

Lecture Notes in Financial Econometrics (MSc course)

Paul Söderlind¹

6 July 2025

¹University of St. Gallen. *Address:* s/bf-HSG, Unterer Graben 21, CH-9000 St. Gallen, Switzerland. *E-mail:* Paul.Soderlind@unisg.ch. Document name: FinEcmtAll.TeX. ©Paul Söderlind.

This text is for a second M.A. course in econometrics. The goal is to present econometrics in a way so that it can readily be applied, but also understood. Point estimates, for instance, OLS coefficients, are often straightforward to calculate and understand, but standard errors can be trickier. The text therefore spends some efforts on analysing correlations (across time or within a cluster of cross-sectional units), explaining how that matters and how the standard errors can be calculated to better reflect the true uncertainty.

The approach is to motivate the key results, although leaving out some details. This means that many proofs lack the proper restrictions on the data generating process, and rather says something like “assuming the limit is well defined and finite.” Instead, the text includes many Monte Carlo simulations to highlight the properties of the estimators, including cases where things go wrong. Unless mentioned otherwise, all Monte Carlo simulations draw 25,000 samples. The text also has very many empirical examples, often illustrated in plots. Optional (often more advanced) material is denoted by a star*).

In implementing numerical computations based on these notes, my students have typically used Julia, Matlab, Python or R. Julia notebooks with numerical examples for each chapter are found at Paul Söderlind’s Github page: <https://github.com/PaulSoderlind/FinancialEconometrics> All calculations in these notes are done in Julia and the plots generated by PyPlot/matplotlib.

When I first set up this course many years ago, I was inspired by the texts of Verbeek and Greene. More recently, I have enjoyed Hansen’s text. Most likely that shows.

My students at the MBF program at the University of St. Gallen have asked many good questions and pointed out mistakes. Also my teaching assistants did the same. I thank them for their input.

Data Sources

The data used in these lecture notes are from the following sources:

1. The website of Kenneth French,
<http://mba.tuck.dartmouth.edu/pages/faculty/ken.french>
2. Bloomberg
3. Datastream

4. Federal Reserve Bank of St. Louis (FRED),
<http://research.stlouisfed.org/fred2/>
5. The website of Robert Shiller,
<http://www.econ.yale.edu/~shiller/data.htm>
6. yahoo! finance, <http://finance.yahoo.com/>
7. OlsenData, <http://www.olsendata.com>

Contents

1 Review of Statistics	8
1.1 Random Variables and Distributions	8
1.2 Moments	15
1.3 Distributions Commonly Used in Tests	22
1.4 Normal Distribution of the Sample Mean	24
1.5 Appendix – Notation*	27
1.6 Appendix – Statistical Tables*	27
2 Least Squares Estimation	32
2.1 Least Squares: The Optimization Problem	32
2.2 Missing Data	44
2.3 The Distribution of $\hat{\beta}$	44
2.4 The Distribution of $\hat{\beta}$: More General Results	53
2.5 Appendix – A Primer in Matrix Algebra*	60
2.6 Appendix – A Primer in Calculus*	66
2.7 Appendix – A Primer in Optimization*	68
3 Betas and Index Models	72
3.1 Single-Index Models	72
3.2 Estimating Beta	73
3.3 Multi-Index Models	75
3.4 Principal Component Analysis*	77
4 Least Squares: Testing	80
4.1 Testing a Single Coefficient: A t -test	80
4.2 Confidence Bands	83
4.3 Power and Size*	84

4.4	Testing A Linear Combination	86
4.5	Joint Test of Several Coefficients	87
4.6	Confidence Bands for a Forecast*	91
4.7	Testing Nonlinear Hypotheses (Delta Method)	92
5	Least Squares: Non-iid Residuals	97
5.1	Heteroskedasticity	97
5.2	Autocorrelation	103
5.3	Cross-Sectional Correlations (“Clustering”)	108
6	A System of OLS Regressions	111
6.1	A System of Two OLS Regressions	111
6.2	A System of n OLS Regressions	113
7	Testing CAPM and Multifactor Models	116
7.1	Market Model	116
7.2	Several Factors	123
8	Model Selection and Other Topics	126
8.1	Model Selection I	126
8.2	Model Selection II	128
8.3	Weighted Least Squares	132
8.4	Comparing Non-Nested Models	134
8.5	Non-Linear Models	134
8.6	Outliers	136
8.7	Estimation on Subsamples	138
9	Asymptotic Results on OLS	144
9.1	Motivation of Asymptotics	144
9.2	Consistency	144
9.3	When OLS Is Inconsistent	148
9.4	Asymptotic Normality	154
9.5	Spurious Regressions	158
9.6	Appendix – Details on Demand and Supply*	164

10 Simulating the Finite Sample Properties	166
10.1 Introduction	166
10.2 Monte Carlo Simulations	166
10.3 Bootstrapping	171
11 Portfolio Sorts	176
11.1 Overview	176
11.2 Univariate Sorts	177
11.3 Bivariate Sorts	178
11.4 Orthogonalisation	183
12 Financial Panel Data	185
12.1 Introduction to Panel Data	185
12.2 Calendar Time Regressions	186
12.3 An Overview of Panel Data Models	187
12.4 Pooled OLS	188
12.5 The Within Estimator	193
12.6 The First-Difference Estimator	198
12.7 Differences-in-Differences Estimator	199
12.8 Fama-MacBeth	200
12.9 Appendix – Random Effects Model*	203
13 Instrumental Variables Method (IV)	205
13.1 Instrumental Variables Method	205
13.2 Two-stages-least squares (2SLS)	208
13.3 Hausman’s Specification Test	213
13.4 Appendix – Asymptotics of the IV and 2SLS Estimators*	215
14 GMM	217
14.1 The Basic GMM	217
14.2 GMM with a Suboptimal Weighting Matrix	223
14.3 GMM without a Loss Function	223
14.4 Appendix – Proofs	225
15 Time Series Analysis	226
15.1 Sample Autocorrelations	226

15.2 Stationarity	228
15.3 White Noise	229
15.4 Moving Average (MA)	230
15.5 Autoregressions	232
15.6 ARMA(p,q)	238
15.7 Approximating MA and ARMA Models with an AR Model	239
15.8 VAR(p)	241
15.9 VAR(1)	243
15.10 Non-stationary Processes	248
16 Predicting Asset Returns	256
16.1 Autocorrelations and Autoregression	256
16.2 Other Predictors and Methods	261
16.3 Out-of-Sample Forecasting Performance	263
16.4 Forecast Averaging	271
16.5 Evaluating Forecasting Performance	272
16.6 Security Analysts	275
17 Event Studies	280
17.1 Basic Structure of Event Studies	280
17.2 Models of Normal Returns	282
17.3 Testing the Abnormal Return	285
17.4 Quantitative Events	288
18 Distributions and Option Pricing	289
18.1 Estimating and Testing Distributions	289
18.2 Option Pricing	299
19 Maximum Likelihood Estimation	304
19.1 Maximum Likelihood	304
19.2 Key Properties of MLE	309
19.3 QMLE	312
20 ARCH and GARCH	314
20.1 Heteroskedasticity	314
20.2 ARCH Models	318

20.3 GARCH Models	321
20.4 Non-Linear Extensions	324
20.5 Multivariate (G)ARCH	325
21 Risk Measures	333
21.1 Value at Risk	333
21.2 Backtesting a VaR Model	336
21.3 Expected Shortfall	338
21.4 Semi-Variance and Maximum Drawdown	341
22 Non-Parametric Regressions	345
22.1 Kernel Regressions	345
22.2 Local Linear Regressions	351
23 LAD and Quantile Regressions	358
23.1 LAD	358
23.2 Quantile Regressions	360
24 Binary Choice and Truncated Models	366
24.1 Binary Choice Model	366
24.2 Truncated Regression Model	371
24.3 Censored Regression Model	373
24.4 A Sample Selection Model	376

Chapter 1

Review of Statistics

1.1 Random Variables and Distributions

1.1.1 The Distribution of a Random Variable

A univariate distribution of a random variable x describes the probability of different values. If $f(x)$ is the probability density function (pdf), then the probability that x is between A and B is calculated as the area under the density function from A to B

$$\Pr(A < x \leq B) = \int_A^B f(x)dx. \quad (1.1)$$

See Figure 1.1 for illustrations of normal (gaussian) distributions.

Remark 1.1 If $x \sim N(\mu, \sigma^2)$, then the probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

This is a bell-shaped curve centred on the mean μ and where the standard deviation σ determines the “width” of the curve. See Figure 1.1 for illustrations.

Remark 1.2 Notice that if $\phi(z)$ is the pdf of an $N(0, 1)$ variable z , then the pdf of a $N(\mu, \sigma^2)$ variable x can be calculated as $\phi((x - \mu)/h) / h$. The logarithm, $\ln \phi(x) = -\ln(2\pi)/2 - x^2/2$, is often used in maximum likelihood estimation.

The probability that $x \leq B$ (that is, $-\infty < x \leq B$) is measured by the *cumulative distribution function*, $\text{cdf}(B)$. For instance, if x has a $N(0, 1)$ distribution, then $\Pr(x \leq -1.645) = 0.05$ and $\Pr(x \leq 0) = 0.5$. Once you have the cdf, you can calculate the probability of $B < x$ as $1 - \text{cdf}(B)$. See Figure 1.2 for an illustration.

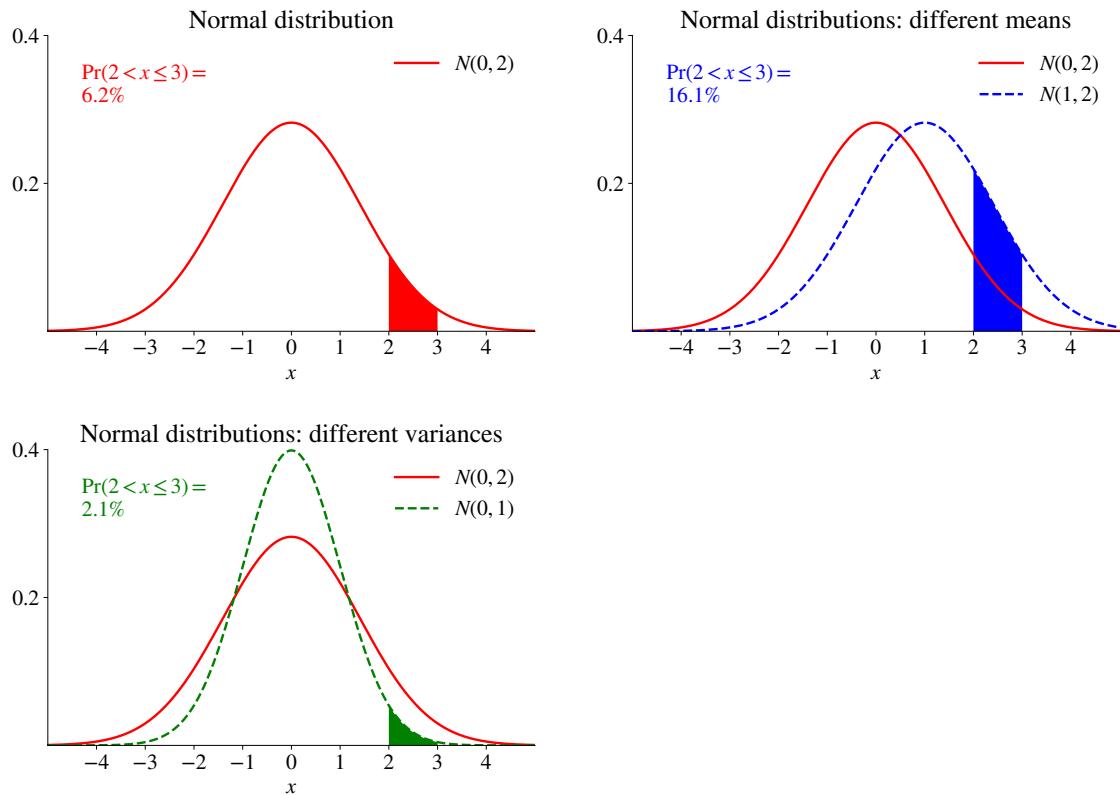


Figure 1.1: A few different normal distributions

If we invert the cdf, then we get the *quantiles* (percentiles) of the random variable. For instance, the 0.05th quantile (5th percentile) of a $N(0, 1)$ variable is -1.645 , while the 0.5th quantile (also called the median) is 0.

1.1.2 The Joint Distribution of Several Random Variables

A bivariate distribution of the random variables x and y contains the same information as the two respective univariate distributions, but also information on how x and y are related. Let $h(x, y)$ be the joint density function, then the probability that x is between A and B and y is between C and D is calculated as the volume under the surface of the bivariate density function

$$\Pr(A < x \leq B \text{ and } C < y \leq D) = \int_A^B \int_C^D h(x, y) dy dx. \quad (1.2)$$

See Figure 1.3 for an example of a bivariate density function.

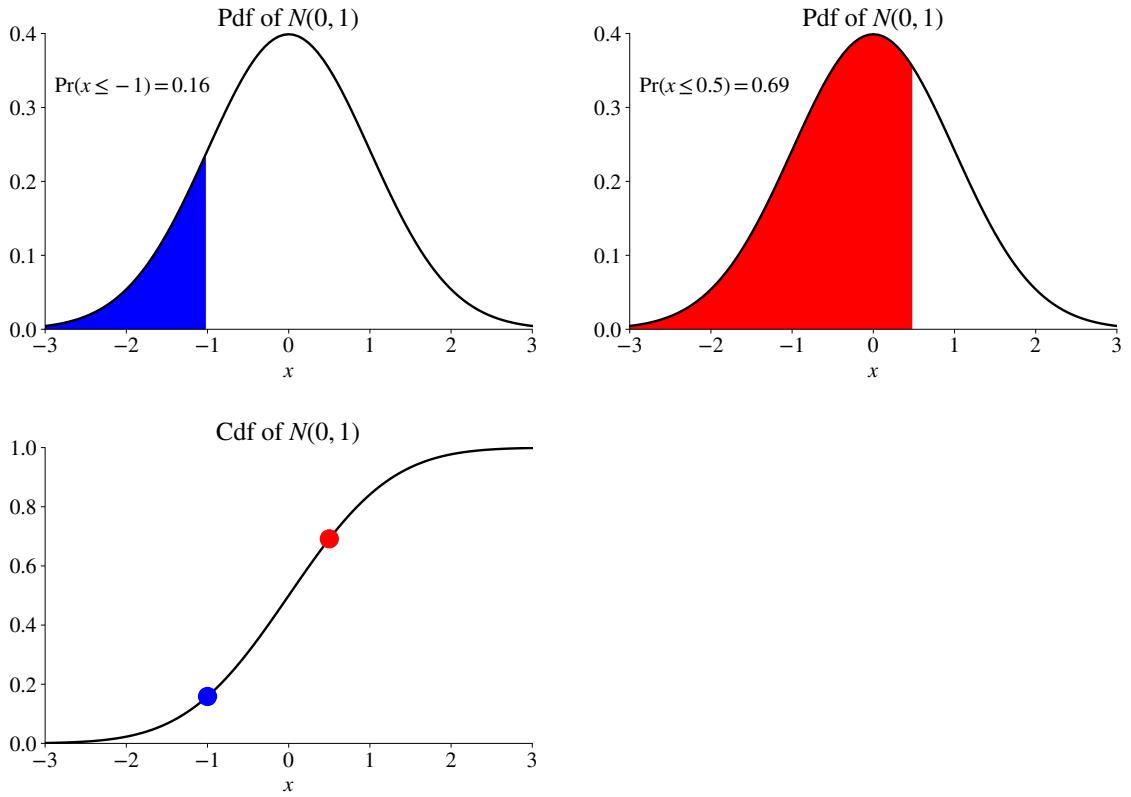


Figure 1.2: Pdf and cdf of $N(0, 1)$

A joint normal distribution is completely described by the means and the variance-covariance matrix

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \right), \quad (1.3)$$

where μ_x and μ_y denote the means of x and y , σ_x^2 and σ_y^2 denote the variances of x and y and σ_{xy} denotes their covariance. Sometimes alternative notations are used: $E x$ for the mean, $Std(x)$ for the standard deviation, $Var(x)$ for the variance and $Cov(x, y)$ for the covariance. See Figure 1.3 for an example.

Clearly, if the covariance σ_{xy} is zero, then the variables are (linearly) unrelated to each other. Otherwise, information about x can help us to make a better estimate/prediction of y . The correlation of x and y is defined as

$$\rho_{xy} = \sigma_{xy}/(\sigma_x \sigma_y). \quad (1.4)$$

The correlation is important for the shape of the distribution, again see Figure 1.3, which

shows both the bivariate pdf and its contours. The latter is sometimes easier to display.

If two random variables are independent of each other, then the joint density function is just the product of the two univariate densities (here denoted $f(x)$ and $k(y)$)

$$h(x, y) = f(x) k(y) \text{ if } x \text{ and } y \text{ are independent.} \quad (1.5)$$

This is useful in many cases, for instance, when we construct likelihood functions for maximum likelihood estimation.

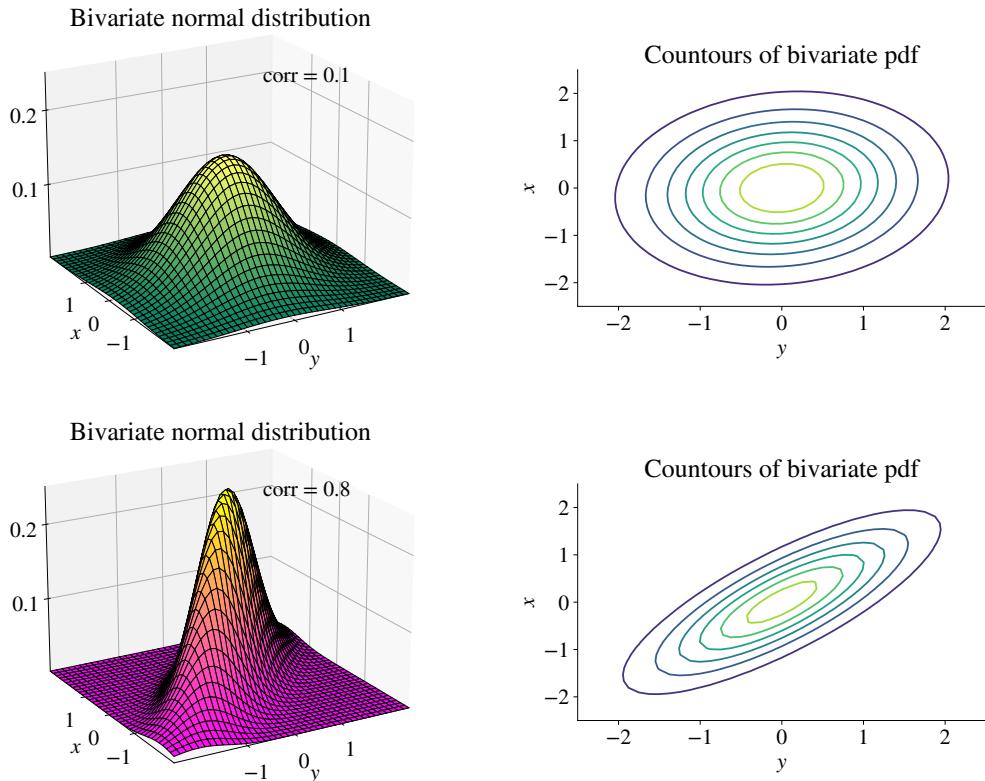


Figure 1.3: Density function bivariate normal distributions

1.1.3 Conditional Distributions*

If $h(x, y)$ is the joint density function and $f(x)$ the (marginal) density function of x , then the conditional density function is

$$g(y|x) = h(x, y)/f(x). \quad (1.6)$$

Notice that the conditional mean can be interpreted as the best estimate/prediction of y given that we know x . Similarly, the conditional variance can be interpreted as the variance of the forecast error (using the conditional mean as the forecast). The conditional and marginal distribution coincide if x and y are independent. (This follows directly from combining (1.5) and (1.6).)

For the bivariate normal distribution (1.3) we have the distribution of y conditional on a given value of x as

$$y|x \sim N(\mu_y + \beta(x - \mu_x), \sigma^2), \text{ with } \beta = \sigma_{xy}/\sigma_x^2 \text{ and } \sigma^2 = \sigma_y^2 - \sigma_{xy}\sigma_{xy}/\sigma_x^2. \quad (1.7)$$

In this case, the mean depends on x , while the variance does not. Also notice that the variance is lower than in the unconditional distribution (we have more information). Independence of x and y would here mean a zero covariance: set $\sigma_{xy} = 0$ in (1.7) to see that the conditional and unconditional distributions coincide. See Figure 1.4 for an illustration based on the bivariate distributions displayed in Figure 1.3. Notice how the location of the conditional distribution of y changes as a function of the correlation and the value of x . In contrast, the width changes as a function of the correlation, not x .

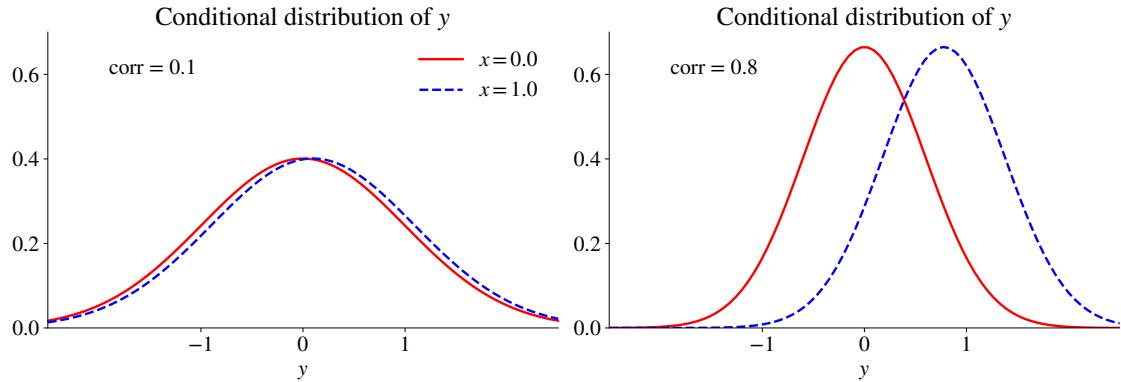


Figure 1.4: Conditional density functions of normal distributions, based on the bivariate distributions in Figure 1.3.

Remark 1.3 (*Relation of (1.7) to a linear regression**) Suppose you regress $y = a + bx + u$. The mean in (1.7) is the same as $a + bx$ and the variance is the same as $\text{Var}(u)$.

1.1.4 Illustrating a Distribution

If we know the type of distribution (uniform, normal, etc) a variable has, then the best way of illustrating the distribution is to estimate its parameters (mean, variance and whatever more—see below) and then draw the density function.

In case we are not sure about which distribution to use, the first step is typically to draw a histogram: it shows the relative frequencies for different bins (intervals). For instance, it could show the relative frequencies of a variable x being in each of the follow intervals: -0.5 to 0, 0 to 0.5 and 0.5 to 1.0. Clearly, the relative frequencies should sum to unity (or 100%), but they are sometimes normalized so the area under the histogram has an area of unity, similar to a probability density function.

Empirical Example 1.4 (*Histogram of equity returns*) See Figure 1.5.

