jasiak, J. (2001). *Financial Econometrics*. Princeton University Press.

Granger, C. W. and Newbold, P. (1971). Some comments on a paper of Coen, Gomme and Kendall. *Journal of the Royal Statistical Society Series A*, 134:229–240.

Granger, C. W. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120.

Greene, W. H. (2000). *Econometric Analysis*. Prentice Hall, 4th edition.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, 5th edition.

Hamilton, J. D. (1994). *Time Series Analysis*. Princton University Press.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054.

Hansen, L. P. and Singleton, K. J. (1996). Efficient estimation of linear asset-pricing models with moving average errors. *Journal of Business and Economic Statistics*, 14:53–68.

Harvey, A. C. (1984). A unified view of statistical forecasting procedures. *Journal of Forecasting*, 3:245–275.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Harvey, A. C., Ruiz, E., and Shepard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, 61:247–64.

Hastings, N. and Peacock, J. (1975). *Statistical Distributions*. Butterworth.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Ingersoll, J. E. (1987). *Theory of Financial Decision Making*. Rowman & Littlefield.

Jarrow, R. A. and Rudd, A. (1983). *Option Pricing*. Dow Jones-Irwin.

Kiefer, N. (1988). Economic duration data and hazard functions. *Journal of Economic Literature*, 26:646–679.

Kiel, K. A. and McClain, K. T. (1995). House prices during siting decision stages: The case of an incinerator from rumor through operation. *Journal of Environmental Economics and Management*, 28:241–255.

Kirby, C. (1997). Measuring the predictable variation in stock and bond returns. *Review of Financial Studies*, 10:579–630.

Kmenta, J. (1971). *Elements of Econometrics*. Macmillan.

Kon, S. J. (1984). Models of stock returns – a comparison. *Journal of Finance*, 39:147–165.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 52:159–178.

Lamoureux, C. and Lastrapes, W. (1990). Heteroscedasticity in stock return data: volume versus GARCH effects. *Journal of Finance*, 45:221–229.

Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review*, 87(4):270–290.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer.

Mills, T. C. (1993). *The Econometric Modelling of Financial Time Series*. Cambridge University Press.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*, 20(4):111–132.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 2nd edition.

Roberts, M. R. and Whited, T. M. (2012). *Endogeneity in Empirical Corporate Finance*, volume 2A of *Handbook of the Economics of Finance*. Elsevier.

Roll, R. and Ross, S. A. (1980). An empirical investigation of the arbitrage pricing theory. *Journal of Finance*, 35:1073–1103.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13:341–360.

Ruiz, E. (1994). Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics*, 63:289–306.

SAS (1995). *Stock Market Analysis using the SAS System*. SAS Institute.

Shanken, J. (1992). On the estimation of beta-pricing models. *Review of Financial Studies*, 5:1–33.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, 54:375–421.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20(4):518–529.

Studenmund, A. (2001). *Using Econometrics*. Addison Wesley Longman.

Thomas, R. (1997). *Modern Econometrics*. Addison Wesley.

Tsay, R. S. (2002). *Analysis of Financial Time Series*. Wiley.

Valkanov, R. (2003). Long-horizon regressions: Theoretical results and applications. *Journal of Financial Economics*, 68:201–232.

Verbeek, M. (2004). *Modern Econometrics*. Wiley, 2nd edition.

Wang, P. (2003). *Financial Econometrics*. Routledge.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2003). *Introductory Econometrics*. Thomson, 2nd edition.

Yogo, M. (2004). Estimating the elasticity of intertemporal substitution when instruments are weak. *The Review of Economics and Statistics*, 86(3):797–810.