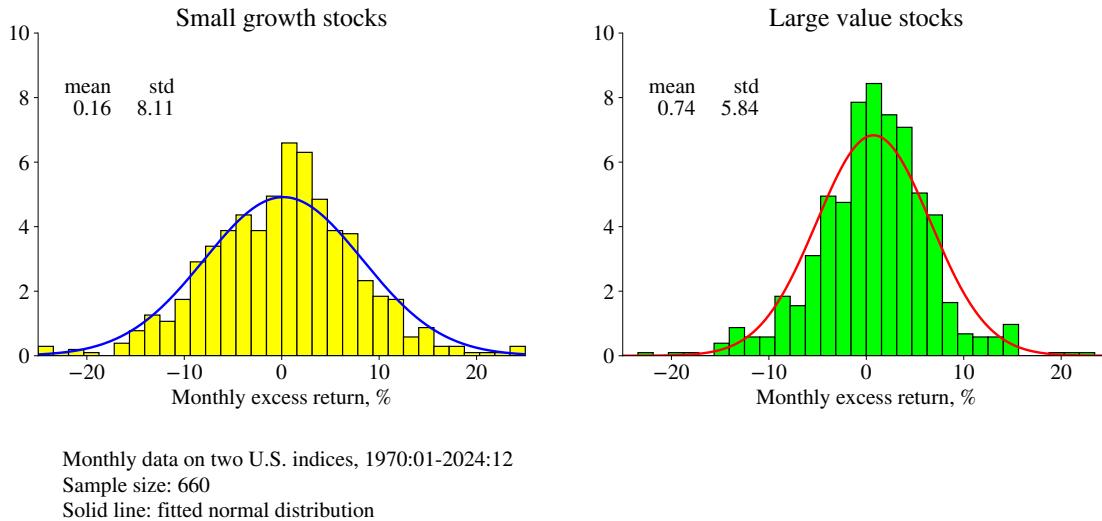


Figure 1.5: Histogram of returns, the curve is a normal distribution with the same mean and standard deviation as the return series

1.1.5 Confidence Bands and t-tests

For a symmetric distribution, a 90% (two-sided) confidence band is constructed by finding a critical value c such that

$$\Pr(\mu - c < x \leq \mu + c) = 0.9. \quad (1.8)$$

Replace 0.9 by 0.95 to get a 95% confidence band—and similarly for other confidence levels. In particular, if $x \sim N(\mu, \sigma^2)$, then

$$\begin{aligned}\Pr(\mu - 1.64\sigma < x \leq \mu + 1.64\sigma) &= 0.9 \text{ and} \\ \Pr(\mu - 1.96\sigma < x \leq \mu + 1.96\sigma) &= 0.95.\end{aligned}\tag{1.9}$$

As an example, suppose x is not a data series but a regression coefficient (denoted $\hat{\beta}$)—and we know that the standard error equals some number σ . We could then construct a 90% confidence band around the point estimate ($\hat{\beta}$) as

$$[\hat{\beta} - 1.64\sigma, \hat{\beta} + 1.64\sigma],\tag{1.10}$$

provided $\hat{\beta}$ is normally distributed. In case this band does not include the null hypothesis $\beta = q$ ($q = 0$ is a commonly used special case), then we would be at least 90% sure that the (true) regression coefficient is different from q .

Alternatively, suppose we instead construct the 90% confidence band around q as

$$[q - 1.64\sigma, q + 1.64\sigma].\tag{1.11}$$

If this band does not include the point estimate ($\hat{\beta}$), then we are also at least 90% sure that the (true) regression coefficient is different from q .

A third way to create a confidence band is to first create a standardized variable

$$t = (\hat{\beta} - q)/\sigma,\tag{1.12}$$

and then notice that we are 90% sure that t is in the interval

$$[-1.64, 1.64].\tag{1.13}$$

(Provided the null hypothesis is true, that is, $\beta = q$.) This is a t -test. Testing the null hypothesis by using (1.10), (1.11) or (1.13) should give the same answer to the question: is there sufficient statistical evidence against the null hypothesis.

1.1.6 The Idea behind Confidence Bands and t-tests

Suppose we have estimated a parameter ($\hat{\beta}$) from a particular sample of data (observations 1 to T). The parameter could, for instance, be the mean or a regression coefficient. This estimate is a random variable, since the sample is randomly drawn. If we are willing to

assume that data for every possible sample would have similar statistical properties, then we can estimate the confidence band as in (1.10) to describe the dispersion *across potential samples*. For instance, the variance of samples averages is (under the assumption of iid data) equal to the variance of data divided by the sample length.

1.1.7 Hypothesis Testing

We are here interested in testing the null hypothesis that $\beta = q$, where q is a number of interest (0.27, say). A null hypothesis is often denoted H_0 . (Econometric programs often automatically report results for $H_0: \beta = 0$.) We here consider the alternative hypothesis (denoted H_1 or perhaps H_A) that $\beta \neq q$. This leads to a two-sided (or two-tailed) test.

Typically, we assume that the estimates are normally distributed. To be able to easily compare with printed tables of probabilities, transform to a $N(0, 1)$ variable. In particular, if the true coefficient is really q , then $\hat{\beta} - q$ should have a zero mean (recall that $E \hat{\beta}$ equals the true value). Dividing by the standard error (deviation) of $\hat{\beta}$, we should have

$$t = (\hat{\beta} - q) / \text{Std}(\hat{\beta}) \sim N(0, 1) \quad (1.14)$$

In case $|t|$ is very large (say, 1.64 or larger), then our estimate $\hat{\beta}$ is a very unlikely outcome if $E \hat{\beta}$ (which equals the true coefficient value, β) is indeed q . We therefore draw the conclusion that the true coefficient is not q , that is, we *reject the null hypothesis*.

1.2 Moments

1.2.1 Expected Value and the Variance of a Random Variable

The expected value (or mean) of a random variable x is defined as

$$E x = \sum_{s=1}^S \pi_s x_s \text{ or } \int f(x) x dx, \quad (1.15)$$

for a discrete and continuous random variable, respectively. For the former, π_s denotes the probability of outcome x_s , and for the latter $f(x)$ represents the probability density function (pdf). The probabilities must sum to unity; therefore $\sum_{s=1}^S \pi_s = 1$ and $\int f(x) dx = 1$. The expected value is sometimes denoted μ .

The expectation can be extended to a function $g(x)$ of the random variable as

$$E g(x) = \sum_{s=1}^S \pi_s g(x_s) \text{ or } \int f(x) g(x) dx. \quad (1.16)$$

A typical case is $g(x) = (x - \mu)^2$, which gives the variance

$$\text{Var}(x) = \sum_{s=1}^S \pi_s (x_s - \mu)^2 \text{ or } \int f(x)(x - \mu)^2 dx. \quad (1.17)$$

We often use σ^2 to denote the variance. The standard deviation is the square root of the variance, $\text{Std}(x) = \text{Var}(x)^{1/2}$.

In most of the portfolio analysis, these concepts refer to the expectations of investors as of the time of investment. This implies that they may change over time and also differ from the historical average returns.

If a and b are two constants, then the previous expressions directly show that

$$E(a + bx) = a + b E x \quad (1.18)$$

$$\text{Var}(a + bx) = b^2 \text{Var}(x). \quad (1.19)$$

1.2.2 Estimates of the Mean and Standard Deviation

The mean and variance of a series are estimated as

$$\bar{x} = \sum_{t=1}^T x_t / T \text{ and } \hat{\sigma}^2 = \sum_{t=1}^T (x_t - \bar{x})^2 / T. \quad (1.20)$$

The standard deviation (the square root of the variance) is the most common measure of volatility. (Sometimes we use $T - 1$ in the denominator of the sample variance instead of T .) See Figure 1.5 for an illustration.

A sample mean is normally distributed if x_t is normally distributed, $x_t \sim N(\mu, \sigma^2)$. The reason is that a linear combination of normally distributed variables is (typically) also normally distributed. However, a sample average is often approximately normally distributed even if the variable is not (discussed below). If x_t is iid (independently and identically distributed), then the variance of a sample mean is

$$\text{Var}(\bar{x}) = \sigma^2 / T, \text{ if } x_t \text{ is iid.} \quad (1.21)$$

Sometimes, we instead report the variance of $\sqrt{T}\bar{x}$, which is

$$\text{Var}(\sqrt{T}\bar{x}) = \sigma^2, \text{ if } x_t \text{ is iid.} \quad (1.22)$$

(To see the link, factor out T from the left hand side of (1.22) to get $T \text{Var}(\bar{x})$. Then divide by sides by T to get (1.21).)

A sample average is (typically) *unbiased*, that is, the expected value of the sample

average equals the population mean, that is,

$$\mathbb{E} \bar{x} = \mathbb{E} x_t = \mu. \quad (1.23)$$

Since sample averages tend to be normally distributed in large samples, we have

$$\bar{x} \sim N(\mu, \sigma^2/T), \quad (1.24)$$

so we can construct a *t-stat* as

$$t = (\bar{x} - \mu)/(\sigma/\sqrt{T}), \quad (1.25)$$

which has an $N(0, 1)$ distribution. An alternative way of expressing (1.24) is

$$\sqrt{T}(\bar{x} - \mu) \sim N(0, \sigma^2). \quad (1.26)$$

This is often used in the context of the central limit theorem.

Proof (of (1.21)–(1.23)) To prove (1.21), notice that

$$\begin{aligned} \text{Var}(\bar{x}) &= \text{Var}(\sum_{t=1}^T x_t / T) \\ &= \sum_{t=1}^T \text{Var}(x_t / T) \\ &= T \text{Var}(x_t) / T^2 \\ &= \sigma^2 / T. \end{aligned}$$

The first equality is just a definition and the second equality follows from the assumption that x_t and x_s are independently distributed. This means, for instance, that $\text{Var}(x_2 + x_3) = \text{Var}(x_2) + \text{Var}(x_3)$ since the covariance is zero. The third equality follows from the assumption that x_t and x_s are identically distributed (so their variances are the same). The fourth equality is a trivial simplification.

To prove (1.23)

$$\begin{aligned} \mathbb{E} \bar{x} &= \mathbb{E} \sum_{t=1}^T x_t / T \\ &= \sum_{t=1}^T \mathbb{E} x_t / T \\ &= \mathbb{E} x_t. \end{aligned}$$

The first equality is just a definition and the second equality is always true (the expectation of a sum is the sum of expectations), and the third equality follows from the assumption of identical distributions which implies identical expectations. \square

1.2.3 Skewness and Kurtosis

The skewness, kurtosis and Jarque-Bera test for normality are useful diagnostic tools. First, let $z_t = (x_t - \mu)/\sigma$ and then note

	Test statistic	Distribution	
skewness	$\sum_{t=1}^T z_t^3 / T$	$N(0, 6/T)$	(1.27)
kurtosis	$\sum_{t=1}^T z_t^4 / T$	$N(3, 24/T)$	
Jarque-Bera	$(T/6)\text{skewness}^2 + (T/24)(\text{kurtosis} - 3)^2$	χ_2^2 .	

This is implemented by using the estimated mean and standard deviation. See Figure 1.5 for an illustration.

The distributions stated on the right hand side of (1.27) are under the null hypothesis that x_t is iid $N(\mu, \sigma^2)$. The “excess kurtosis” is defined as the kurtosis minus 3. The test statistic for the normality test (Jarque-Bera) can be compared with 4.6 or 6.0, which are the 10% and 5% critical values of a χ_2^2 distribution.

Clearly, we can test the skewness or kurtosis by traditional t-stats as in

$$t = \text{skewness}/\sqrt{6/T} \text{ and } t = (\text{kurtosis} - 3)/\sqrt{24/T}, \quad (1.28)$$

which both have $N(0, 1)$ distribution under the null hypothesis of a normal distribution.

1.2.4 Covariance and Correlation

A covariance of two variables (here x and y) is defined as

$$\begin{aligned} \text{Cov}(x, y) &= E[(x - E x)(y - E y)] \\ &= E xy - E x E y. \end{aligned} \quad (1.29)$$

If $E x = 0$ or $E y = 0$, then $\text{Cov}(x, y) = E xy$. When $x = y$, then we get $\text{Var}(x) = E x^2 - (E x)^2$. These results hold for sample moments too.

The variance of a linear combination of x and y is

$$\text{Var}(a + bx + cy) = b^2 \text{Var}(x) + c^2 \text{Var}(y) + 2bc \text{Cov}(x, y), \quad (1.30)$$

where (a, b, c) are constants.

The covariance is typically estimated as

$$\hat{\sigma}_{xy} = \sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y}) / T. \quad (1.31)$$

(Sometimes we use $T - 1$ in the denominator of the sample covariance instead of T .)

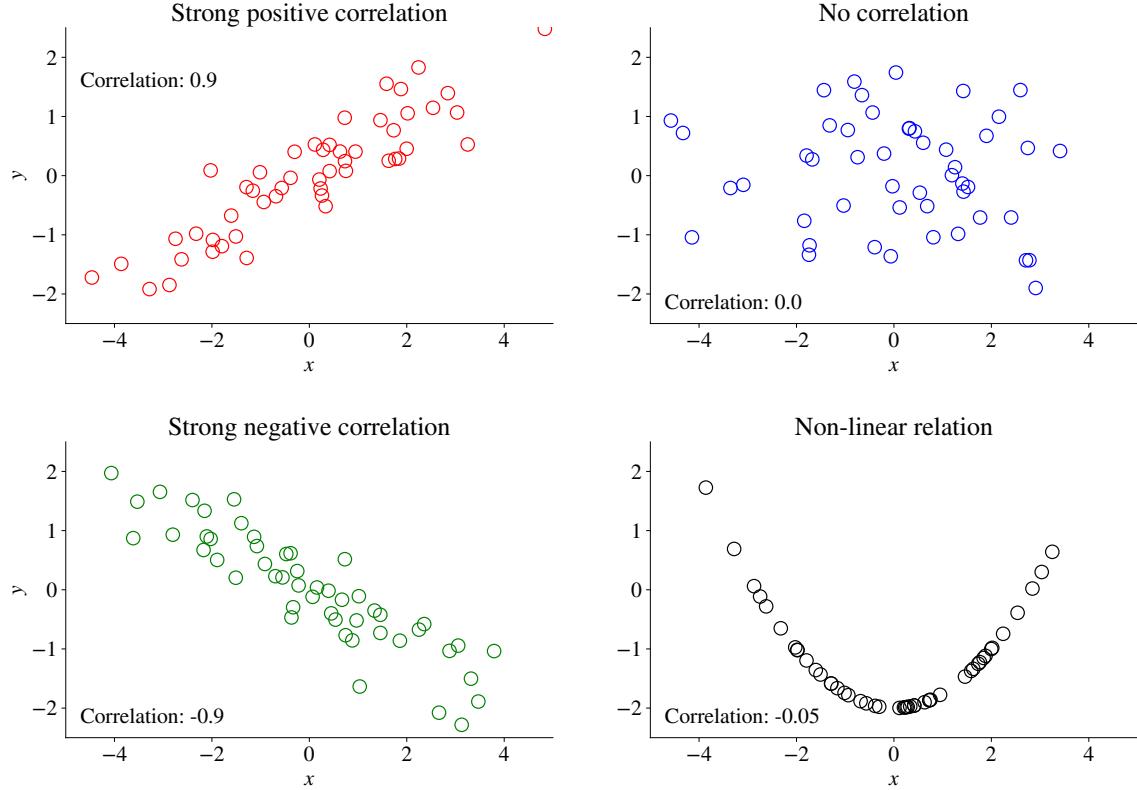


Figure 1.6: Example of correlations.

The correlation of two variables is then estimated as

$$\hat{\rho}_{xy} = \hat{\sigma}_{xy}/(\hat{\sigma}_x \hat{\sigma}_y), \quad (1.32)$$

where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the estimated standard deviations. A correlation must be between -1 and 1 . Note that covariance and correlation measure the degree of *linear* relation only. This is illustrated in Figure 1.6.

Empirical Example 1.5 (*Scatter plot of equity returns*) See Figure 1.7.

Under the null hypothesis of no correlation, and if the data is approximately normally distributed, then

$$\hat{\rho}/\sqrt{1 - \hat{\rho}^2} \sim N(0, 1/T), \quad (1.33)$$

so we can form a t-stat as

$$t = \sqrt{T} \hat{\rho} / \sqrt{1 - \hat{\rho}^2}, \quad (1.34)$$

which has an $N(0, 1)$ distribution.

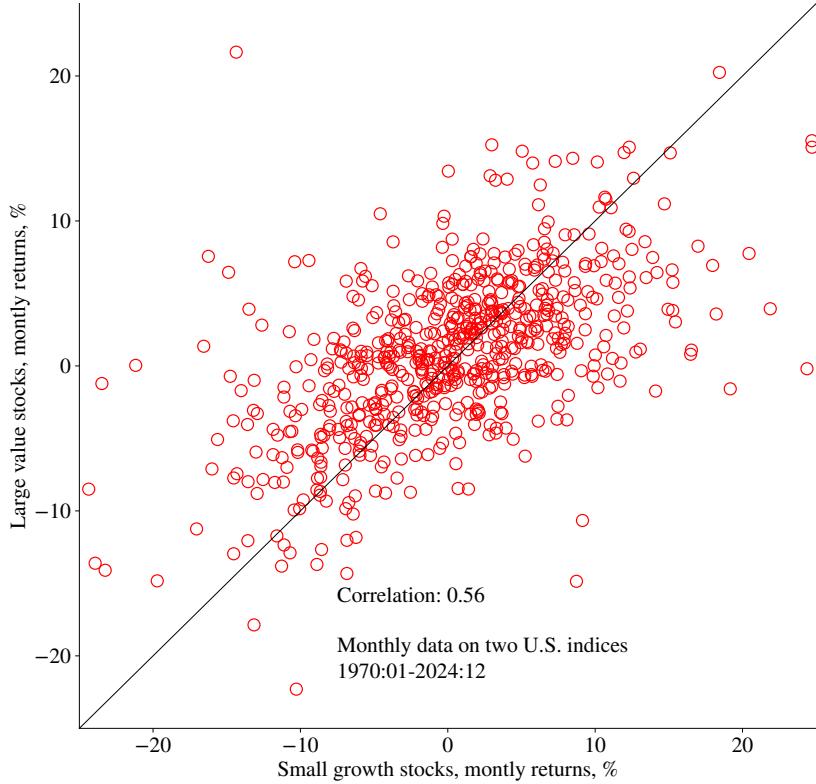


Figure 1.7: Scatter plot of two different portfolio returns

1.2.5 Moments of Vectors

Moments of vectors are straightforward extensions of the previous discussion. For instance, if $x = [x_1, x_2]$ is a vector of the two random variables x_1 and x_2 (the subscripts here indicate different variables), then the mean of x is just a vector (with 2 elements) of the means of the two variables

$$\mathbb{E}x = \begin{bmatrix} \mathbb{E}x_1 \\ \mathbb{E}x_2 \end{bmatrix}. \quad (1.35)$$

Also, the (2×2) variance-covariance matrix of x is

$$\text{Var}(x) = \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) \end{bmatrix}. \quad (1.36)$$

Clearly, the matrix is symmetric (since the two covariances are the same).

Also, if $y = [y_1, y_2, y_3]$ is a vector of the three random variables y_1 , y_2 and y_3 , then the (2×3) covariance matrix of the x and y vectors is

$$\text{Cov}(x, y) = \begin{bmatrix} \text{Cov}(x_1, y_1) & \text{Cov}(x_1, y_2) & \text{Cov}(x_1, y_3) \\ \text{Cov}(x_2, y_1) & \text{Cov}(x_2, y_2) & \text{Cov}(x_2, y_3) \end{bmatrix}. \quad (1.37)$$

This is not a symmetric matrix, and not even the first two columns define a symmetric matrix.

1.2.6 Correlations vs. Causality

Notice that a correlation between x and y does not say anything about causality. There are several possibilities, including

$$\begin{aligned} (x, \varepsilon) &\Rightarrow y \\ (y, u) &\Rightarrow x \\ (z, u) &\Rightarrow x \text{ and } (z, \varepsilon) \Rightarrow y \end{aligned} \quad (1.38)$$

In the first case, x and some other variables (here labelled ε) are indeed causing y , so changes in x are likely to be accompanied by changes in y . The second case shows the opposite: y is causing x . The third case is when some other variable z is driving the correlation between x and y . However, an independent move in x (due to u) will not lead to moves in y . Therefore, regressions (which are based on correlation analysis) need to be combined with other types of information to explore causality. In contrast, forecasting models are more focused on the correlation as such.

1.2.7 Correlations and the Variance of a Sample Average

The result in (1.21) that $\text{Var}(\bar{x}) = \sigma^2/T$ does not hold if x_t and x_{t-s} are correlated. To see that, consider the case when the sample has just two observations (indicated by subscripts)

$$\begin{aligned} \bar{x} &= (x_1 + x_2)/2 \text{ and} \\ \text{Var}(\bar{x}) &= (\sigma^2 + \sigma^2 + \sigma_{12} + \sigma_{21})/4. \end{aligned} \quad (1.39)$$

In the iid case we *assume* that $\sigma_{12} = \sigma_{21} = 0$, so $\text{Var}(\bar{x}) = \sigma^2/2$. In the other extreme case of perfect correlations, $\sigma_{12} = \sigma^2$ so the variance of the sample average is the same as for the data (no precision is gained by averaging), $\text{Var}(\bar{x}) = \sigma^2$.

More generally, with T observations, we have

$$\text{Var}(\bar{x}) = \sum_{i=1}^T \sum_{j=1}^T \sigma_{ij} / T^2, \quad (1.40)$$

which is sum of all the elements of the covariance matrix, divided by T^2 . This can be written as

$$\text{Var}(\bar{x}) = (\bar{\sigma}^2 - \bar{\sigma}_{ij}) / T + \bar{\sigma}_{ij}, \quad (1.41)$$

where $\bar{\sigma}^2$ is the average variance and $\bar{\sigma}_{ij}$ the average covariance of any two observations (x_t and x_{t-s} , say). This insight carries over to the variance of regression coefficients—when the residuals are correlated (over time or over cross sectional units).

Example 1.6 (Covariance matrix with $T = 2$) The covariance matrix of x_1 and x_2 is

$$\begin{bmatrix} \sigma^2 & \sigma_{12} \\ \sigma_{21} & \sigma^2 \end{bmatrix},$$

if we assume that x_1 and x_2 have the same variance (σ^2). Also, notice that $\sigma_{12} = \sigma_{21}$.

Example 1.7 ($\text{Var}(\bar{x})$) Assume $\bar{\sigma}^2 = 1$, then $\text{Var}(\bar{x})$ is

$$\begin{array}{ccc} \bar{\sigma}_{ij} = 0 & \bar{\sigma}_{ij} = 0.10 \\ \hline T = 10 & 0.1 & 0.19 \\ T = 100 & 0.01 & 0.109 \end{array}$$

1.3 Distributions Commonly Used in Tests

1.3.1 Standard Normal Distribution, $N(0, 1)$

Suppose the random variable x has a $N(\mu, \sigma^2)$ distribution. Then, the the *standardized variable* $(x - \mu)/\sigma$ has a standard normal distribution

$$t = (x - \mu)/\sigma \sim N(0, 1). \quad (1.42)$$

To see this, notice that $x - \mu$ has a mean of zero and that x/σ has a standard deviation of unity.

1.3.2 t -distribution

If we instead need to estimate σ to use in (1.42), then the test statistic has t_n -distribution

$$t = (x - \mu)/\hat{\sigma} \sim t_n, \quad (1.43)$$

where n denotes the “degrees of freedom,” that is the number of observations minus the number of estimated parameters. For instance, if we have a sample with T data points and only estimate the mean, then $n = T - 1$.

The t -distribution has more probability mass in the tails than an $N(0, 1)$ distribution. It therefore gives a more “conservative” test (harder to reject the null hypothesis), but the difference vanishes as the degrees of freedom (sample size) increase. See Figure 1.8 for a comparison and Table 1.1 for critical values.

Example 1.8 (t -distribution) If $t = 2.0$ and $n = 50$, then this is larger than the 10% critical value (but not the 5% critical value) for a 2-sided test in Table 1.1.

1.3.3 Chi-square Distribution

If $z \sim N(0, 1)$, then $z^2 \sim \chi_1^2$, that is, z^2 has a chi-square distribution with one degree of freedom. This can be generalized in several ways. For instance, if $x \sim N(\mu_x, \sigma_{xx})$ and $y \sim N(\mu_y, \sigma_{yy})$ and they are uncorrelated, then $[(x - \mu_x)/\sigma_x]^2 + [(y - \mu_y)/\sigma_y]^2 \sim \chi_2^2$.

More generally, we have

$$v' \Sigma^{-1} v \sim \chi_n^2, \text{ if the } n \times 1 \text{ vector } v \sim N(0, \Sigma). \quad (1.44)$$

See Figure 1.8 for an illustration and Table 1.2 for critical values.

Example 1.9 (χ_2^2 distribution) Suppose x is a 2×1 vector

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N \left(\begin{bmatrix} 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix} \right).$$

If $x_1 = 3$ and $x_2 = 5$, then

$$\begin{bmatrix} 3 - 4 \\ 5 - 2 \end{bmatrix}' \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 3 - 4 \\ 5 - 2 \end{bmatrix} \approx 6.1$$

has a χ_2^2 distribution. Notice that 6.1 is higher than the 5% critical value (but not the 1% critical value) in Table 1.2.

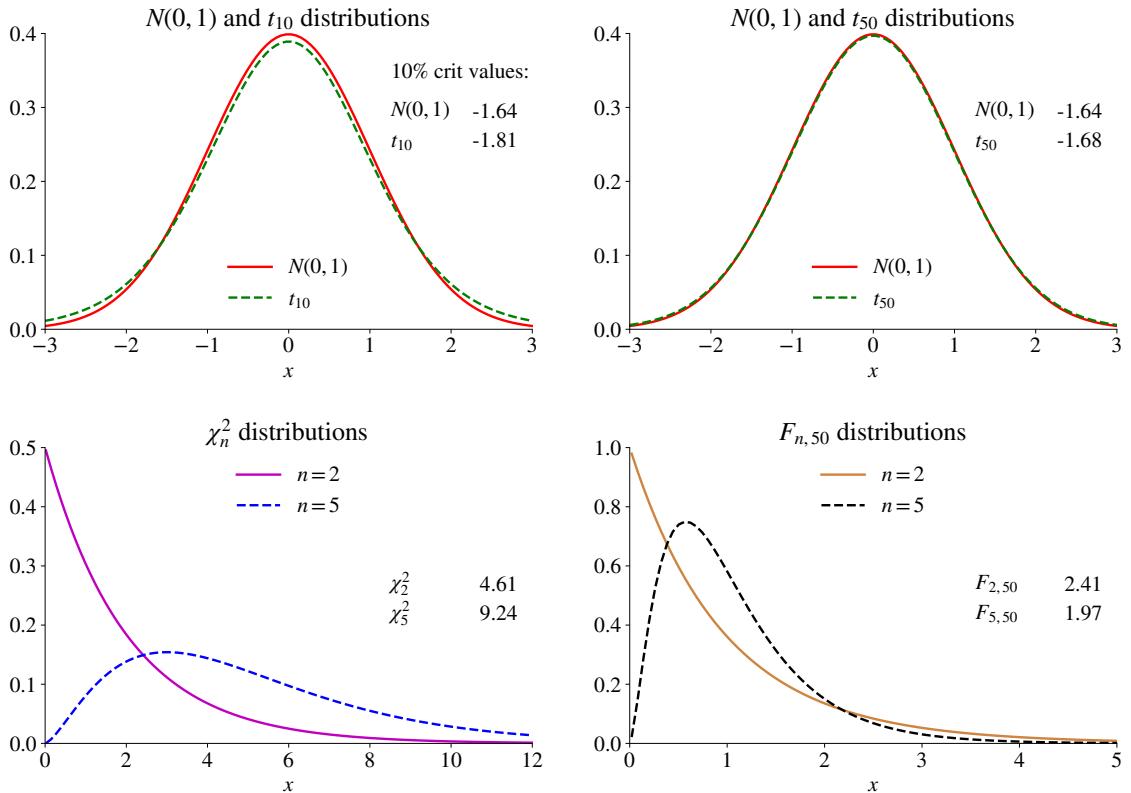


Figure 1.8: Probability density functions

1.3.4 F -distribution

If $x \sim \chi^2_{n_1}$ and $y \sim \chi^2_{n_2}$, then $(x/n_1)/(y/n_2)$ has an F_{n_1, n_2} distribution with (n_1, n_2) degrees of freedom. See Figure 1.8 for an illustration and Tables 1.3–1.4 for critical values.

1.4 Normal Distribution of the Sample Mean

In many cases, it is unreasonable to assume that a random variable x_t is normally distributed. The nice thing with a sample mean (or sample average), here denoted \bar{x} , is that it has very useful properties (in a reasonably large sample). This section gives a short summary of what happens to sample means as the sample size increases (often called “asymptotic theory”).

The *law of large numbers* (LLN) says that the sample mean converges to the true population mean as the sample size goes to infinity. This holds for a very large class

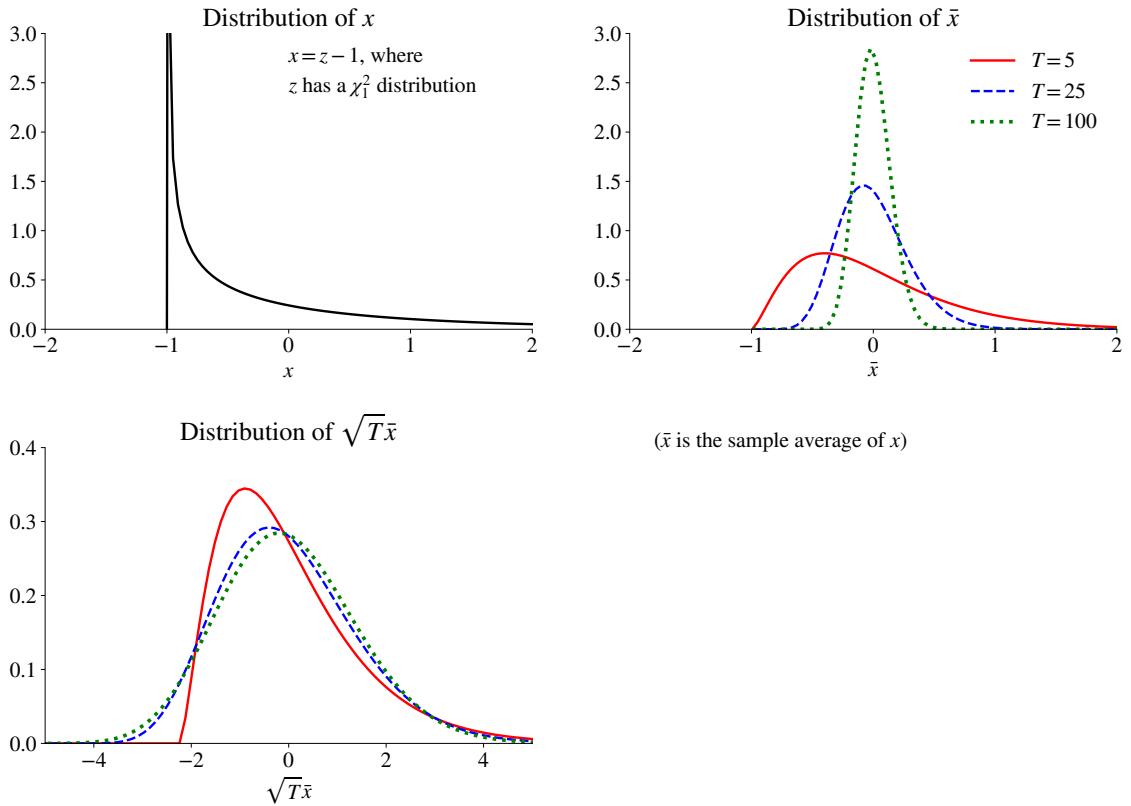


Figure 1.9: Sampling distributions

of random variables, but there are exceptions. A sufficient (but not necessary) condition for this convergence is that the sample average is unbiased (as in (1.23)) and that the variance goes to zero as the sample size goes to infinity (as in (1.21)). (This is also called convergence in mean square.) To see the LLN in action, see Figure 1.9.

The *central limit theorem* (CLT) says that $\sqrt{T}\bar{x}$ converges in distribution to a normal distribution as the sample size increases. See Figure 1.9 for an illustration. This also holds for a large class of random variables—and it is a very useful result since it allows us to test hypotheses by assuming that $\sqrt{T}\bar{x}$ is normally distributed. Most estimators (including least squares and other methods) are effectively some kind of sample average, so the CLT can potentially be applied. If the population mean is μ , then this is often written

$$\sqrt{T}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2), \quad (1.45)$$

where \xrightarrow{d} denotes that the distribution converges to a normal as T goes to infinity. If x is

iid, then σ^2 is the same as the variance of x . This is often expressed as

$$\bar{x} \xrightarrow{a} N(\mu, \sigma^2/T),$$

where \xrightarrow{a} means “is asymptotically distributed as”. (This follows from: (a) if the variance of $\sqrt{T}\bar{x}$ is σ^2 , then the variance of \bar{x} is σ^2/T ; (b) adding μ to $\bar{x} - \mu$ increases the mean by μ .)

Further Reading

See also Verbeek (2017) Appendix B for an introduction to statistics. Hansen (2022b) is a detailed treatment.

1.5 Appendix – Notation*

Here is an incomplete list of the notation used in these notes.

- $\Sigma_{t=1}^T x_t$ denotes the sum $x_1 + \dots + x_T$. In the running text, this is sometimes written as just $\Sigma_t x_t$. It also happens that Σ is used to denote a variance-covariance matrix. The distinction should be clear from the context.
- $\text{Var}(x)$ denotes either the variance of a single random variable, or the $n \times n$ variance-covariance matrix of a n -vector x .
- $\text{Cov}(x, z)$ denotes either the covariance of single random variables x and z , or the $n_x \times n_z$ matrix of covariances when x is an n_x -vector and z is an n_z -vector.
- $E x$ denotes the expectation (population mean) of x (which could be a single random variable, or a vector/matrix). When x is a vector/matrix, then $E x$ is the vector/matrix of expectation of each element of x . Sample means are typically denoted by \bar{x} .

1.6 Appendix – Statistical Tables*

<u>n</u>	Significance level		
	10%	5%	1%
10	1.81	2.23	3.17
20	1.72	2.09	2.85
30	1.70	2.04	2.75
40	1.68	2.02	2.70
50	1.68	2.01	2.68
60	1.67	2.00	2.66
70	1.67	1.99	2.65
80	1.66	1.99	2.64
90	1.66	1.99	2.63
100	1.66	1.98	2.63
Normal	1.64	1.96	2.58

Table 1.1: Critical values (two-sided test) of t distribution (different degrees of freedom) and normal distribution.

<u>n</u>	Significance level		
	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21

Table 1.2: Critical values of chisquare distribution (different degrees of freedom, n).

<u>n_1</u>	<u>n_2</u>					$\chi^2_{n_1}/n_1$
	10	30	50	100	300	
1	4.96	4.17	4.03	3.94	3.87	3.84
2	4.10	3.32	3.18	3.09	3.03	3.00
3	3.71	2.92	2.79	2.70	2.63	2.60
4	3.48	2.69	2.56	2.46	2.40	2.37
5	3.33	2.53	2.40	2.31	2.24	2.21
6	3.22	2.42	2.29	2.19	2.13	2.10
7	3.14	2.33	2.20	2.10	2.04	2.01
8	3.07	2.27	2.13	2.03	1.97	1.94
9	3.02	2.21	2.07	1.97	1.91	1.88
10	2.98	2.16	2.03	1.93	1.86	1.83

Table 1.3: 5% Critical values of F_{n_1, n_2} distribution (different degrees of freedom).

<u>n_1</u>	<u>n_2</u>					<u>$\chi^2_{n_1}/n_1$</u>
	10	30	50	100	300	
1	3.29	2.88	2.81	2.76	2.72	2.71
2	2.92	2.49	2.41	2.36	2.32	2.30
3	2.73	2.28	2.20	2.14	2.10	2.08
4	2.61	2.14	2.06	2.00	1.96	1.94
5	2.52	2.05	1.97	1.91	1.87	1.85
6	2.46	1.98	1.90	1.83	1.79	1.77
7	2.41	1.93	1.84	1.78	1.74	1.72
8	2.38	1.88	1.80	1.73	1.69	1.67
9	2.35	1.85	1.76	1.69	1.65	1.63
10	2.32	1.82	1.73	1.66	1.62	1.60

Table 1.4: 10% Critical values of F_{n_1,n_2} distribution (different degrees of freedom).

	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002
-2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
-2.8	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
-2.7	0.003	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.005
-2.6	0.005	0.005	0.005	0.005	0.005	0.005	0.006	0.006	0.006	0.006
-2.5	0.006	0.006	0.007	0.007	0.007	0.007	0.008	0.008	0.008	0.008
-2.4	0.008	0.008	0.009	0.009	0.009	0.009	0.010	0.010	0.010	0.010
-2.3	0.011	0.011	0.011	0.012	0.012	0.012	0.013	0.013	0.013	0.014
-2.2	0.014	0.014	0.015	0.015	0.015	0.016	0.016	0.017	0.017	0.017
-2.1	0.018	0.018	0.019	0.019	0.020	0.020	0.021	0.021	0.022	0.022
-2.0	0.023	0.023	0.024	0.024	0.025	0.026	0.026	0.027	0.027	0.028
-1.9	0.029	0.029	0.030	0.031	0.031	0.032	0.033	0.034	0.034	0.035
-1.8	0.036	0.037	0.038	0.038	0.039	0.040	0.041	0.042	0.043	0.044
-1.7	0.045	0.046	0.046	0.047	0.048	0.049	0.051	0.052	0.053	0.054
-1.6	0.055	0.056	0.057	0.058	0.059	0.061	0.062	0.063	0.064	0.066
-1.5	0.067	0.068	0.069	0.071	0.072	0.074	0.075	0.076	0.078	0.079
-1.4	0.081	0.082	0.084	0.085	0.087	0.089	0.090	0.092	0.093	0.095
-1.3	0.097	0.099	0.100	0.102	0.104	0.106	0.107	0.109	0.111	0.113
-1.2	0.115	0.117	0.119	0.121	0.123	0.125	0.127	0.129	0.131	0.133
-1.1	0.136	0.138	0.140	0.142	0.145	0.147	0.149	0.152	0.154	0.156
-1.0	0.159	0.161	0.164	0.166	0.169	0.171	0.174	0.176	0.179	0.181
-0.9	0.184	0.187	0.189	0.192	0.195	0.198	0.200	0.203	0.206	0.209
-0.8	0.212	0.215	0.218	0.221	0.224	0.227	0.230	0.233	0.236	0.239
-0.7	0.242	0.245	0.248	0.251	0.255	0.258	0.261	0.264	0.268	0.271
-0.6	0.274	0.278	0.281	0.284	0.288	0.291	0.295	0.298	0.302	0.305
-0.5	0.309	0.312	0.316	0.319	0.323	0.326	0.330	0.334	0.337	0.341
-0.4	0.345	0.348	0.352	0.356	0.359	0.363	0.367	0.371	0.374	0.378
-0.3	0.382	0.386	0.390	0.394	0.397	0.401	0.405	0.409	0.413	0.417
-0.2	0.421	0.425	0.429	0.433	0.436	0.440	0.444	0.448	0.452	0.456
-0.1	0.460	0.464	0.468	0.472	0.476	0.480	0.484	0.488	0.492	0.496

Table 1.5: Values of the standard normal cumulative distribution function at x where x is the sum of the values in the first column and the first row.

	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999

Table 1.6: Values of the standard normal cumulative distribution function at x where x is the sum of the values in the first column and the first row.

Chapter 2

Least Squares Estimation

2.1 Least Squares: The Optimization Problem

2.1.1 Simple Regression

The simple regression model is

$$y_t = \beta_0 + \beta_1 x_t + u_t, \text{ where } E u_t = 0 \text{ and } \text{Cov}(x_t, u_t) = 0, \quad (2.1)$$

where we can observe (have data on) the dependent variable y_t and the regressor x_t but not the residual u_t . The residual accounts for movements in y_t not explained by x_t . The subscript t refers to observation t , which could represent period t (when data is a time series) or investor t (when data is a cross-section). In the latter case, it is common to instead use i as subscript.

Remark 2.1 (*On notation*) These notes sometimes use alternative notations for the regression equation, for instance, $y_t = \alpha + \beta x_t + u_t$ (as is typical in CAPM regressions) or $y_i = a + b x_i + u_i$.

The two key assumptions in (2.1) are: (i) the mean of the residual is zero; and (ii) the residual is not correlated with the regressor, x_t . This implies that the residual represents pure noise. In contrast, if the average of u_t was non-zero, then $\beta_0 + \beta_1 x_t$ would get the general level of y_t wrong. Also, if x_t and u_t were correlated, then the best guess of y_t based on x_t would not be $\beta_0 + \beta_1 x_t$.

Suppose you do not know β_0 or β_1 , and that you have a sample of data: y_t and x_t for $t = 1, \dots, T$. The LS estimator of β_0 and β_1 minimizes the loss function

$$\sum_{t=1}^T (y_t - b_0 - b_1 x_t)^2 = (y_1 - b_0 - b_1 x_1)^2 + (y_2 - b_0 - b_1 x_2)^2 + \dots \quad (2.2)$$

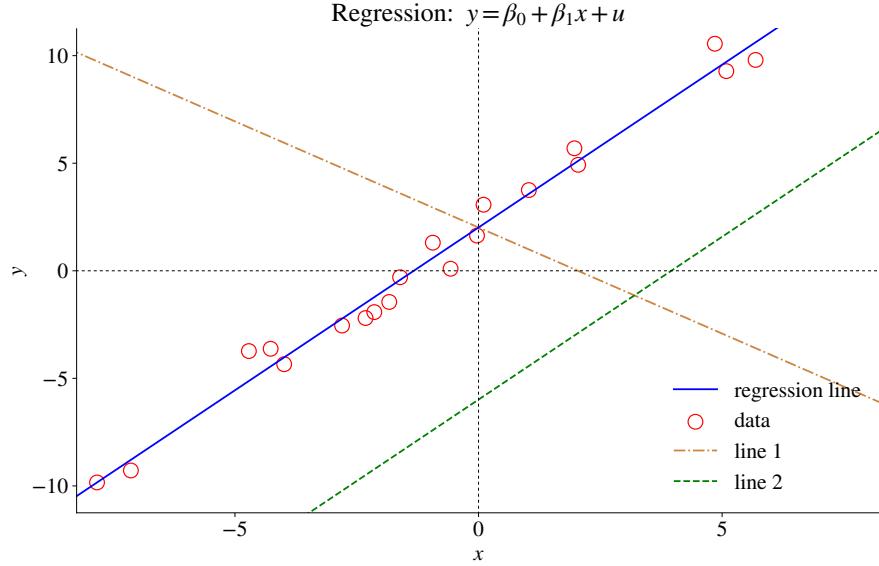


Figure 2.1: Example of OLS

by choosing b_0 and b_1 to make the loss function value as small as possible. This means fitting the model to the data as closely as possible, in the sense of smallest possible squared errors. See Figures 2.1 –2.2 for illustrations. (Other methods use different loss function, which will be discussed in later chapters.)

Remark 2.2 (*On notation*) These notes use $\sum_{t=1}^T x_t$ to denote the sum $x_1 + \dots + x_T$. The symbol Σ is also used to denote a variance-covariance matrix; the distinction should be clear from the context.

Remark 2.3 Note that β_i is the true (unobservable) value which we estimate to be $\hat{\beta}_i$. Whereas β_i is an unknown (deterministic) number, $\hat{\beta}_i$ is a random variable since it is calculated as a function of the random sample of y_t and x_t . We use b_i as an argument in the loss function (so we contemplate different values of b_i), and the optimal value is $\hat{\beta}_i$.

Remark 2.4 (*First order condition for minimizing a differentiable function*). We want to find the value of b , which makes the value of the differentiable function $f(b)$ as small as possible. The answer is a value of b where $df(b)/db = 0$. For a convex function like the quadratic, this value is unique. See Figure 2.3.

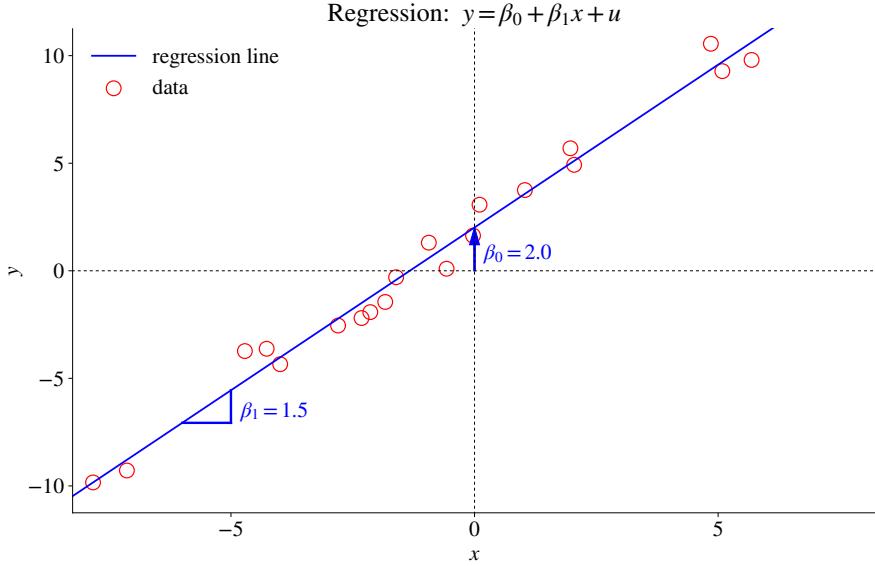


Figure 2.2: Example of OLS

The first order conditions for a minimum are that the derivatives of the loss function with respect to b_0 and b_1 equal zero. Notice that for a given observation t

$$\frac{\partial}{\partial b_0} (y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t) \quad (2.3)$$

$$\frac{\partial}{\partial b_1} (y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)x_t. \quad (2.4)$$

Let $(\hat{\beta}_0, \hat{\beta}_1)$ be the values of (b_0, b_1) where the derivatives of the loss function are zero

$$\frac{\partial}{\partial \beta_0} \sum_{t=1}^T (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2 = -2 \sum_{t=1}^T (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t) = 0 \quad (2.5)$$

$$\frac{\partial}{\partial \beta_1} \sum_{t=1}^T (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2 = -2 \sum_{t=1}^T x_t (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t) = 0, \quad (2.6)$$

which are two equations in two unknowns ($\hat{\beta}_0$ and $\hat{\beta}_1$), which must be solved simultaneously. These equations show that both the constant and x_t should be *orthogonal* to the fitted residuals, $\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t$. This characteristic of least squares reflects the sample analogues of the assumptions outlined in (2.1): $E u_t = 0$ and $\text{Cov}(x_t, u_t) = 0$. Equation (2.5) says that the sample average of \hat{u}_t should be zero. Similarly, (2.6) says that the sample cross moment of \hat{u}_t and x_t , that is, $\sum_{t=1}^T \hat{u}_t x_t / T$, should also be zero, which implies that the sample covariance is zero as well since \hat{u}_t has a zero sample mean (recall

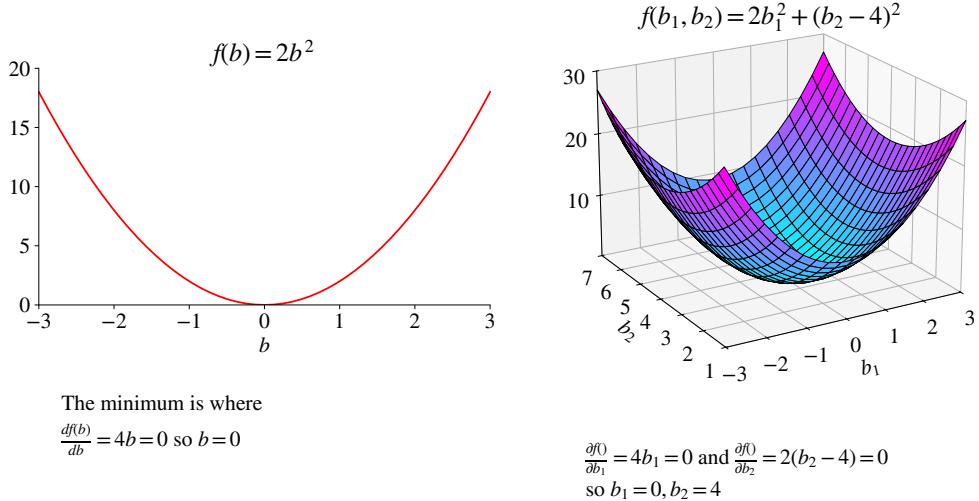


Figure 2.3: Quadratic loss function. Subfigure a: 1 coefficient; Subfigure b: 2 coefficients

that $\text{Cov}(x, u) = \text{E } xu - \text{E } x \text{ E } u$.

When the means of y and x are zero, then we know that intercept is zero ($\beta_0 = 0$). In this case, (2.6) with $\hat{\beta}_0 = 0$ immediately gives

$$\sum_{t=1}^T x_t y_t = \hat{\beta}_1 \sum_{t=1}^T x_t x_t \text{ or}$$

$$\hat{\beta}_1 = \frac{\sum_{t=1}^T x_t y_t / T}{\sum_{t=1}^T x_t x_t / T}. \quad (2.7)$$

In this case, the coefficient estimator is the sample covariance (recall: means are zero) of y_t and x_t , divided by the sample variance of the regressor x_t (this statement is actually true even if the means are not zero and a constant is included on the right hand side).

Empirical Example 2.5 (CAPM regressions) See Table 2.1 and Figure 2.5 for CAPM regressions for two industry portfolios. The betas clearly differ.

Example 2.6 (Simple regression) Consider the simple regression model (2.1). Suppose we have the following data

<u>t</u>	<u>x</u>	<u>y</u>
1	-1	-1.5
2	0	-0.6
3	1	2.1

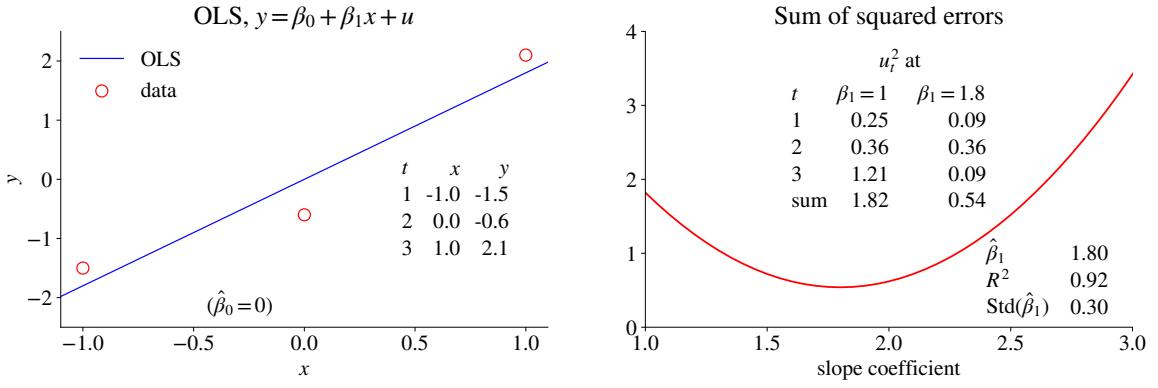


Figure 2.4: Example of OLS estimation

	HiTec	Utils
constant	-0.05 (-0.41)	0.23 (1.70)
market return	1.24 (39.02)	0.52 (14.57)
R^2	0.76	0.33
obs	660	660

Table 2.1: CAPM regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats.

To calculate the LS estimate according to (2.7), we note that

$$\begin{aligned}\sum_{t=1}^T x_t x_t &= (-1)^2 + 0^2 + 1^1 = 2 \text{ and} \\ \sum_{t=1}^T x_t y_t &= (-1)(-1.5) + 0(-0.6) + 1 \times 2.1 = 3.6\end{aligned}$$

This gives

$$\hat{\beta}_1 = 3.6/2 = 1.8.$$

The fitted residuals are

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.6 \\ 2.1 \end{bmatrix} - 1.8 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ -0.6 \\ 0.3 \end{bmatrix}.$$

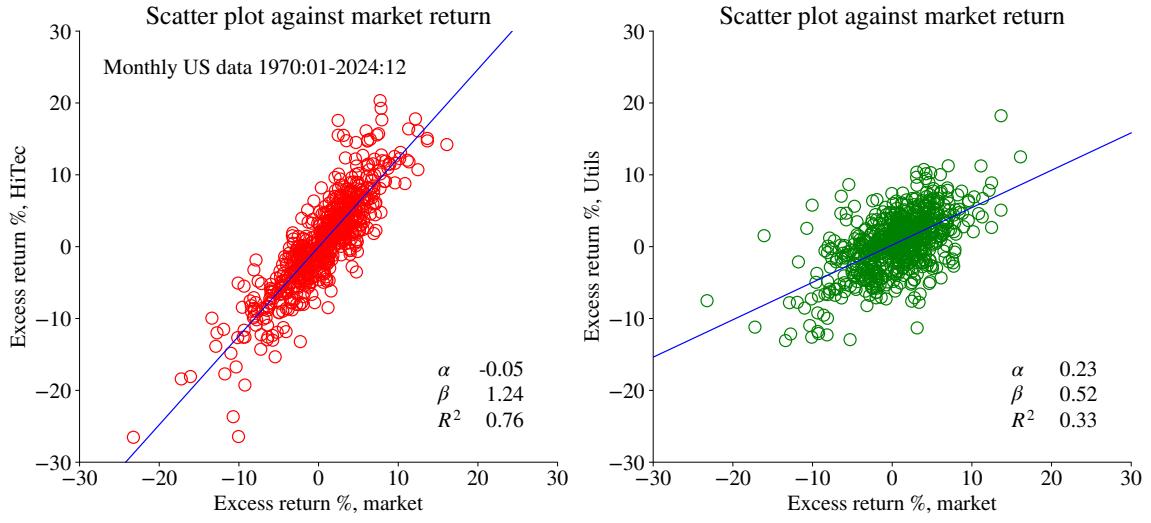


Figure 2.5: Scatter plot against market return

The fitted residuals indeed obey the first order conditions since the mean is zero and

$$\sum_{t=1}^T x_t \hat{u}_t = (-1) \times 0.3 + 0(-0.6) + 1 \times 0.3 = 0.$$

See Figure 2.4 for an illustration.

Example 2.7 Using the same data as in Example 2.6 we can also calculate the sums of squared residuals for different values of the slope coefficient. For instance, with $\beta = 1.8$ we get

t	\underline{u}_t	\underline{u}_1^2
1	$-1.5 - 1.8 \times (-1) = 0.3$	0.09
2	$-0.6 - 1.8 \times 0 = -0.6$	0.36
3	$2.1 - 1.8 \times 1 = 0.3$	0.09
sum	0	0.54

See Figure 2.4 for details. Among these alternatives, $\beta = 1.8$ has the lowest sum of squared residuals.

2.1.2 Multiple Regression

All the previous results hold also in a multiple regression, with suitable reinterpretations of the notation. Consider the linear model

$$\begin{aligned} y_t &= x_{1t}\beta_1 + x_{2t}\beta_2 + \cdots + x_{kt}\beta_k + u_t \\ &= x'_t\beta + u_t, \end{aligned} \tag{2.8}$$

where y_t and u_t are scalars, x_t a $k \times 1$ vector, and β is a $k \times 1$ vector of the true coefficients. In this expression, one of the elements of x_t is typically a constant equal to one (and the intercept is its coefficient).

Example 2.8 (of $x'_t\beta$)

$$x_t = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \beta = \begin{bmatrix} 0 \\ 1.8 \end{bmatrix} \rightarrow x'_t\beta = 1 \times 0 + (-1) \times 1.8 = -1.8$$

Remark 2.9 (On notation) These notes typically denote a vector of regression coefficients by β . The distinction from the $y_t = \alpha + \beta x_t + u_t$ notation (sometimes used for simple regressions) should be clear from the context.

Least squares minimizes the sum of the squared fitted residuals

$$\sum_{t=1}^T (y_t - x'_t b)^2, \tag{2.9}$$

by choosing the vector b . The first order conditions (zero derivatives) hold at the (optimal) values $\hat{\beta}$, and can then be written

$$\mathbf{0}_k = \sum_{t=1}^T x_t (y_t - x'_t \hat{\beta}), \tag{2.10}$$

where $\mathbf{0}_k$ denotes a k -vector of zeros. Solve this as

$$\hat{\beta} = \left(\sum_{t=1}^T x_t x'_t \right)^{-1} \sum_{t=1}^T x_t y_t. \tag{2.11}$$

If the regressors are orthogonal (for instance, $\sum_{t=1}^T x_{1t} x_{2t} = 0$) then the results from the multiple regression (2.11) are the same as those from a series of simple regressions: y_t regressed on x_{1t} , y_t regressed on x_{2t} , etc. (This is easy to see since in this case $\sum_{t=1}^T x_t x'_t$ is a diagonal matrix which carries over to the inverse.) This is an unlikely case, unless the regressors have been pre-processed to indeed be orthogonal.

Example 2.10 ($x_t = 1$) With a constant as the only regressor, $x_t y_t = y_t$ and $x_t x'_t = 1$. We then get $\hat{\beta} = \sum_{t=1}^T y_t / T$, that is, the sample average of y_t .

Empirical Example 2.11 (Factor model) Table 2.2 shows results from estimating a multi-factor model for the returns of two sector indices.

	HiTec	Utils
constant	0.14 (1.32)	0.11 (0.87)
market return	1.13 (38.20)	0.60 (17.75)
SMB	0.18 (3.76)	-0.20 (-4.08)
HML	-0.51 (-11.16)	0.31 (5.78)
R^2	0.82	0.41
obs	660	660

Table 2.2: Fama-French regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats.

Remark 2.12 (Matrix notation*) Let X be a $T \times k$ matrix where row t is filled with the elements of x_t and let Y be a $T \times 1$ vector where element t is y_t . Then, $X'X = \sum_{t=1}^T x_t x'_t$ and $X'Y = \sum_{t=1}^T x_t y_t$, so (2.11) can also be written $\hat{\beta} = (X'X)^{-1} X'Y$.

Example 2.13 (OLS with 2 regressors) With 2 regressors ($k = 2$) denoted x_{1t} and x_{2t} ,

$$x_t y_t = \begin{bmatrix} x_{1t} y_t \\ x_{2t} y_t \end{bmatrix} \text{ and } x_t x'_t = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \begin{bmatrix} x_{1t} & x_{2t} \end{bmatrix} = \begin{bmatrix} x_{1t} x_{1t} & x_{1t} x_{2t} \\ x_{2t} x_{1t} & x_{2t} x_{2t} \end{bmatrix}.$$

This means that (2.10) is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sum_{t=1}^T \begin{bmatrix} x_{1t}(y_t - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2) \\ x_{2t}(y_t - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2) \end{bmatrix}$$

and (2.11) is

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left(\sum_{t=1}^T \begin{bmatrix} x_{1t} x_{1t} & x_{1t} x_{2t} \\ x_{2t} x_{1t} & x_{2t} x_{2t} \end{bmatrix} \right)^{-1} \sum_{t=1}^T \begin{bmatrix} x_{1t} y_t \\ x_{2t} y_t \end{bmatrix}.$$

Example 2.14 (*OLS in terms of covariance matrices**) Consider the multiple regression $y_t = \alpha + z'_t \gamma + u_t$, where z_t is a vector with $k - 1$ elements. In terms of the first order conditions (2.10), we have $x_t = [1, z_t]$, that is (after dividing by T)

$$\begin{bmatrix} 0 \\ \mathbf{0}_{k-1} \end{bmatrix} = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} y_t - \hat{\alpha} - z'_t \hat{\gamma} \\ z_t(y_t - \hat{\alpha} - z'_t \hat{\gamma}) \end{bmatrix}.$$

The first line implies that $\hat{\alpha} = \bar{y}_t - \bar{z}'_t \hat{\gamma}$. Use this in the second line to replace $\hat{\alpha}$ and notice that it does not matter if the term outside the parenthesis is z_t or $z_t - \bar{z}_t$ (since the term in parenthesis is zero on average) to get

$$\begin{aligned} \mathbf{0}_{k-1} &= \frac{1}{T} \sum_{t=1}^T (z_t - \bar{z}_t)[(y_t - \bar{y}_t) - (z_t - \bar{z}_t)' \hat{\gamma}], \text{ or} \\ &= \widehat{\text{Cov}}(z_t, y_t) - \widehat{\text{Var}}(z_t) \hat{\gamma}, \end{aligned}$$

where $\widehat{\text{Var}}(z_t)$ denotes the sample variance-covariance matrix of z_t and $\widehat{\text{Cov}}(z_t, y_t)$ the vector of sample covariances of z_t and y_t . We can thus solve as $\hat{\gamma} = \widehat{\text{Var}}(z_t)^{-1} \widehat{\text{Cov}}(z_t, y_t)$ and $\hat{\alpha} = \bar{y}_t - \bar{z}'_t \hat{\gamma}$.

Example 2.15 (*Regression with an intercept and slope*) Suppose we have the following data:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.6 \\ 2.1 \end{bmatrix}, x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This is clearly the same as in Example 2.6, except that we allow for an intercept, which turns out to be zero in this particular example. The notation we need to solve this problem is the same as for a general multiple regression. Therefore, calculate the following:

$$\begin{aligned} \sum_{t=1}^T x_t x_t' &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
\sum_{t=1}^T x_t y_t &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-1.5) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (-0.6) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} 2.1 \\
&= \begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix} + \begin{bmatrix} -0.6 \\ 0 \end{bmatrix} + \begin{bmatrix} 2.1 \\ 2.1 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 3.6 \end{bmatrix}
\end{aligned}$$

To calculate the LS estimate, notice that the inverse of the $\Sigma_{t=1}^T x_t x'_t$ is

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix},$$

which can be verified by

$$\begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The LS estimate is therefore

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{t=1}^T x_t x'_t \right)^{-1} \sum_{t=1}^T x_t y_t \\
&= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 3.6 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 1.8 \end{bmatrix}.
\end{aligned}$$

The Frisch-Waugh-Lovell theorem*

Split up x_t into the vectors x_{1t} and x_{2t} and write (2.8) as

$$y_t = x'_{1t} \beta_1 + x'_{2t} \beta_2 + u_t. \quad (2.12)$$

First, regress y_t on x_{1t} and get the residuals \tilde{e}_{yt} . Second, regress x_{2t} on x_{1t} and get the residuals \tilde{e}_{2t} . (If x_{2t} is a vector, then this is one regression per element in x_{2t} and \tilde{e}_{2t} is the corresponding vector of residuals.) Third, regress \tilde{e}_{yt} on \tilde{e}_{2t} . This gives the same estimates as $\hat{\beta}_2$ from the multiple regression of y_t on both x_{1t} and x_{2t} . (The proof is a straightforward reshuffling of the first order conditions, see, for instance, Greene (2018) 3.)

	HiTec	Utils
SMB residual	0.18 (3.76)	-0.20 (-4.08)
HML residual	-0.51 (-11.16)	0.31 (5.78)
R^2	0.27	0.12
obs	660	660

Table 2.3: 3rd step in Frisch-Waugh regressions, monthly returns, %, US data 1970:01-2024:12. 1st step: regress the sector returns on (c,market return) and extract the residuals. 2nd step: regress SMB and HML returns on (c,market return) and extract the residuals. 3rd step: regress step 1 residuals on step 2 residuals. Numbers in parentheses are t-stats.

Empirical Example 2.16 (*Factor model regression*) *Table 2.3.* Shows the results from the third step for the same data as used in *Table 2.2*. Both the point estimates and the standard errors should be the same.

The perhaps most common application of this is when x_{1t} contains various dummy variables (for instance, for different cross-sectional units) and x_{2t} are the variables of key interest. It can then be convenient to apply this 3-step approach. This is used in the fixed effects estimator for panel data.

2.1.3 Least Squares: Goodness of Fit

The quality of a regression model is often measured in terms of its ability to explain the movements of the dependent variable.

Let \hat{y}_t be the fitted (predicted) value of y_t , $\hat{y}_t = x'_t \hat{\beta}$. If a constant is included in the regression (or the means of y and x are zero), then a check of the *goodness of fit* of the model is given by the fraction of the variation in y_t that is explained by the model

$$R^2 = \frac{\widehat{\text{Var}}(\hat{y}_t)}{\widehat{\text{Var}}(y_t)} = 1 - \frac{\widehat{\text{Var}}(\hat{u}_t)}{\widehat{\text{Var}}(y_t)}, \quad (2.13)$$

where $\widehat{\text{Var}}()$ denotes a sample variance. This shows that the variance of the residuals and R^2 are negatively related, so it follows that a high R^2 will be associated with low standard errors of the estimates.

Example 2.17 (R^2) From Example 2.6 we have $\widehat{\text{Var}}(\hat{u}_t) = 0.18$ and $\widehat{\text{Var}}(y_t) = 2.34$, so

$$R^2 = 1 - 0.18/2.34 \approx 0.92.$$

See Figure 2.4.

The R^2 can also be rewritten as the squared (sample) correlation of the actual and fitted values

$$R^2 = \widehat{\text{Corr}}(y_t, \hat{y}_t)^2. \quad (2.14)$$

To understand this result, suppose that x_t has no explanatory power, so R^2 should be zero. How does that happen? Well, if x_t is uncorrelated with y_t , then slope coefficients are zero. As a consequence, \hat{y}_t equals the intercept (a constant). This means that R^2 in (2.13) is zero, since the fitted residual has the same variance as the dependent variable (\hat{y}_t captures nothing of the movements in y_t). Similarly, R^2 in (2.14) is also zero, since a constant is always uncorrelated with anything else, as correlations measure comovements around the means.

Proof (of (2.13)–(2.14)) Write the regression equation as

$$y_t = \hat{y}_t + \hat{u}_t,$$

where hats denote fitted values. Since \hat{y}_t and \hat{u}_t are uncorrelated (always true in OLS provided the regression includes a constant), we have

$$\hat{v}(y_t) = \hat{v}(\hat{y}_t) + \hat{v}(\hat{u}_t),$$

where $\hat{v}()$ is compact notation for a sample variance. R^2 is defined as the fraction of $\hat{v}(y_t)$ that is explained by the model

$$R^2 = \frac{\hat{v}(\hat{y}_t)}{\hat{v}(y_t)} = \frac{\hat{v}(y_t) - \hat{v}(\hat{u}_t)}{\hat{v}(y_t)} = 1 - \frac{\hat{v}(\hat{u}_t)}{\hat{v}(y_t)}.$$

Equivalently, we can rewrite R^2 by noting that

$$\hat{c}(y_t, \hat{y}_t) = \hat{c}(\hat{y}_t + \hat{u}_t, \hat{y}_t) = \hat{v}(\hat{y}_t),$$

where $\hat{c}()$ is compact notation for a sample covariance. Use this in the numerator of R^2 and multiply by $\hat{c}(y_t, \hat{y}_t)/\hat{v}(\hat{y}_t) = 1$

$$R^2 = \frac{\hat{c}(y_t, \hat{y}_t)^2}{\hat{v}(y_t)\hat{v}(\hat{y}_t)} = \widehat{\text{Corr}}(y_t, \hat{y}_t)^2.$$

□

Remark 2.18 (R^2 from simple regression*) Suppose $\hat{y}_t = \beta_0 + \beta_1 x_t$, so (2.14) becomes

$$R^2 = \widehat{\text{Corr}}(y_t, x_t)^2.$$

The R^2 can never decrease as we add more regressors, which might make it attractive to add more and more regressors. To avoid that, some researchers advocate using an ad hoc punishment for many regressors, the *adjusted R^2* : $\bar{R}^2 = 1 - (1 - R^2)(T - 1)/(T - k)$, where k is the number of regressors (including the constant). This measure can be negative.

Empirical Example 2.19 (*CAPM regressions*) See Table 2.1 for CAPM regressions for two industry portfolios where the R^2 values clearly differ. This is seen also from the dispersion around the regression line in Figure 2.5.

2.2 Missing Data

It is often the case that some data is missing. For instance, we may not have data on regressor 3 for observation $t = 7$. If data is *missing in a random way*, then we can simply exclude (y_t, x_t) for observation t if it has some missing data. In contrast, if data is missing in a non-random way, for instance, depending on the value of y_{it} , then we have to apply more sophisticated sample-selection models (not discussed in this chapter).

Remark 2.20 (*Replacing missing values with 0**) Instead of excluding (y_t, x_t) for the t with some missing data, we could set $(y_t, x_t) = (0, \mathbf{0}_k)$. Notice that also the constant should be set to 0. This would not change the estimates. However, we must then be careful with how to estimate the variance of the residuals (σ^2): the estimation would divide by the effective number of observations (non-zero y_t values).

2.3 The Distribution of $\hat{\beta}$

The estimated coefficients are random variables since they depend on which particular sample has been “drawn.” It is important to remember that we always assume that there are some true (but unknown) parameter values that would be the same across samples. The main reason why the estimates differ across samples is that the model is not perfect: there are residuals and they differ randomly across different samples. It could also be the case that the regressors differ across samples. See Figure 2.6 for an illustration from a computer simulation (Monte Carlo simulation).

We usually do not have several samples, so the variation across samples is not directly observable. However, we can, under some assumption, use the *variation within our sample to figure out how the variation across samples ought to be*. This can help us testing hypotheses about the coefficients, for instance, that $\beta_1 = 0$.

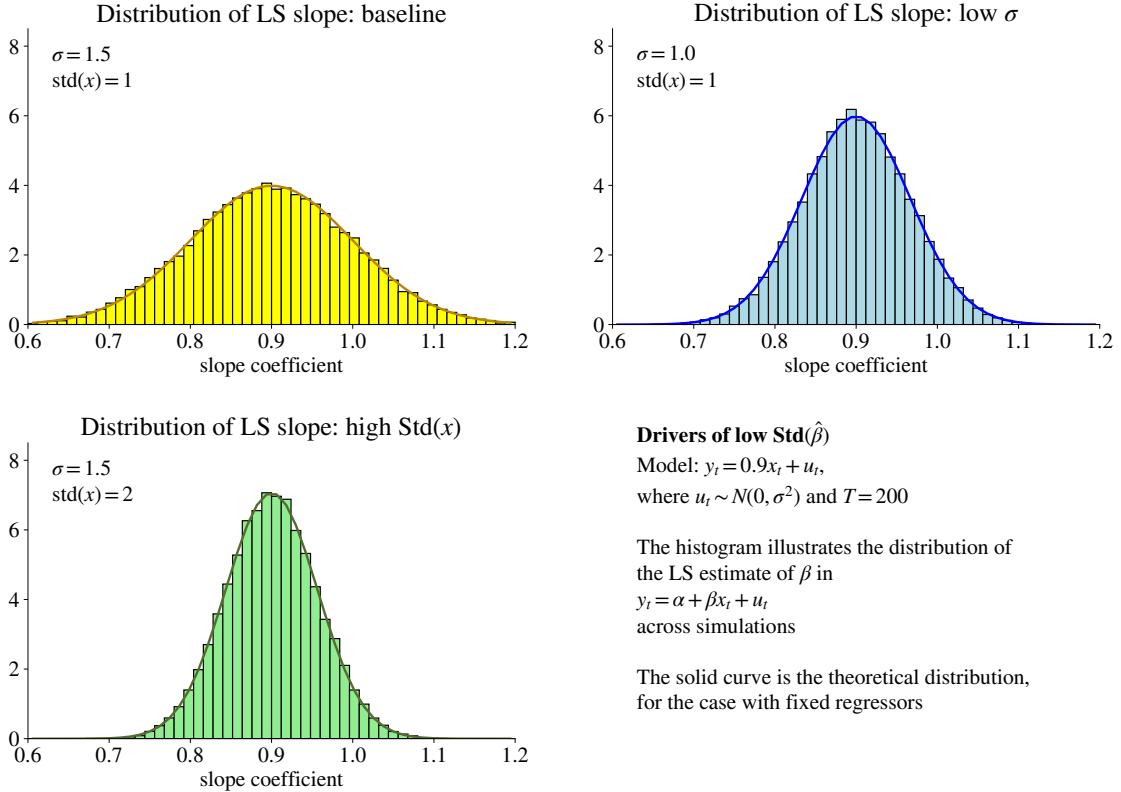


Figure 2.6: Distribution of OLS estimate, from simulation and theory

Use (2.8) in (2.11) to substitute for y_t and simplify

$$\hat{\beta} = \beta + \left(\sum_{t=1}^T x_t x_t' \right)^{-1} \sum_{t=1}^T x_t u_t \text{ or} \quad (2.15)$$

$$= \beta + \left(\sum_{t=1}^T x_t x_t' / T \right)^{-1} \sum_{t=1}^T x_t u_t / T \quad (2.16)$$

where $\sum_{t=1}^T x_t x_t'$ is a $k \times k$ matrix and $\sum_{t=1}^T x_t u_t$ is a $k \times 1$ vector. This shows that so the OLS estimates, the $\hat{\beta}$ vector, equals the true value, β , plus a term that involves the residuals, u_t . Equation (2.15) will give different values of $\hat{\beta}$ when we use different samples, that is, different draws of the random variable y_t (and thus u_t), and perhaps also different values of x_t if the regressors are also random. The distribution of these estimates across samples would describe the uncertainty about the true value. If we are willing to assume that the statistical properties of u_t (and possibly x_t) are the same across samples, then our observed sample can help us extrapolate and thus assess this uncertainty.

The first conclusion from (2.16) is that, with $u_t = 0$ the estimate would always

be perfect. In contrast, with large movements in u_t we will see large movements in $\hat{\beta}$ across samples. The second conclusion is that a small sample (small T) will also lead to large random movements in $\hat{\beta}$, in contrast to a large sample where the randomness in $\sum_{t=1}^T x_t u_t / T$ is averaged out more effectively (it should be zero in a large sample).

There are three main routes to learn more about the distribution of $\hat{\beta}$: (i) set up a small “experiment” in the computer and simulate the distribution (Monte Carlo or bootstrap simulations); (ii) pretend that the regressors can be treated as fixed numbers (or treat all results as being conditional on the particular sample of regressor values that we have) and then assume something about the distribution of the residuals; or (iii) use the asymptotic (large sample) distribution as an approximation.

The simulation approach has the advantage of giving a precise answer, but the disadvantage of requiring a very precise question (must write computer code that is tailor-made for the particular model, including specific parameter values).

In contrast, asymptotic theory gives more general results, but only deliver useful results for large samples. Treating the regressors as fixed is easier, and is often good enough for illustrating some basic properties of the estimation method (a teaching device), but fails to capture some important aspects of stochastic regressors that might be important. The latter will be discussed later on.

The typical outcome of all three approaches will (under strong assumptions, to be discussed below) be that

$$\hat{\beta} \sim N(\beta, S_{xx}^{-1}\sigma^2), \text{ where } S_{xx} = \sum_{t=1}^T x_t x_t' \quad (2.17)$$

and where $\sigma^2 = \text{Var}(u_t)$ denotes the variance of the residuals. In practice, we calculate/estimate both $\sum_{t=1}^T x_t x_t'$ and σ^2 from the available data and fitted residuals. See Table 2.1 for an empirical example and Figure 2.6 for an illustration of how the results depend on σ and the standard deviation of x_t . The variance-covariance matrix in (2.17) is $k \times k$ and will be denoted $\text{Var}(\hat{\beta})$, while the standard errors (or standard deviations), denoted $\text{Std}(\hat{\beta})$, are the square root of the diagonal elements.

Example 2.21 (*The variance-covariance matrix in (2.17)) with two regressors.*) With two regressors we have

$$\text{Var}(\hat{\beta}) = \text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) = \begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) \end{bmatrix}.$$

Remark 2.22 (*Matrix notation**) Let X be a $T \times k$ matrix where row t is filled with

the elements of x_t . Then, the variance-covariance matrix in (2.17) can also be written $(X'X)^{-1}\sigma^2$.

2.3.1 The Distribution of $\hat{\beta}$ with Fixed Regressors

Assuming fixed regressors is mostly a simplifying (teaching) device, since econometrics is typically not applied to controlled experiments. However, it is easy to derive results for this case, and those results are often similar to what asymptotic theory gives.

A variant of the “fixed regressors” assumption is to treat all results as *conditional* on the regressor values in our particular sample. This does not rule out the possibility that samples of x_t differ, but the results only apply to the subset of samples that have the same x_t values as our current sample. This is effectively treating x_t as fixed.

Remark 2.23 (*Notation for conditional on X) Some texts emphasize the “conditional on X ” interpretation instead of the fixed regressor interpretation. The notation is then often $E(\hat{\beta}|X)$ and $\text{Var}(\hat{\beta}|X)$ where the “ $|X$ ” denotes that the expectation/variance is conditional on the sample of x_1, \dots, x_T at hand.

The results derived below are based on the *Gauss-Markov assumptions* : (a) the residuals have zero means, (b) have constant variances and (c) are not correlated across observations. In other words, the *residuals are assumed to be zero mean iid variables*.

For notational convenience, write (2.16) as

$$\hat{\beta} = \beta + S_{xx}^{-1} (x_1 u_1 + x_2 u_2 + \dots + x_T u_T), \text{ where } S_{xx} = \sum_{t=1}^T x_t x_t'. \quad (2.18)$$

Since x_t is assumed to be fixed, the expected value of this expression is

$$E \hat{\beta} = \beta + S_{xx}^{-1} (x_1 E u_1 + x_2 E u_2 + \dots + x_T E u_T) = \beta, \quad (2.19)$$

which follows from assuming that the residuals have zero means (see (2.1)). This says that OLS is *unbiased* when the regressors are fixed. The interpretation is that we can expect OLS to give a correct answer, at least on average. That is, if we could draw many different samples and estimate the slope coefficients in each of them, then the average of those estimates would be the correct numbers (β). Clearly, this is something we want from an estimation method. In contrast, a method that was systematically wrong would not be very attractive. Unfortunately, the unbiasedness does not always carry over to the case of stochastic regressors (to be discussed later).

Remark 2.24 (*Linear combination of normally distributed variables.*) If the random variables z_t and v_t are normally distributed and independent of each other, then $a + bz_t + cv_t$ is normally distributed with a mean of $a + b\mu_z + c\mu_v$ and a variance of $b^2\sigma_z^2 + c^2\sigma_v^2$.

Suppose $u_t \sim N(0, \sigma^2)$ and that the residuals are independent of each other, then (2.18) shows that $\hat{\beta}$ is *normally distributed*. The reason is that $\hat{\beta}$ is just a constant (β) plus a linear combination of independent normally distributed residuals. It is clear that the mean of this normal distribution is β (the true value), since $\hat{\beta}$ is unbiased.

Finding the *variance-covariance matrix* of $\hat{\beta}$ is slightly more complicated. Remember that we treat x_t as fixed numbers and assume that the residuals are iid: they are uncorrelated with each other (follows from independently distributed) and have the same variances (follows from identically distributed). We also notice that the variance-covariance matrix of $x_t u_t$ equals

$$\text{Var}(x_t u_t) = x_t x'_t \sigma_t^2. \quad (2.20)$$

where $\sigma_t^2 = \text{Var}(u_t)$ and where we use the fact that the vector x_t is non-random.

Example 2.25 (of (2.20)) With

$$x_t = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \sigma_t^2 = 0.18, \text{ we get}$$

$$\text{Var}(x_t u_t) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} \times 0.18 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \times 0.18.$$

The variance of (2.18) can then be written

$$\begin{aligned} \text{Var}(\hat{\beta}) &= S_{xx}^{-1} \text{Var}(x_1 u_1 + x_2 u_2 + \dots + x_T u_T) S_{xx}^{-1} \\ &= S_{xx}^{-1} (x_1 x'_1 \sigma_1^2 + x_2 x'_2 \sigma_2^2 + \dots + x_T x'_T \sigma_T^2) S_{xx}^{-1} \\ &= S_{xx}^{-1} (x_1 x'_1 \sigma^2 + x_2 x'_2 \sigma^2 + \dots + x_T x'_T \sigma^2) S_{xx}^{-1} \\ &= S_{xx}^{-1} \left(\sum_{t=1}^T x_t x'_t \right) \sigma^2 S_{xx}^{-1} \\ &= S_{xx}^{-1} \sigma^2. \end{aligned} \quad (2.21)$$

The first line follows directly from (2.18), since β is a constant. The second line follows from assuming that the residuals are uncorrelated with each other ($\text{Cov}(u_i, u_j) = 0$ if $i \neq j$), so all cross terms ($x_i x_j \text{Cov}(u_i, u_j)$) are zero. The third line follows from

assuming that the variances are the same across observations ($\sigma_i^2 = \sigma_j^2 = \sigma^2$). The fourth and fifth lines are just algebraic simplifications which use the definition of S_{xx} .

This expression shows that there are three main ways of getting a low uncertainty (low $\text{Var}(\hat{\beta})$). First, a large sample (T is large), decreases the S_{xx}^{-1} factor (since $S_{xx} = \sum_{t=1}^T x_t x'_t$ increases with T) while σ^2 stays constant: a larger sample gives a smaller uncertainty about the estimate. Second, large movements in the regressors (large value of S_{xx}) should help us estimate the link between x and y since the movements in y driven by x should then dominate over the movements in y driven by the residual. Third, a lower volatility of the residuals (lower σ^2) also gives a lower uncertainty about the estimate. See Figures 2.6 and 2.7.

A key assumption in regression analysis is that our sample is “representative” of the population. In practice, this means that *we can estimate σ^2* (and S_{xx} if the regressors are stochastic) in (2.21) *from the data in the sample*. This is the main “trick” behind using our (one and only) sample to inform us about how the distribution of $\hat{\beta}$ (across samples) looks like. This is a plausible assumption when our sample is a random draw from the population (say, 700 out of a total of 10,000 firms). This assumption is more stringent when the sample is a time series of data points. Then we effectively assume that the past (before the sample) and the future (after the sample) will have the same structure. In case you are not willing to accept those assumptions, the t -stats are useless.

2.3.2 Examples of the Variance-Covariance Matrix*

This section presents some examples of the variance covariance matrix (2.21). This provides some insights to the properties of OLS and also prepares for some later applications.

Example 2.26 (Applying (2.21)) *When the regressor is just a constant (equal to one) $x_t = 1$, then we have*

$$\sum_{t=1}^T x_t x'_t = \sum_{t=1}^T 1 \times 1' = T \text{ so } \text{Var}(\hat{\beta}) = \sigma^2 / T.$$

This is the classical expression for the variance of a sample mean, assuming iid data.

Example 2.27 (Applying (2.21)) *When the regressor is a single zero mean variable, then we have*

$$\sum_{t=1}^T x_t x'_t = \widehat{\text{Var}}(x)T \text{ so } \text{Var}(\hat{\beta}) = \sigma^2 / (\widehat{\text{Var}}(x)T),$$

where $\widehat{\text{Var}}(x)$ is the sample variance. Clearly, $\text{Var}(\hat{\beta})$ is increasing in σ^2 , but decreasing in both T and $\widehat{\text{Var}}(x)$. See Figure 2.7 for an illustration.

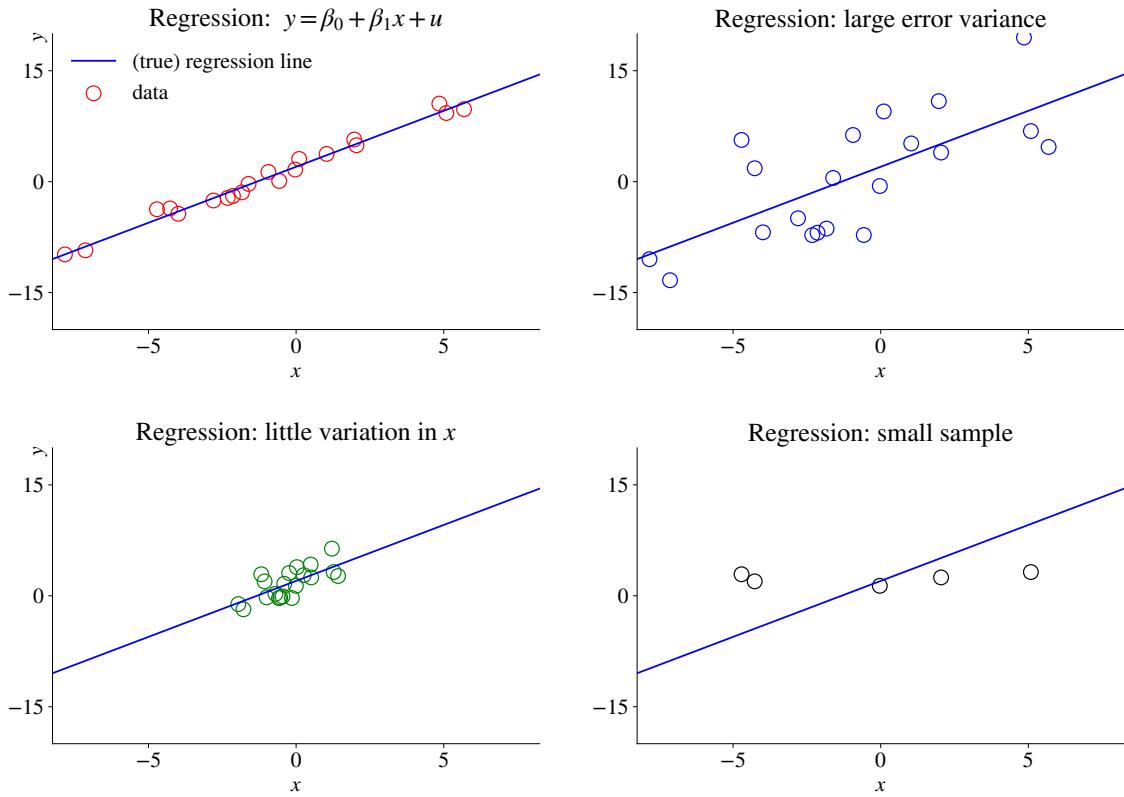


Figure 2.7: Regressions: importance of error variance and variation of regressor

Example 2.28 (*Applying (2.21)) When the regressor is just a constant (equal to one) and one variable regressor with zero mean, z_t , so $x_t = [1, z_t]$, then we have

$$\sum_{t=1}^T x_t x_t' = \sum_{t=1}^T \begin{bmatrix} 1 & z_t \\ z_t & z_t^2 \end{bmatrix} = T \begin{bmatrix} 1 & 0 \\ 0 & \widehat{\text{Var}}(z) \end{bmatrix}, \text{ so}$$

$$\text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) = \sigma^2 \left(\sum_{t=1}^T x_t x_t' \right)^{-1} = \frac{\sigma^2}{T} \begin{bmatrix} 1 & 0 \\ 0 & 1/\widehat{\text{Var}}(z) \end{bmatrix},$$

where $\widehat{\text{Var}}(z)$ is the sample variance. This is a combination of the two preceding examples.

Example 2.29 (Distribution of slope coefficient) From Example 2.6 we have $\text{Var}(\hat{u}) = \sigma^2 = 0.18$ and $\sum_{t=1}^T x_t x_t = 2$, so $\text{Var}(\hat{\beta}_1) = 0.18/2 = 0.09$, which gives $\text{Std}(\hat{\beta}_1) = 0.3$.

Example 2.30 (Covariance matrix of $\hat{\beta}_1$ and $\hat{\beta}_2$) From Example 2.15

$$\sum_{t=1}^T x_t x'_t = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \text{ and } \sigma^2 = 0.18, \text{ then}$$

$$\text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} 0.18 = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.09 \end{bmatrix}.$$

The standard deviations (also called standard errors) are therefore

$$\begin{bmatrix} \text{Std}(\hat{\beta}_1) \\ \text{Std}(\hat{\beta}_2) \end{bmatrix} = \begin{bmatrix} 0.24 \\ 0.3 \end{bmatrix}.$$

Example 2.31 (*Applying (2.21)) When the regressor is just a constant (equal to one) and one variable regressor with a non-zero mean, z_t , so $x_t = [1, z_t]$, then we have

$$\sum_{t=1}^T x_t x'_t = \sum_{t=1}^T \begin{bmatrix} 1 & z_t \\ z_t & z_t^2 \end{bmatrix} = T \begin{bmatrix} 1 & \bar{z} \\ \bar{z} & \bar{z}^2 \end{bmatrix}.$$

Using the standard expression for an inverse of a 2×2 matrix, we get

$$\text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) = \sigma^2 \left(\sum_{t=1}^T x_t x'_t \right)^{-1} = \frac{\sigma^2}{T} \frac{1}{\widehat{\text{Var}}(z)} \begin{bmatrix} \widehat{\text{Var}}(z) + \bar{z}^2 & -\bar{z} \\ -\bar{z} & 1 \end{bmatrix},$$

where we also use the fact that $\widehat{\text{Var}}(z) = \bar{z}^2 - \bar{z}^2$. The variance of the slope coefficient ($\hat{\beta}_2$) is the same as in the previous examples, but the variance of the intercept ($\hat{\beta}_1$) now depends on the properties of the regressor z_t , in particular, the uncertainty about its mean.

2.3.3 Multicollinearity

When the regressors in a multiple regression are highly correlated, then we have a practical problem: the standard errors of individual coefficients tend to be large, even if the R^2 suggests that the regression does fairly well.

As a simple example, consider the regression

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t, \quad (2.22)$$

where (for simplicity) the dependent variable and the regressors have zero means. In this

case, the variance (assuming iid errors) is

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{T \widehat{\text{Var}}(x_{2t})} \frac{1}{1 - \gamma^2}, \quad (2.23)$$

where γ is the sample correlation of x_{1t} and x_{2t} . If the regressors are highly correlated, then the uncertainty about the slope coefficient is high. The basic reason is that we see that the regressors have an effect on y_t , but it is hard to tell if that effect is from regressor one or two (since they are so similar). This can well lead to a situation where the R^2 is high and a joint test easily rejects the null hypothesis that all slopes are zero, but each individual slope coefficient is insignificant.

Remark 2.32 ($\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$) To simplify the notation, let the sample variances of x_{1t} and x_{2t} be equal to one. It can then be shown (see proof below) that

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2}{T} \frac{\gamma}{1 - \gamma^2}.$$

This says that when the regressors are positively correlated ($\gamma > 0$), then the coefficients tend to be negatively correlated (notice the leading minus sign).

More generally, in the multiple regression

$$y_t = x'_t \beta + u_t, \quad (2.24)$$

it is straightforward to show that for all slope coefficients (not the intercept)

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{T \widehat{\text{Var}}(x_{it})} \frac{1}{1 - R_i^2}, \quad (2.25)$$

where R_i^2 is the R^2 value obtained from regressing x_{it} on the *other* regressors (including a constant). The last term ($1/(1 - R_i^2)$) is often called the *variance inflation factor* and some regression packages report the maximum across the regressors, and a value of 10 or larger ($R_i^2 \geq 0.9$) is considered highly problematic. The name variance inflation factor is meant to indicate how much the variance increases compared to a simple regression, assuming σ^2 is unchanged. In practice, the estimated σ^2 often change considerably.

Proof (of (2.23) and Remark 2.32*). Recall that for a 2×2 matrix we have

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

For the regression (2.22) we get

$$\begin{aligned}\sum_{t=1}^T x_t x_t' &= \begin{bmatrix} T\hat{v}(x_{1t}) & T\hat{c}(x_{1t}, x_{2t}) \\ T\hat{c}(x_{1t}, x_{2t}) & T\hat{v}(x_{2t}) \end{bmatrix}^{-1} \\ &= \frac{1}{T^2\hat{v}(x_{1t})\hat{v}(x_{2t}) - T^2\hat{c}(x_{1t}, x_{2t})^2} \begin{bmatrix} T\hat{v}(x_{2t}) & -T\hat{c}(x_{1t}, x_{2t}) \\ -T\hat{c}(x_{1t}, x_{2t}) & T\hat{v}(x_{1t}) \end{bmatrix},\end{aligned}$$

where $\hat{v}()$ and $\hat{c}()$ are short hand notation for sample variances and covariances. The variance of the second slope coefficient is σ^2 time the lower right element of this matrix

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{T} \frac{\hat{v}(x_{1t})}{\hat{v}(x_{1t})\hat{v}(x_{2t}) - \hat{c}(x_{1t}, x_{2t})^2},$$

Divide both numerator and denominator by $\hat{v}(x_{1t})\hat{v}(x_{2t})$ to get

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{T} \frac{1/\hat{v}(x_{2t})}{1 - \widehat{\text{Corr}}(x_{1t}, x_{2t})^2},$$

which is the same as (2.23). Similarly, we get

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2}{T} \frac{\hat{c}(x_{1t}, x_{2t})}{\hat{v}(x_{1t})\hat{v}(x_{2t}) - \hat{c}(x_{1t}, x_{2t})^2},$$

which gives the result in Remark 2.32. \square

2.4 The Distribution of $\hat{\beta}$: More General Results

2.4.1 Problems with the Gauss-Markov (iid) and Normality Assumptions

The previous results on the distribution of $\hat{\beta}$ have several weak points, which will be briefly discussed here (a later chapter presents more details).

First, the Gauss-Markov assumptions of iid residuals (constant volatility and no correlation across observations) are likely to be false in many cases.

Second, the idea of fixed regressor is clearly just a simplifying assumption, and unlikely to be relevant for economics and financial data. If the regressors are random variables then we typically not rule out that u_t and x_{t+s} are correlated, for instance, when the regressors include the lagged dependent variable.

Third, there are no particularly strong reasons for why the residuals should be normally distributed. If not, the estimates are unlikely to be normally distributed in small samples, but may well be in large samples (due to the central limit theorem).

The next few sections introduce these issues, but later chapters will discuss them in more detail.

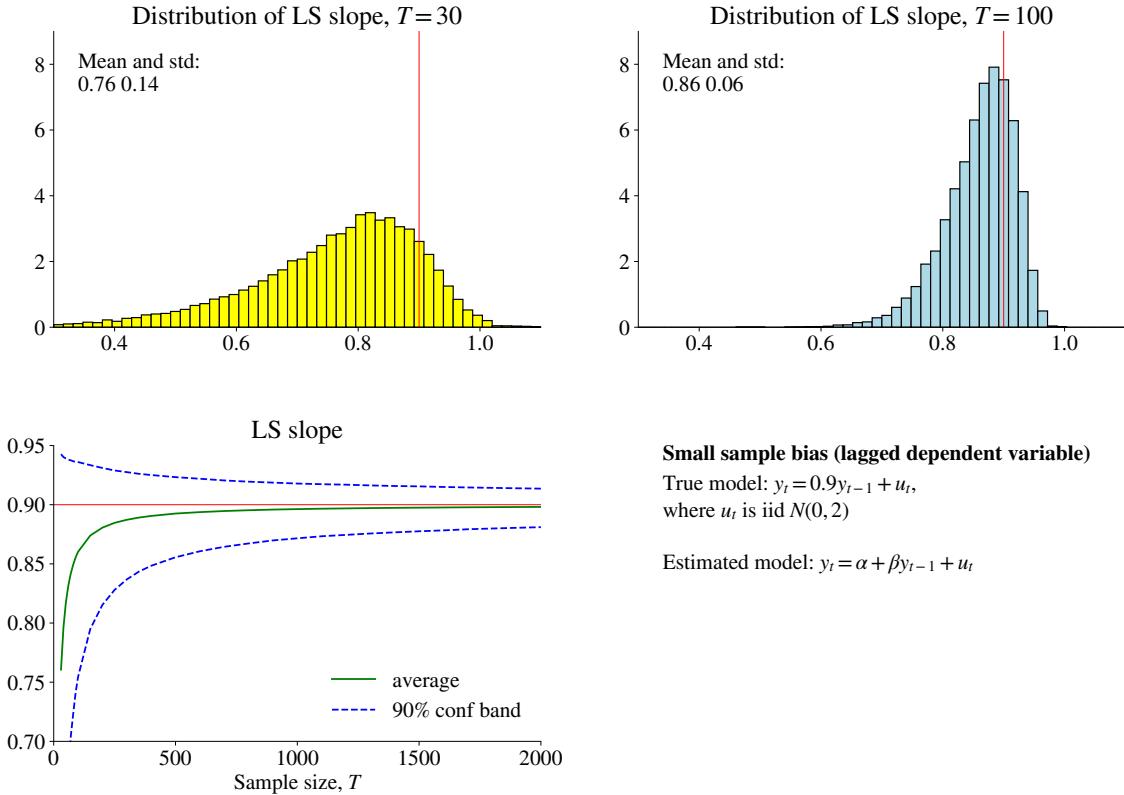


Figure 2.8: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

2.4.2 Failure of the Gauss-Markov Assumptions

If the residuals are not iid, then we have to stop at the first line of (2.21), so

$$\text{Var}(\hat{\beta}) = S_{xx}^{-1} S S_{xx}^{-1}, \text{ where } S = \text{Var}(\Sigma_{t=1}^T x_t u_t). \quad (2.26)$$

The S matrix is estimated in different ways (for instance, using White's or Newey-West's methods, to be discussed later) depending on the properties of the residuals (heteroskedasticity or autocorrelation).

2.4.3 Bias

If an estimation method is *biased*, then it produces systematically wrong (say, too low) coefficients. Sometimes, this bias disappears in large samples, so it is only a small sample bias.

Figure 2.8 illustrates some simulation results from estimating an AR(1)

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t \quad (2.27)$$

on artificial samples where “data” is generated from

$$y_t = 0.9y_{t-1} + u_t, \text{ where } u_t \text{ is iid.} \quad (2.28)$$

In this case, the regressor is a (stochastic) random variable (not fixed). Figure 2.8 suggests that the estimates are biased, that is, not centered on the true value, in small samples, but the bias appears to decrease as the sample size increases.

To understand these results, recall that (2.16) says that

$$\hat{\beta} = \beta + \left(\sum_{t=1}^T x_t x_t' / T \right)^{-1} \sum_{t=1}^T x_t u_t / T \quad (2.29)$$

where u_t are the true residuals. (We will never observe the true residuals, so (2.29) can only be used for a conceptual discussion.)

To get *unbiased estimates* ($E \hat{\beta} = \beta$), the second term of the right hand side of (2.29) should have an expectation of zero. This would happen when u_t and x_{t+s} (for all s) are independent. This is hard to guarantee when the regressors are random variables. For instance, in the AR(1) example, then u_t affects x_{t+1} so there is an interaction between the numerator and denominator. This is probably most easily investigated by simulations.

Remark 2.33 (*Bias of AR(1)) It can be shown (see, for instance, Pesaran (2015) 14) that a bias corrected estimate of the AR(1) coefficient can be calculated as $(1 + T \hat{\beta}_1) / (T - 3)$.

2.4.4 Consistency

If an estimation method is *inconsistent*, then it produces systematically wrong (say, too low) coefficients also in very large samples (actually, in the limit as $T \rightarrow \infty$). Figure 2.8 suggests that the problem with the AR(1) estimation vanishes as the sample size increases.

To get *consistent estimates* (which is defined as the bias and the variance of $\hat{\beta}$ go to zero as $T \rightarrow \infty$), then it is enough if x_t and u_t (in the same period/for the same observation) are uncorrelated. This is indeed the case in the AR(1) simulations discussed before. To see this from (2.29), notice that a “law of large numbers” makes the numerator ($\sum_{t=1}^T x_t u_t / T$) converge to the population covariance of u_t and x_t . (Also, the denominator converges to a fixed number, so we can focus on the numerator.)

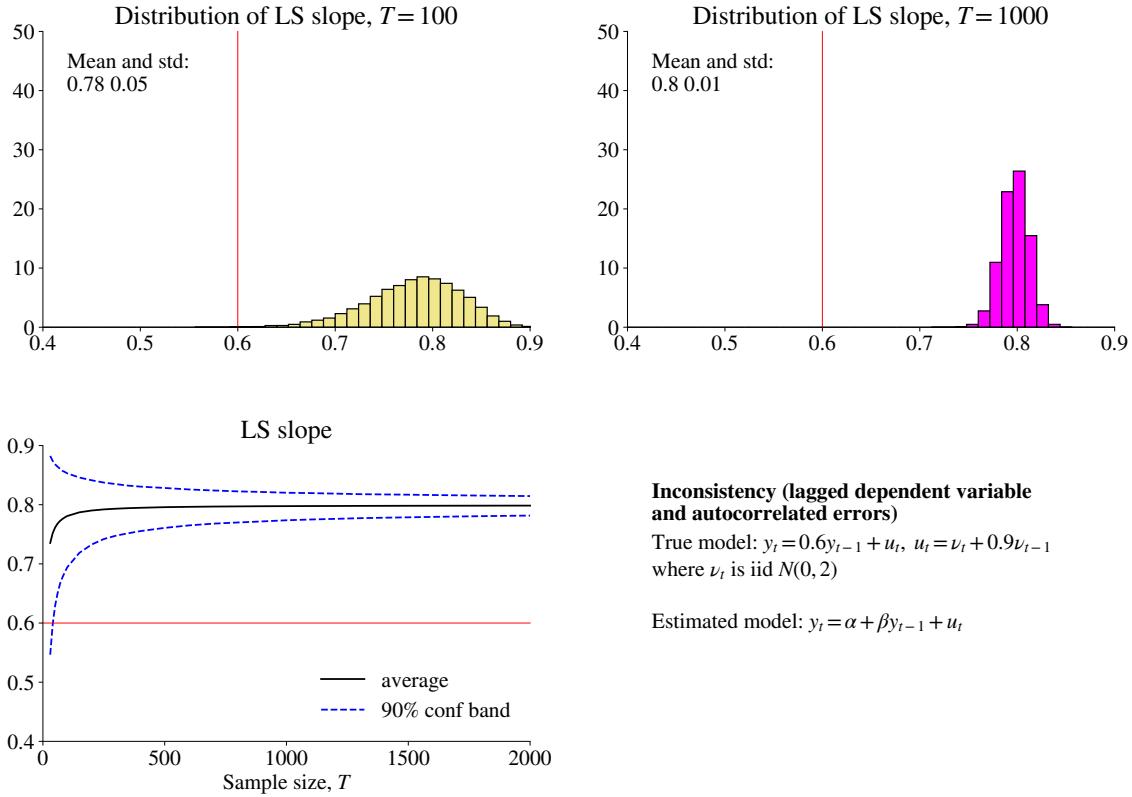


Figure 2.9: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

This means that *if* we knew that $\text{Cov}(x_t, u_t) = 0$ (in the population), then we would also know that OLS is consistent. However, since the true errors are never observed, this cannot be shown by empirical methods. (Recall OLS always construct fitted errors so they are uncorrelated with the regressors.) Instead, we have to rely on theoretical arguments that make it plausible to *believe or not* in consistency.

To make matters worse, it is often the case that $\hat{\beta}$ converges (as T increases), but perhaps not to what you hoped for. As an illustration of how tricky this can be, consider the case in Figure 2.9. It estimates the same AR(1) as in (2.27) but where the simulated “data” now follows

$$y_t = \rho y_{t-1} + u_t, \text{ where } u_t = v_t + \theta v_{t-1} \text{ where } v_t \text{ is iid.} \quad (2.30)$$

In this case, the residuals are themselves autocorrelated. The figure clearly shows that the OLS estimate of the slope β_1 in (2.27) does *not* converge to the true value ρ as the

sample size increases: OLS is inconsistent. The reason in this case is that u_t and y_{t-1} (the regressor) both depend on v_{t-1} so they are correlated.

An *a priori* argument for why OLS should be able to estimate a model consistently thus require a careful discussion of the model properties: how can we explain that the residuals are uncorrelated with the regressors? (Alternatively, we use an instrumental variables technique, which is discussed later on.) This typically involves a discussion of the following points.

1. Have we excluded (omitted) some relevant regressors? If so, their effect is captured by the residual. If these excluded regressors are correlated to some of the included regressors, then we have a problem.
2. Do we use a lagged dependent variable as regressor at the same time as the residual is autocorrelated? (This is the previous example.)
3. Does y_t affect x_t ? If so a shock to the equation that explains y_t also drives x_t and we get a correlation between the regressor (x_t) and the residual. A classical case is when we try to estimate how the demand for a product depends on its price. In fact, such an equation actually estimates a mix between the demand and supply elasticities.
4. Is the regressor measured without (important) errors? If not, we again have a correlation between (the used) regressor and the residual.

2.4.5 Normality

If the regressors x_t are fixed numbers and u_t is normally distributed, then the second term in (2.29) shows that the normality carries over to $\hat{\beta}$ also in small samples. We can test the assumption of normally distributed residuals by using a Jarque-Bera test

$$JB = \frac{T}{6} \text{skewness}^2 + \frac{T}{24} (\text{kurtosis} - 3)^2, \quad (2.31)$$

which has χ^2_2 distribution under the null hypothesis that both the skewness and excess kurtosis (that is, kurtosis–3) are zero.

2.4.6 Asymptotic Normality

Even if the normality test fails or if the regressors are stochastic, we can often still hope for a (close to) normal distribution of $\hat{\beta}$ if the sample is large—due to the central limit

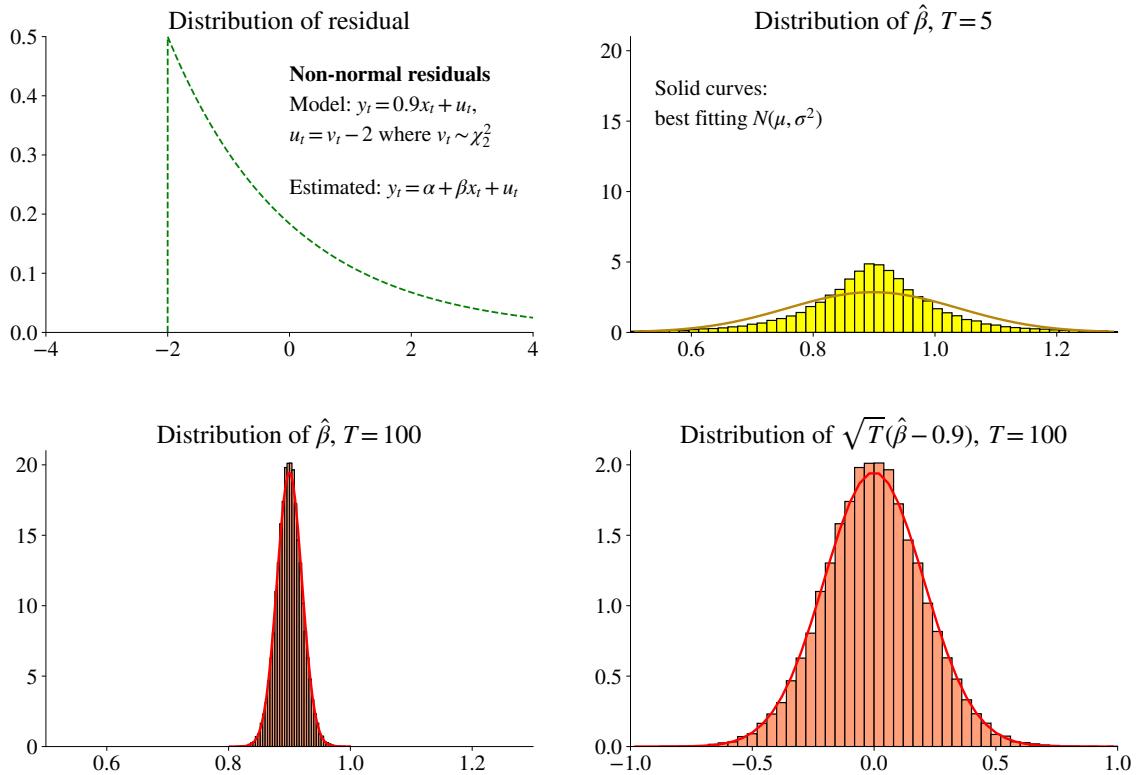


Figure 2.10: Distribution of OLS estimate, from simulations

theorem. This is illustrated in Figure 2.10. It is based on simulations where the residual is drawn from a very non-normal distribution. For a small sample, this carries over to $\hat{\beta}$ and the t -stat for the hypothesis that $\beta = 0$. However, in these simulations, already samples of moderate size tend to give an almost normal distribution. This is perhaps most evident when we plot the histogram of $\sqrt{T}(\hat{\beta} - \beta)$ rather than of $\hat{\beta}$.

To understand the theory of this, rewrite (2.29) by subtracting β from both sides and then multiply both sides by \sqrt{T}

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\sum_{t=1}^T x_t x_t' / T \right)^{-1} \sqrt{T} \sum_{t=1}^T x_t u_t / T. \quad (2.32)$$

The inverted term is the sample average of $x_t x_t'$ which will converge to a matrix of fixed numbers (the population mean of $x_t x_t'$) as $T \rightarrow \infty$. We can therefore focus on what happens to the numerator. It is \sqrt{T} times the sample average of $x_t u_t$ (a vector). Under weak conditions a central limit theorem applies to $\sqrt{T} \times$ a sample average: it typically converges to a normal distribution.

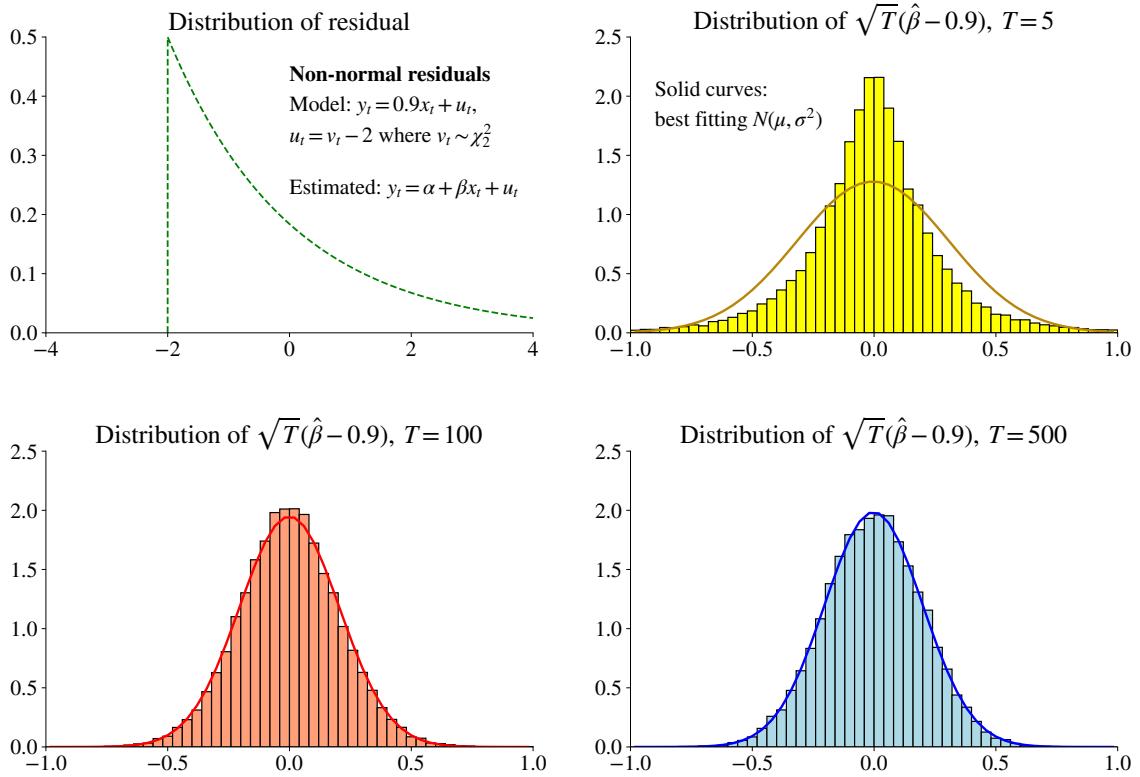


Figure 2.11: Distribution of OLS estimate, from simulations

This shows that $\sqrt{T}(\hat{\beta} - \beta)$ has an *asymptotic normal distribution*

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, W), \quad (2.33)$$

where \xrightarrow{d} indicates that the random variable $\sqrt{T}(\hat{\beta} - \beta)$ has a normal distribution as T increases. This often holds as a reasonable approximation also in moderately sized samples. See Figure 2.11 for an illustration. Actually, it turns out that this is a property of many estimators (not just OLS), basically because most estimators are some kind of sample average.

The variance-covariance matrix W in (2.33) can be understood from (2.32). It has the form

$$W = \left(\frac{S_{xx}^*}{T} \right)^{-1} \frac{S}{T} \left(\frac{S_{xx}^*}{T} \right)^{-1}, \quad (2.34)$$

where S_{xx}^*/T is the limit of $\sum_{t=1}^T x_t x_t' / T$ as $T \rightarrow \infty$ and S/T is the variance-covariance matrix of $\sum_{t=1}^T x_t u_t / \sqrt{T}$, that is the last term in (2.32).

Except for that this expression involves limits and the T terms, it is reminiscent of the expression based on the assumption of fixed regressors (2.26). In fact, cancelling T terms and adding β in (2.33)–(2.34) give

$$\hat{\beta} \xrightarrow{a} N(\beta, S_{xx}^{*-1} S S_{xx}^{*-1}), \quad (2.35)$$

where \xrightarrow{a} means “is asymptotically distributed as”. In practice we estimate this by using the information from the available sample: S_{xx}^* is estimated as $\hat{S}_{xx} = \sum_{t=1}^T x_t x_t'$ and \hat{S} as before, for instance, with White’s or Newey-West’s approach. This gives the same estimate of the variance-covariance matrix as in (2.26), which was derived under the assumption of fixed regressors.

Proof (of (2.35)*) Notice that $W = \text{Var}(\sqrt{T}\hat{\beta}) = T \text{Var}(\hat{\beta})$. Therefore, divide both sides of (2.34) by T to get

$$\text{Var}(\hat{\beta}) = \frac{1}{T} \left(\frac{S_{xx}^*}{T} \right)^{-1} \frac{S}{T} \left(\frac{S_{xx}^*}{T} \right)^{-1}.$$

Cancel the T terms. Finally, add β to shift the distribution from being centered on 0 to β . This is (2.35). \square

Further Reading

See, for instance, Verbeek (2017) 2, Greene (2018) 2-4 and Hansen (2022a) 4-5.

2.5 Appendix – A Primer in Matrix Algebra*

This appendix introduces fundamental concepts of matrix algebra.

2.5.1 Matrix and Scalar Addition and Multiplication

Let c be a scalar and define the matrices

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \text{ and } B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Multiplying a matrix by a scalar means multiplying each element by the scalar

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} c = \begin{bmatrix} A_{11}c & A_{12}c \\ A_{21}c & A_{22}c \end{bmatrix}.$$

Example 2.34 (*Matrix \times scalar*)

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} 10 = \begin{bmatrix} 10 & 30 \\ 30 & 40 \end{bmatrix}.$$

Adding/subtracting a scalar to each element of a matrix is done by

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + c J = \begin{bmatrix} A_{11} + c & A_{12} + c \\ A_{21} + c & A_{22} + c \end{bmatrix},$$

where J is a matrix (of the same size as A) filled with ones. This is sometimes written $A + c$, although that notation is not universally liked. In some applications, $\mathbf{1}_n$ (or just $\mathbf{1}$) is used to represent a vector of n ones.

Example 2.35 (*Matrix \pm scalar*)

$$\begin{aligned} \begin{bmatrix} 10 \\ 11 \end{bmatrix} - 10 \begin{bmatrix} 1 \\ 1 \end{bmatrix} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} + 10 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} &= \begin{bmatrix} 11 & 13 \\ 13 & 14 \end{bmatrix}. \end{aligned}$$

2.5.2 Adding and Multiplying: Two Matrices

Matrix *addition* (or subtraction) of matrices of the same size is element by element

$$A + B = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{bmatrix}.$$

Example 2.36 (*Matrix addition and subtraction*)

$$\begin{aligned} \begin{bmatrix} 10 \\ 11 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix} &= \begin{bmatrix} 8 \\ 6 \end{bmatrix} \\ \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & -2 \end{bmatrix} &= \begin{bmatrix} 2 & 5 \\ 6 & 2 \end{bmatrix} \end{aligned}$$

Matrix *multiplication* requires the two matrices to be conformable: the first matrix has as many columns as the second matrix has rows. Element ij of the result is the multiplication of the i th row of the first matrix with the j th column of the second matrix

$$AB = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Multiplying a square matrix A with a column vector z gives a column vector

$$Az = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} A_{11}z_1 + A_{12}z_2 \\ A_{21}z_1 + A_{22}z_2 \end{bmatrix}.$$

Example 2.37 (*Matrix multiplication*)

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & -2 \end{bmatrix} = \begin{bmatrix} 10 & -4 \\ 15 & -2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 17 \\ 26 \end{bmatrix}$$

2.5.3 Transpose

Transposing a column vector gives a row vector. Similarly, transposing a matrix is like flipping it around the main diagonal

$$A' = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}' = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}.$$

Example 2.38 (*Matrix transpose*)

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' = \begin{bmatrix} 10 & 11 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

2.5.4 Inner and Outer Products, Quadratic Forms

For two column vectors x and z , the product $x'z$ is called the *inner product* (a scalar)

$$x'z = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = x_1z_1 + x_2z_2,$$

and xz' the *outer product* (a matrix)

$$xz' = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} z_1 & z_2 \end{bmatrix} = \begin{bmatrix} x_1z_1 & x_1z_2 \\ x_2z_1 & x_2z_2 \end{bmatrix}.$$

(Notice that xz does not work for two vectors.)

Example 2.39 (*Inner and outer products*)

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 10 & 11 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} = 75$$

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix}' = \begin{bmatrix} 10 \\ 11 \end{bmatrix} \begin{bmatrix} 2 & 5 \end{bmatrix} = \begin{bmatrix} 20 & 50 \\ 22 & 55 \end{bmatrix}$$

If x is a column vector and A a square matrix, then the product $x'Ax$ is a quadratic form (a scalar).

Example 2.40 (*Quadratic form*)

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 10 \\ 11 \end{bmatrix} = 1244$$

2.5.5 Kronecker Product

Let \otimes represent the Kronecker product, that is, if A and B are matrices, then

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Example 2.41 (*Kronecker product*)

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \text{ and } B = \begin{bmatrix} 10 & 11 \end{bmatrix}, \text{ we get } A \otimes B = \begin{bmatrix} 10 & 11 & 30 & 33 \\ 20 & 22 & 40 & 44 \end{bmatrix}.$$

2.5.6 Matrix Inverse

A matrix *inverse* is the closest we get to “dividing” by a square matrix. The inverse of a matrix A , denoted A^{-1} , is such that

$$AA^{-1} = I \text{ and } A^{-1}A = I,$$

where I is the *identity matrix* (ones along the diagonal, and zeros elsewhere). The matrix inverse is useful for solving systems of linear equations, $y = Ax$ as $x = A^{-1}y$.

For a 2×2 matrix we have

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix}.$$

Example 2.42 (*Matrix inverse*) We have

$$\begin{bmatrix} -4/5 & 3/5 \\ 3/5 & -1/5 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ so}$$

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} -4/5 & 3/5 \\ 3/5 & -1/5 \end{bmatrix}.$$

2.5.7 Solving Systems of Linear Equations

If A is $n \times n$ and invertible and b and y are $n \times 1$ vectors, then we can solve

$$Ab = y \text{ as } b = A^{-1}y.$$

This solution is unique. In numerical applications, this system can often be solved (faster and with better precision) without the explicit matrix inverse.

Example 2.43 (*Solving a system of linear equations*)

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 11 \end{bmatrix}, \text{ gives}$$

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 10 \\ 11 \end{bmatrix} = \begin{bmatrix} -1.4 \\ 3.8 \end{bmatrix}.$$

2.5.8 OLS Notation: $X'X$ or $\sum_{t=1}^T x_t x_t'$?

Let x_t be a $K \times 1$ vector of (of data in period t). We can calculate the outer product ($K \times K$) as $x_t x_t'$ and summing each element across T observations gives the $K \times K$ matrix $S_{xx} = \sum_{t=1}^T x_t x_t'$.

Alternatively, let X be a $T \times K$ matrix with x_t' in row t . Then we can also calculate S_{xx} as $X'X$.

Example 2.44 (*Sum of outer product, $\sum_{t=1}^T x_t x_t'$*)

$$x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We then have

$$\begin{aligned}\sum_{t=1}^T x_t x_t' &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.\end{aligned}$$

In this example, the matrix happens to be diagonal, but that is not a general result. However, it will always be symmetric.

Example 2.45 (Sum of outer product, $X'X$) Define

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

It is straightforward to calculate that $X'X = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$.

Also, if y_t is a scalar, then $S_{xy} = \sum_{t=1}^T x_t y_t$ is a $K \times 1$ vector which is calculated as $X'Y$ where Y is a $T \times 1$ vector.

Example 2.46 With $(y_1, y_2, y_3) = (-1.5, -0.6, 2.1)$

$$\sum_{t=1}^T x_t y_t = \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-1.5) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (-0.6) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} 2.1 = \begin{bmatrix} 0 \\ 3.6 \end{bmatrix}$$

2.5.9 Derivatives of Matrix Expressions

Let z and x be $n \times 1$ vectors. The derivative of the inner product is $\partial(x'z)/\partial x = z$.

Example 2.47 (Derivative of an inner product) With $n = 2$

$$x'z = x_1 z_1 + x_2 z_2, \text{ so } \frac{\partial(x'z)}{\partial x} = \frac{\partial(z_1 x_1 + z_2 x_2)}{\begin{bmatrix} \partial x_1 \\ \partial x_2 \end{bmatrix}} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Let x be $n \times 1$ and A a symmetric $n \times n$ matrix. The *derivative of the quadratic form* is $\partial(x'Ax)/\partial x = 2Ax$.

Example 2.48 (*Derivative of a quadratic form*) With $n = 2$, the quadratic form is

$$x'Ax = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 A_{11} + x_2^2 A_{22} + 2x_1 x_2 A_{12}.$$

The derivatives with respect to x_1 and x_2 are

$$\begin{aligned} \partial(x'Ax)/\partial x_1 &= 2x_1 A_{11} + 2x_2 A_{12} \text{ and } \partial(x'Ax)/\partial x_2 = 2x_2 A_{22} + 2x_1 A_{12}, \text{ or} \\ \partial(x'Ax)/\partial x &= \begin{bmatrix} \partial(x'Ax)/\partial x_1 \\ \partial(x'Ax)/\partial x_2 \end{bmatrix} = 2 \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

2.4 Appendix – A Primer in Calculus*

The following derivatives (with respect to x) are often used in these chapters

$$\begin{aligned} \frac{d}{dx}(ax^k + bx) &= akx^{k-1} + b \\ \frac{d}{dx} \ln x &= 1/x \\ \frac{d}{dx} e^x &= e^x. \end{aligned}$$

Derivatives typically depend on at which x value we evaluate them at ($x = 1$ or $x = 2$, say), so the derivatives are themselves functions. The first expression embeds the *sum rule* (the derivative of a sum is the sum of the derivatives).

Example 2.49 (*Derivative of power function*) $3x^2 + 7x$ has the derivative $6x + 7$ which is -5 at $x = -2$ and 13 at $x = 1$.

The *chain rule* says that if $g()$ and $f()$ are two functions, then the derivative of the composite function $g(f(x))$ is

$$\frac{d}{dx} g(f(x)) = g'(u)f'(x), \text{ where } u = f(x),$$

and where $g'(u)$ is short hand (Lagrange's) notation for $\frac{d}{du}g(u)$, and similarly for $f'(x)$. The derivative $g'(u)$ is often referred to as the outer derivative and $f'(x)$ as the inner derivative.

Example 2.50 (*Chain rule*) Let $g(u) = u^2$ and $u = f(x) = 2 - 3x$, so we are considering the composite function $(2 - 3x)^2$. We then get

$$\frac{d}{dx}(2 - 3x)^2 = \underbrace{2(2 - 3x)}_{g'(u)} \underbrace{(-3)}_{f'(x)} = 18x - 12.$$

This derivative is -12 at $x = 0$ and 6 at $x = 1$.

Consider a function of two variables, $f(x, z)$. The *partial derivative* with respect to x is just a standard derivative, treating z as fixed. For instance,

$$\begin{aligned}\frac{\partial}{\partial x} ax^k bz &= akx^{k-1}bz \\ \frac{\partial}{\partial z} ax^k bz &= ax^k b.\end{aligned}$$

Suppose the function $f(x)$ gives a scalar output, but x is a n -vector of inputs (with elements x_1, x_2, \dots, x_n). The *gradient* is then

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n}. \end{bmatrix}$$

Similarly, $\frac{\partial f(x)}{\partial x'}$ is the transpose of this expression.

Example 2.51 (*Gradient*) For the function $f(x) = (x_1 - 2)^2 + (4x_2 + 3)^2$, the gradient is

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} 2(x_1 - 2) \\ 8(4x_2 + 3) \end{bmatrix}.$$

The *Hessian* is the $n \times n$ matrix of second derivatives

$$\frac{\partial^2 f(x)}{\partial x \partial x'} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ & \ddots & \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

(In case the derivatives are continuous, then this matrix is symmetric.)

(Hessian) Using the same function as in Example 2.51, we get

$$\frac{\partial^2 f(x)}{\partial x \partial x'} = \begin{bmatrix} 2 & 0 \\ 0 & 32 \end{bmatrix}.$$

A *first-order Taylor approximation* of a differentiable function $f(x)$ is

$$f(b) \approx f(a) + \frac{\partial f(a)}{\partial x} (b - a),$$

where the derivative is evaluated as $x = a$. For highly non-linear functions, this only works well when a and b are close. For a vector of functions (which depend on a vector of variables x), we instead have

$$\begin{bmatrix} f_1(b) \\ \vdots \\ f_n(b) \end{bmatrix} \approx \begin{bmatrix} f_1(a) \\ \vdots \\ f_n(a) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1(a)}{\partial x_1} & \dots & \frac{\partial f_1(a)}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(a)}{\partial x_1} & \dots & \frac{\partial f_n(a)}{\partial x_m} \end{bmatrix} \begin{bmatrix} b_1 - a_1 \\ \vdots \\ b_m - a_m \end{bmatrix} \text{ or}$$

$$f(b) \approx f(a) + \frac{\partial f(a)}{\partial x'} (b - a).$$

See Figure 2.12 for an illustration.

Example 2.52 (Taylor approximation) Let $f(x) = \ln x$ and consider $(a, b) = (1, 1.2)$, so $f(1.2) \approx 0 + 1 \times 0.2 = 0.2$, when the true value is approximately 0.18. Instead, with $b = 2$, we get the approximation 1 and the true value around 0.69, so the error is considerable.

A related concept is the *mean-value theorem* which says that for a differentiable function $f(x)$,

$$f(b) = f(a) + \frac{\partial f(c)}{\partial x} (b - a),$$

where the derivative is evaluated at value $x = c$ between a and b . A similar expression holds for a vector of functions

$$f(b) = f(a) + \frac{\partial f(c)}{\partial x'} (b - a).$$

Again, see Figure 2.12 for an illustration.

Example 2.53 (Mean-value theorem) Let $f(x) = \ln x$ and consider $(a, b) = (1, 2)$. With $c \approx 1.443$, we have $0.693 \approx 0 + \frac{1}{1.443}(2 - 1)$.

2.5 Appendix – A Primer in Optimization*

Remark 2.54 (First order condition for optimising a differentiable function). We want to find the value of b in the interval $b_{\text{low}} \leq b \leq b_{\text{high}}$, which makes the value of the

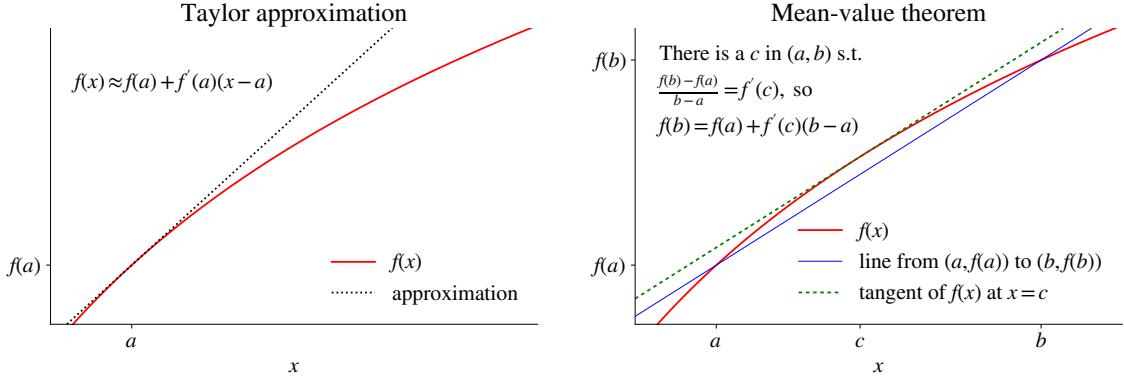


Figure 2.12: Illustration of a Taylor approximation and the mean-value theorem.

differentiable function $f(b)$ as small as possible (a minimization problem). The answer is b_{low} , b_{high} , or a value of b where $df(b)/db = 0$. The latter is a necessary and sufficient condition for an unconstrained problem where $f(b)$ is convex. (If the function is twice differentiable, then convexity means that $f''(b) \geq 0$.) A maximization problem, except that we rather want $f(b)$ to be concave ($f''(b) \leq 0$).

When the goal is to determine x and y that *minimize*

$$L = (x - 2)^2 + (4y + 3)^2,$$

then we have to find the values of x and y that satisfy the *first order conditions* $\partial L / \partial x = 0, \partial L / \partial y = 0$. For L function above, these are

$$\begin{aligned} 0 &= \partial L / \partial x = 2(x - 2) \\ 0 &= \partial L / \partial y = 8(4y + 3), \end{aligned}$$

which clearly requires $x = 2$ and $y = -3/4$. In this particular case, the first order condition with respect to x does not depend on y , but that is not a general property. See Figure 2.13 for the surface of the loss function and the contours.

Also, in this case, there is a unique solution—but in more complicated problems, the first order conditions could be satisfied at different values of x and y .

A *maximization problem* has the same type of first order conditions.

If you want to add a *restriction* to the minimization problem, say

$$x + 2y = 3,$$

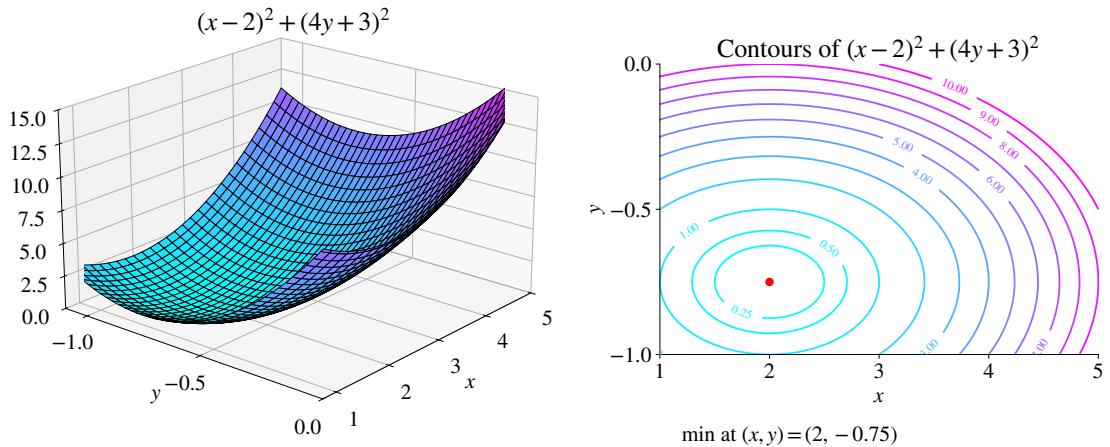


Figure 2.13: Minimization problem

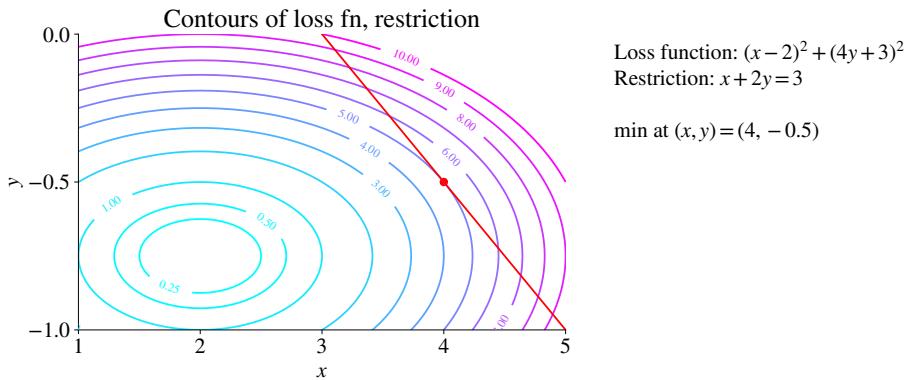


Figure 2.14: Minimization problem with restriction

then we can proceed in two ways. The first is to simply substitute for $x = 3 - 2y$ in L to get

$$L = (1 - 2y)^2 + (4y + 3)^2,$$

with first order condition

$$0 = \partial L / \partial y = -4(1 - 2y) + 8(4y + 3) = 40y + 20,$$

which requires $y = -1/2$, which by implies $x = 4$. (We could equally well have substituted for y). This is also the unique solution. See Figure 2.14. This is an easy way to eliminate an equality restriction.

The second method is to use a *Lagrangian*. The problem is then to choose x , y , and

λ to minimize

$$L = (x - 2)^2 + (4y + 3)^2 + \lambda(x + 2y - 3).$$

(If you instead use $-\lambda()$ or write the restriction as $-x - 2y + 3$, you should get the same result. The interpretation of λ differs, though.)

The term multiplying λ is the restriction. The first order conditions are now

$$\begin{aligned} 0 &= \partial L / \partial x = 2(x - 2) + \lambda \\ 0 &= \partial L / \partial y = 8(4y + 3) + 2\lambda \\ 0 &= \partial L / \partial \lambda = x + 2y - 3. \end{aligned}$$

These are 3 equations in 3 unknowns (x, y, λ) which we have to solve. One way is as follows. The first two conditions say

$$\begin{aligned} x &= 2 - \lambda/2 \\ y &= -3/4 - \lambda/16, \end{aligned}$$

so we need to find λ . To do that, use these latest expressions for x and y in the third first order condition (to substitute for x and y)

$$\begin{aligned} 3 &= 2 - \lambda/2 - 3/2 - \lambda/8 = 1/2 - \lambda/5/8, \text{ so} \\ \lambda &= -4. \end{aligned}$$

Finally, use this to calculate x and y as

$$x = 4 \text{ and } y = -1/2.$$

Notice that this is the same solution as before ($y = -1/2$) and that the restriction holds ($4 + 2(-1/2) = 3$). This second method is clearly a lot clumsier in my example, but it pays off when there are several restrictions and/or when the restriction(s) become complicated.

Chapter 3

Betas and Index Models

3.1 Single-Index Models

The *market model* is a single-index model where the time series of returns of an asset (R_{it}) is regressed on the series of the “market” return (R_{mt})

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it}, \text{ where} \quad (3.1)$$
$$\mathbb{E} \varepsilon_{it} = 0, \text{ Cov}(\varepsilon_{it}, R_{mt}) = 0.$$

This regression may use the net returns (as indicated above), or the returns in excess of a risk-free rate. The results for the β (which is the focus here) are typically very similar.

The two assumptions are the standard assumptions for using ordinary least squares: the residual has a zero mean and is uncorrelated with the non-constant regressor.

Remark 3.1 *(A warning about the notation) When discussing OLS we typically write the regression equation as $y_t = x_t' \beta + u_t$. Comparing with (3.1), we notice that $y_t = R_{it}$. Also, x_t equals the column vector $[1, R_{mt}]'$ and $u_t = \varepsilon_{it}$. Finally, notice that we recycle the β symbol: in OLS it is a vector of all coefficients, corresponding to $[\alpha_i, \beta_i]'$ in the index model.*

Empirical Example 3.2 *(Betas of industry portfolios) See Figure 3.2 and Tables 3.1–3.2 for results on U.S. industry portfolios.*

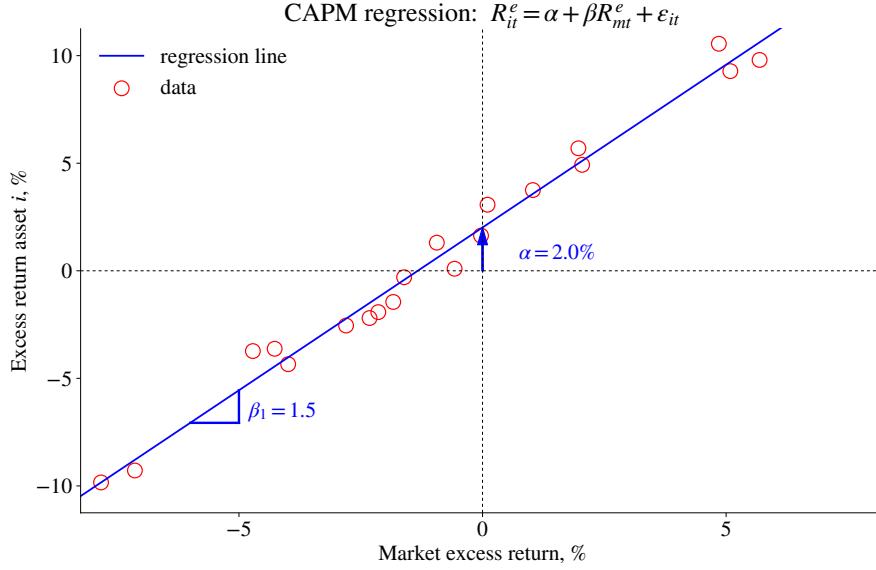


Figure 3.1: CAPM regression

	NoDur	Durbl	Manuf	Enrgy	HiTec
c	0.20 (2.03)	-0.02 (-0.11)	-0.01 (-0.14)	0.18 (0.93)	-0.05 (-0.41)
R_m^e	0.76 (27.51)	1.25 (21.67)	1.03 (51.71)	0.86 (16.07)	1.24 (39.02)
R^2	0.66	0.57	0.87	0.40	0.76
obs	660	660	660	660	660

Table 3.1: CAPM regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats.

3.2 Estimating Beta

3.2.1 Estimating Historical Beta: OLS and Other Approaches

It is sometimes argued that the OLS estimate of beta on a historical sample may not be the best forecast of the beta for a future time periods (see, for instance, Blume (1971)). As a potential solution, we could apply a shrinkage towards the average beta, which is 1,

$$\beta = \eta \hat{\beta}_{OLS} + (1 - \eta)1. \quad (3.2)$$

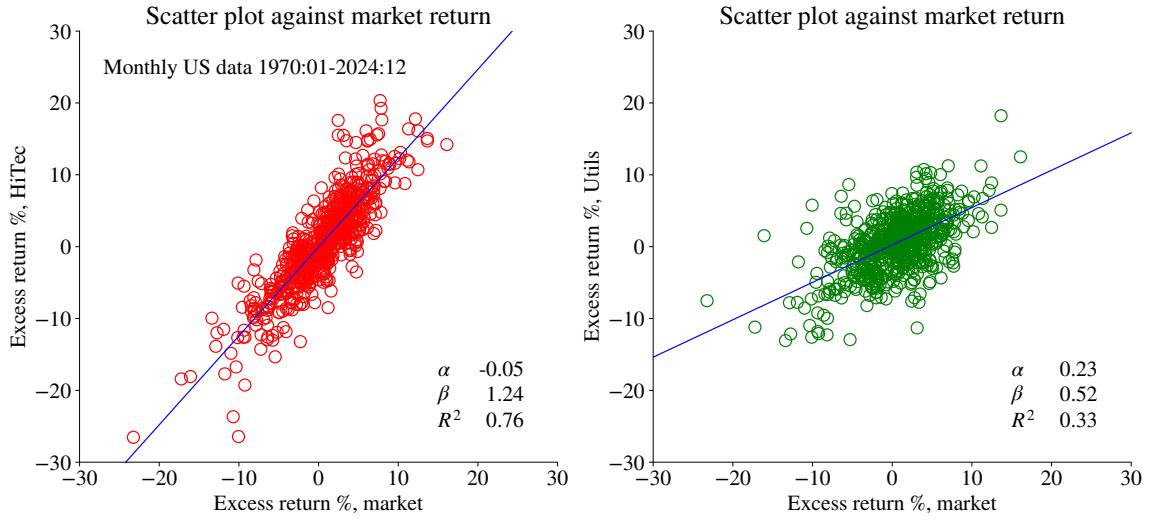


Figure 3.2: Scatter plot against market return

	Telcm	Shops	Hlth	Utils	Other
c	0.05 (0.38)	0.11 (1.14)	0.15 (1.23)	0.23 (1.70)	-0.07 (-0.83)
R_m^e	0.79 (24.75)	1.00 (33.73)	0.81 (21.89)	0.52 (14.57)	1.09 (53.36)
R^2	0.57	0.76	0.58	0.33	0.86
obs	660	660	660	660	660

Table 3.2: CAPM regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats.

This could be motivated by empirical findings or by a Bayesian principle (see Greene (2018) 16). (In the latter case, η would be higher if the sample is long and the fit is good.)

Empirical Example 3.3 *Table 3.3 for an evaluation of several methods: most are of the form (3.2), but we also consider a method where the OLS estimated is replaced by an exponentially weighted moving average estimate (EWMA), which is just a weighted OLS where an observation s periods ago gets the weight λ^s where λ is close to one (for instance, 0.95).*

	OLS adj $0.67\hat{b} + 0.33$	OLS adj $0.5\hat{b} + 0.5$	OLS adj $0.33\hat{b} + 0.67$	1	EWMA $0.5\hat{b} + 0.5$
error in β	-6.6	-5.9	-2.4	11.0	-10.5

Table 3.3: Absolute forecast errors of future betas, as a percentage difference to OLS: the average $|\text{next 2 year } \beta - \text{predicted } \beta|$ compared to the results from OLS. A negative number is better performance than OLS. The models are estimated on moving 10-year windows and EWMA uses $\lambda = 0.95$. 49 industry portfolios, monthly data for 1970:01-2024:12.

3.2.2 Fundamental Betas

Another way to improve the forecasts of beta over a future period is to incorporate information about fundamental firm variables. This is particularly useful when there is little historical data on returns, for instance, because the asset was not traded before.

It is often found that betas are related to fundamental variables as follows (with signs in parentheses indicating the effect on the beta): Dividend payout (-), Asset growth (+), Leverage (+), Liquidity (-), Asset size (-), Earning variability (+), Earnings Beta (slope in earnings regressed on economy wide earnings) (+). Such relations can be used to make an educated guess about the beta of an asset without historical data on the returns—but with data on (at least some of) these fundamental variables.

3.3 Multi-Index Models

3.3.1 Overview

The multi-index model is just a multivariate extension of the single-index model

$$R_{it} = a_i + b_i' f_t + \varepsilon_{it}, \text{ where} \quad (3.3)$$

$$\mathbb{E} \varepsilon_{it} = 0, \text{ Cov}(\varepsilon_{it}, f_t) = \mathbf{0}.$$

As an example, there could be two indices: the stock market return and an interest rate.

It is often found that while several indices are needed for a reasonable approximation, a single-index model performs equally well in forecasting the covariance pattern over a future period. This is much like the classical trade-off between in-sample fit which requires a large model, and forecasting which is often better with a small model.

Remark 3.4 (*Fama-French factors*) *Fama and French (1993)* use three factors: the market

excess return, the return on a portfolio of small stocks minus the return on a portfolio of big stocks (*SMB*), and the return on a portfolio with a high ratio of book value to market value minus the return on a portfolio with a low ratio (*HML*). All three are excess returns (although only the first is in excess of a risk-free return), since they are long-short portfolios.

Empirical Example 3.5 (3-factor model for the 10 industry portfolios) See Tables 3.4 –3.5 for regressions of 10 industry portfolios on the three Fama-French factors.

	NoDur	Durbl	Manuf	Enrgy	HiTec
c	0.15 (1.52)	-0.14 (-0.77)	-0.08 (-1.15)	-0.01 (-0.07)	0.14 (1.32)
R_m^e	0.80 (30.38)	1.25 (22.12)	1.06 (57.45)	0.95 (17.61)	1.13 (38.20)
R_{SMB}	-0.12 (-2.88)	0.24 (2.74)	0.01 (0.44)	-0.09 (-1.40)	0.18 (3.76)
R_{HML}	0.15 (3.50)	0.33 (3.92)	0.19 (6.51)	0.50 (5.38)	-0.51 (-11.16)
R^2	0.68	0.59	0.88	0.46	0.82
obs	660	660	660	660	660

Table 3.4: Fama-French regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats.

	Telcm	Shops	Hlth	Utils	Other
c	-0.01 (-0.08)	0.13 (1.25)	0.25 (2.10)	0.11 (0.87)	-0.21 (-3.28)
R_m^e	0.84 (26.97)	0.98 (31.56)	0.81 (24.70)	0.60 (17.75)	1.14 (61.07)
R_{SMB}	-0.16 (-3.57)	0.08 (1.38)	-0.22 (-4.05)	-0.20 (-4.08)	0.01 (0.37)
R_{HML}	0.15 (3.04)	-0.03 (-0.51)	-0.26 (-4.66)	0.31 (5.78)	0.40 (13.21)
R^2	0.59	0.76	0.62	0.41	0.91
obs	660	660	660	660	660

Table 3.5: Fama-French regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats.

3.3.2 Multi-Index Model as a Method for Portfolio Choice

The factor loadings (betas) are often used directly in portfolio choice. The reason is simple: the betas summarize how different assets are exposed to the big risk factors/return drivers. The betas therefore provide a way of understanding the broad features of even complicated portfolios.

3.4 Principal Component Analysis*

Principal component analysis (PCA) can help us determine how many factors that are needed to explain a cross-section of asset returns.

Let $z_t = R_t - \bar{R}_t$ be an $n \times 1$ vector of demeaned returns with covariance matrix Σ . The first principal component (pc_{1t}) is the (normalized) linear combination of z_t that accounts for as much of the variability as possible, and we denote its variance λ_1 . The j th ($j \geq 2$) principal component (pc_{jt}) is similar (and its variance is denoted λ_j), except that it must be uncorrelated with all lower principal components. Remark 3.6 gives a formal definition.

Remark 3.6 (*Principal component analysis*) Consider the zero mean $N \times 1$ vector z_t with covariance matrix Σ . The first (sample) principal component is $pc_{1t} = w_1' z_t$, where w_1 is the eigenvector associated with the largest eigenvalue (λ_1) of Σ . This value of w_1 solves the problem $\max_w w' \Sigma w$ subject to the normalization $w'w = 1$. The eigenvalue λ_1 equals $\text{Var}(pc_{1t}) = w_1' \Sigma w_1$. The j th principal component solves the same problem, but under the additional restriction that $w_i' w_j = 0$ for all $i < j$. The solution is the eigenvector associated with the j th largest eigenvalue λ_j (which equals $\text{Var}(pc_{jt}) = w_j' \Sigma w_j$).

Let the i th eigenvector be the i th column of the $n \times n$ matrix

$$W = [w_1 \ \cdots \ w_n]. \quad (3.4)$$

We can then calculate the $n \times 1$ vector of principal components as

$$pc_t = W' z_t. \quad (3.5)$$

Since the eigenvectors are orthogonal it can be shown that $W' = W^{-1}$, so the expression can be inverted as

$$z_t = W pc_t. \quad (3.6)$$

This shows that the i th eigenvector (the i th column of W) can be interpreted as the effect of the i th principal component on each of the elements in z_t . However, the sign of column j of W can be changed without any effects (except that the pc_{jt} also changes sign), so we can always reinterpret a negative coefficient as a positive exposure (to $-pc_{jt}$).

Empirical Example 3.7 (*PCA for the 25 FF portfolios*) See Figure 3.3 for the three most important eigenvectors for the 25 FF portfolios. The first pc seems to capture the average (market), while the other are related to size and growth/value.

Example 3.8 (*PCA with 2 series*) Let $w_i^{(j)}$ be the t th element in the i th eigenvector. With two series we have

$$\begin{aligned} pc_{1t} &= \begin{bmatrix} w_1^{(1)} \\ w_1^{(2)} \end{bmatrix}' \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ and } pc_{2t} = \begin{bmatrix} w_2^{(1)} \\ w_2^{(2)} \end{bmatrix}' \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ or} \\ \begin{bmatrix} pc_{1t} \\ pc_{2t} \end{bmatrix} &= \begin{bmatrix} w_1^{(1)} & w_2^{(1)} \\ w_1^{(2)} & w_2^{(2)} \end{bmatrix}' \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ and} \\ \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} &= \begin{bmatrix} w_1^{(1)} & w_2^{(1)} \\ w_1^{(2)} & w_2^{(2)} \end{bmatrix} \begin{bmatrix} pc_{1t} \\ pc_{2t} \end{bmatrix} \end{aligned}$$

For instance, for the two elements in the second eigenvector, $w_2^{(1)}$ shows how pc_{2t} affects z_{1t} , while $w_2^{(2)}$ shows how the same pc_{2t} affects z_{2t} .

Remark 3.9 (*Data in matrices**) Transpose (3.5) to get $pc_t' = z_t' W$, where the dimensions are $1 \times n$, $1 \times n$ and $n \times n$ respectively. If we form a $T \times n$ matrix of data Z by putting z_t in row t , then the $T \times N$ matrix of principal components can be calculated as $PC = ZW$.

Notice that (3.6) shows that all n data series in z_t can be written in terms of the n principal components. Since the principal components are uncorrelated ($\text{Cov}(pc_{it}, pc_{jt}) = 0$)), and it can be shown that $\sum_{i=1}^n \lambda_i = \sum_{i=1}^n \text{Var}(z_{it})$, we can think of the sum of their variances ($\sum_{i=1}^n \lambda_i$) as the “total variation” of the series in z_t . In practice, it is common to report the relative importance of principal component j as

$$\text{relative importance of } pc_j = \lambda_j / \sum_{i=1}^n \lambda_i. \quad (3.7)$$

For instance, if the first two principal components account for 75% of the total variation among many asset returns, a two-factor model is likely to provide a good approximation.

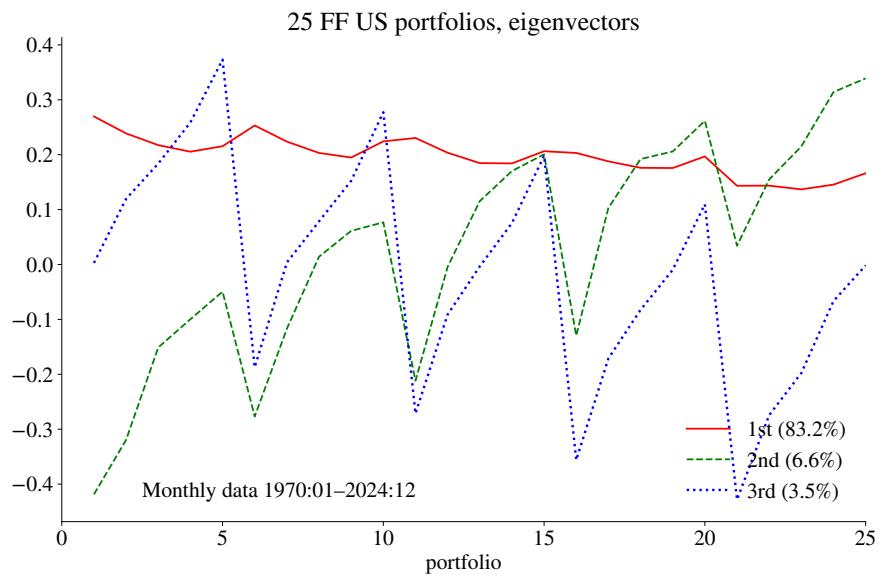


Figure 3.3: Eigenvectors for US portfolio returns

Further Reading

See Elton, Gruber, Brown, and Goetzmann (2014) 7–8 for discussions of how to estimate and use beta estimates.

Chapter 4

Least Squares: Testing

4.1 Testing a Single Coefficient: A t -test

We are interested in testing the null hypothesis (H_0) that a single coefficient $\beta = q$, where q is a number of interest. (Econometric programs typically report results for $H_0: \beta = 0$.) Here, the alternative hypothesis is that $\beta \neq q$, so this is a two-sided (also called “two-tailed”) test.

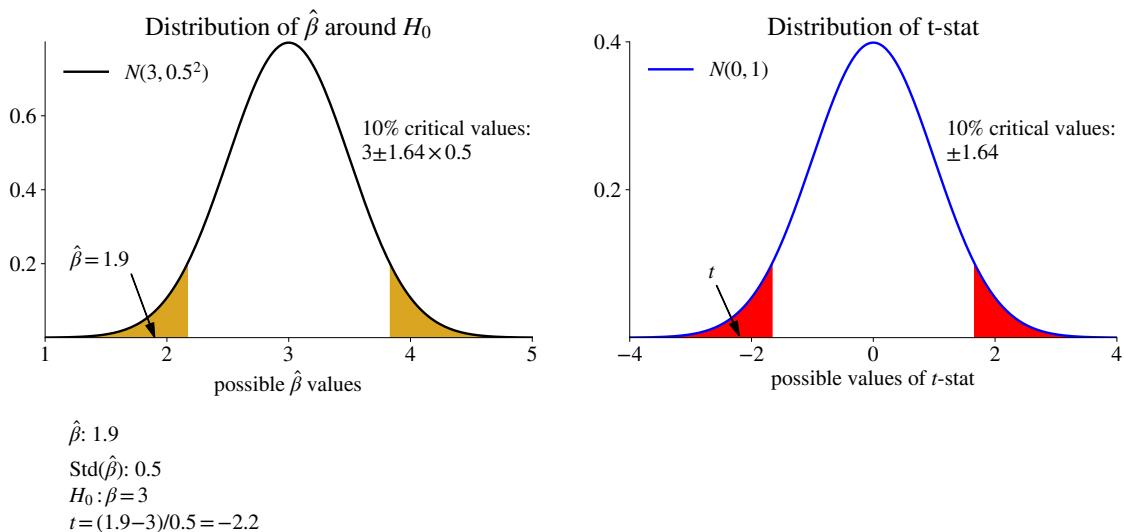


Figure 4.1: Distribution of $\hat{\beta}$ and t-stat

We assume that the estimates are normally distributed, which may be a good approximation when the sample is large, because of the central limit theorem. If the null hypothesis is true, then

$$\hat{\beta} \sim N(q, \text{Var}(\hat{\beta})). \quad (4.1)$$

The estimate $\hat{\beta}$ could be from an OLS regression or from some other method, and the variance $\text{Var}(\hat{\beta})$ depends on the properties of data (heteroskedasticity/autocorrelation) and on the estimation method.

To be able to easily compare with printed tables of probabilities, we transform to a $N(0, 1)$ variable. In particular, if the true coefficient is really q , then $\hat{\beta} - q$ should have a zero mean. Dividing by the standard error (deviation) of $\hat{\beta}$, we should have

$$t = (\hat{\beta} - q) / \text{Std}(\hat{\beta}) \sim N(0, 1) \quad (4.2)$$

We reject the null hypothesis when $|t|$ is very large, for instance, if $|t| > 1.64$.

This decision is driven by (a) how far $\hat{\beta}$ is from q ; (b) how uncertain $\hat{\beta}$ is (as measured by $\text{Std}(\hat{\beta})$); (c) and how we define the cutoff (here 1.64). The latter is typically done by first choosing a *significance level* (for instance, 10%) which defines a *critical value* (1.64 for the 10% significance level): reject the null hypothesis if $|t|$ is larger than the critical value (1.64 on the 10% level, 1.96 on the 5% level). See Figure 4.2 for an illustration of the probabilities according to an $N(0, 1)$ distribution.

The significance level represents the probability, in a random sample, of falsely rejecting a null hypothesis that is actually true. See Figure 4.1 for an illustration. A lower significance level (5%, which corresponds to a critical value of 1.96) is therefore a more conservative test in the sense that we require stronger evidence to reject the null hypothesis, and thus the risk of a false rejection is lower. Therefore, the significance level is a trade-off between actually being able to reject the null hypothesis and sometimes doing it wrongly.

Otherwise, when $|t|$ is not very large (for instance, $|t| < 1$), then evidence is not sufficient to reject the null hypothesis. (You may compare with a court of law where the null hypothesis is that the accused is not guilty.)

Example 4.1 (*t-test*) Let $\hat{\beta} = 1.9$, $\text{Std}(\hat{\beta}) = 0.5$ and $q = 3$. Then, $t = (1.9 - 3)/0.5 = -2.2$ so $|t| > 1.64$ and also $|t| > 1.96$. The null hypothesis is thus rejected at both the 10% and the 5% significance levels.

Empirical Example 4.2 (*CAPM regressions for industry portfolios*) See Table 4.1.

Empirical Example 4.3 (*Multi-factor regressions for industry portfolios*) See Table 4.2.

The *p-value* is a related concept. It is the lowest significance level at which we can reject the null hypothesis: a lower number is a stronger rejection. See Figure 4.3 for an illustration.

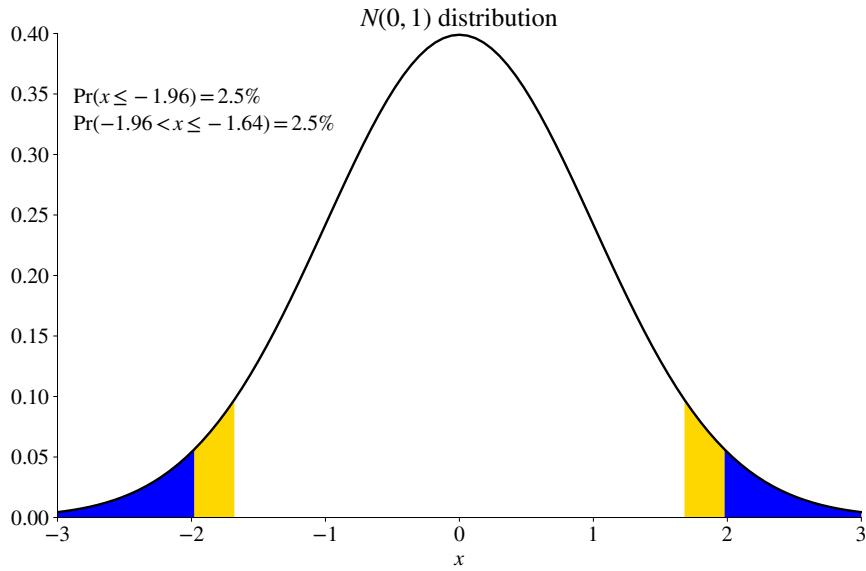


Figure 4.2: Density function of a standard normal distribution

Example 4.4 (*p*-value) Continuing Example 4.1, notice that according to an $N(0, 1)$ -distribution, the probability of -2.2 or lower is 1.4% , so the *p*-value is 2.8% . We thus reject the null hypothesis at the 10% significance level and also at the 5% significance level.

We sometimes compare with a *t*-distribution instead of an $N(0, 1)$ -distribution, especially when the sample is small. For instance, with 22 data points and two estimated coefficients (so there are 20 degrees of freedom), the 10% critical value of a *t*-distribution is 1.72 (while it is 1.64 for the standard normal distribution). However, for samples of more than 30–40 data points, the difference is trivial.

Remark 4.5 (*One-sided test**) As an example of a one-sided test let $H_0 : \beta \leq q$ and $H_1 : \beta > q$. Sometimes the null hypothesis is written $\beta = q$, but that makes little practical difference. We then reject the null hypothesis at the 10% significance level if $t > 1.28$ which is the 0.90 quantile of a $N(0, 1)$ distribution. Conversely, when $H_0 : \beta \geq q$ and $H_1 : \beta < q$, then we reject the null hypothesis if $t < -1.28$. Since 1.28 is the 20% critical value in a two-sided test, we can actually use a two sided test (for instance, from a regression package) to also do a one-sided test: (a) if we reject the null hypothesis on the 20% level in a double sided test; (b) and the sign is right; then (c) this is a rejection on the 10% level in a one-sided test.

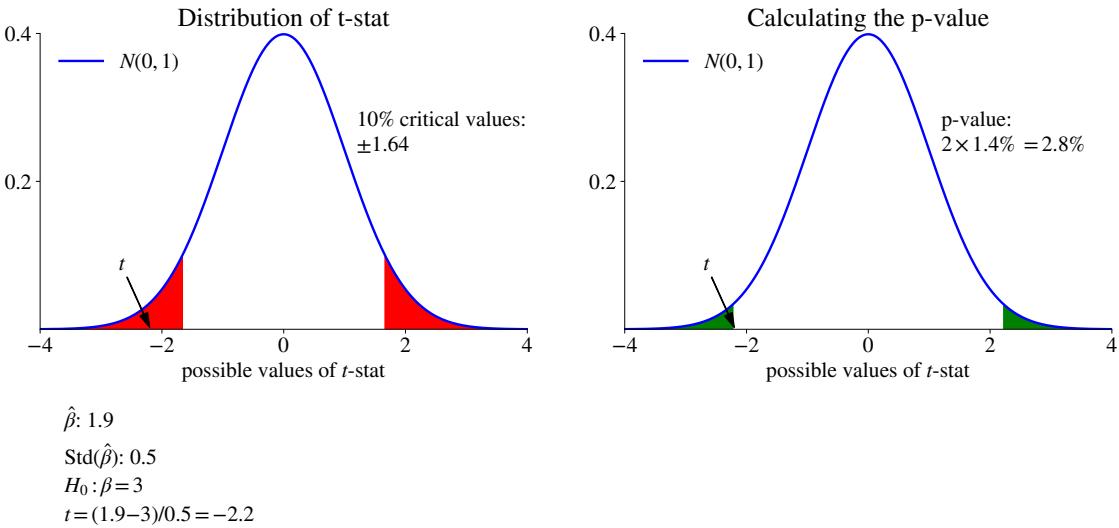


Figure 4.3: Calculating the p-value

4.2 Confidence Bands

A significance level of 10% means that there is, if the null hypothesis is true, a 90% probability that the t value in (4.2) from a random sample is within the interval (band) $(-1.64, 1.64)$, that is,

$$\Pr(-1.64 \leq t \leq 1.64) = 90\%. \quad (4.3)$$

The t -test discussed above rejects the null hypothesis ($\beta = q$) when t is outside this confidence band. Notice that

$$t \text{ is outside } [-1.64, 1.64] \iff \quad (4.4)$$

$$\hat{\beta} \text{ is outside } [q - 1.64 \text{ Std}(\hat{\beta}), q + 1.64 \text{ Std}(\hat{\beta})] \text{ and} \quad (4.5)$$

$$q \text{ is outside } [\hat{\beta} - 1.64 \text{ Std}(\hat{\beta}), \hat{\beta} + 1.64 \text{ Std}(\hat{\beta})]. \quad (4.6)$$

The interval in (4.5) is a 90% confidence band of β *centered on the null hypothesis*, while the confidence band in (4.6) is *centered on the point estimate*. These are alternative ways of doing a hypothesis test. See Figures 4.1 and 4.4 for illustrations.

Proof (that t and $\hat{\beta}$ are outside their confidence bands at the same time) For $\hat{\beta}$ to be outside the band we must have

$$\hat{\beta} < q - 1.64 \text{ Std}(\hat{\beta}) \text{ or } \hat{\beta} > q + 1.64 \text{ Std}(\hat{\beta}).$$

Rearrange this by subtracting q from both sides of the inequalities and then divide both

	HiTec	Utils
constant	-0.05 (-0.41)	0.23 (1.70)
market return	1.24 (39.02)	0.52 (14.57)
R^2	0.76	0.33
Autocorr	0.32	0.93
White	0.05	0.00
All slopes	0.00	0.00
obs	660	660

Table 4.1: CAPM regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats. Autocorr is the p-value for no autocorrelation; White is the p-value for homoskedasticity; All slopes is the p-value for all slope coefficients being zero.

sides by $\text{Std}(\hat{\beta})$

$$(\hat{\beta} - q) / \text{Std}(\hat{\beta}) < -1.64 \text{ or } (\hat{\beta} - q) / \text{Std}(\hat{\beta}) > 1.64.$$

□

Example 4.6 (*t-test and confidence band around q*) With $\text{Std}(\hat{\beta}) = 0.5$ and $q = 3$, the 90% confidence band is $3 \pm 1.64 \times 0.5$, that is, $[2.18, 3.82]$. Notice that $\hat{\beta} = 1.90$ is outside this band, so we reject the null hypothesis. Equivalently, $t = (1.9 - 3)/0.5 = -2.2$ is outside the band $[-1.64, 1.64]$.

Example 4.7 (*t-test and confidence band around $\hat{\beta}$*) With $\text{Std}(\hat{\beta}) = 0.5$ and $\hat{\beta} = 1.9$, the 90% confidence band is $1.9 \pm 1.64 \times 0.5$, that is, $[1.08, 2.72]$. Notice that $q = 3$ is outside this band, so we reject the null hypothesis.

4.3 Power and Size*

The *size* is the probability of rejecting a true H_0 , sometimes called the type I error. That is, the probability of a wrong rejection It should be low. Provided you use a valid test (correct standard error, etc), the *size* is the significance level you have chosen (which defines the critical values). For instance, with a *t*-test with critical values $(-1.64, 1.64)$, the size is 10%. This means that we run a 10% chance of wrongly rejecting a true null hypothesis. See Table 4.3.

	HiTec	Utils
constant	0.14 (1.32)	0.11 (0.87)
market return	1.13 (38.20)	0.60 (17.75)
SMB	0.18 (3.76)	-0.20 (-4.08)
HML	-0.51 (-11.16)	0.31 (5.78)
R^2	0.82	0.41
Autocorr	0.54	0.60
White	0.00	0.00
All slopes	0.00	0.00
obs	660	660

Table 4.2: Fama-French regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats. Autocorr the p-value for no autocorrelation; White is the p-value for homoskedasticity; All slopes is the p-value for all slope coefficients being zero.

	H_0 not rejected	H_0 rejected
H_0 is true	1 - size	size
H_0 is false	1 - power	power

Table 4.3: Size and power

The *power* is the probability of rejecting a false H_0 . That is, the probability of a correct rejection. It should be high. Typically, it cannot be controlled, but some tests are better than others. This power depends on how false H_0 is, a fact we will never know. All we can do is to create (artificial) examples to get an idea of what the power would be for different tests and for different values of the true parameter β . For instance, with a t -test using the critical values -1.64 and 1.64 , the power would be

$$\text{power} = \Pr(t \leq -1.64) + \Pr(t \geq 1.64). \quad (4.7)$$

($1 - \text{power}$ is sometimes called the type II error. This is the probability of not rejecting a false H_0 .) Again, see Table 4.3.

To make this more concrete, suppose we test the null hypothesis that the coefficient is equal to q , but the true value happens to be β . Since the OLS estimate, $\hat{\beta}$ is distributed as

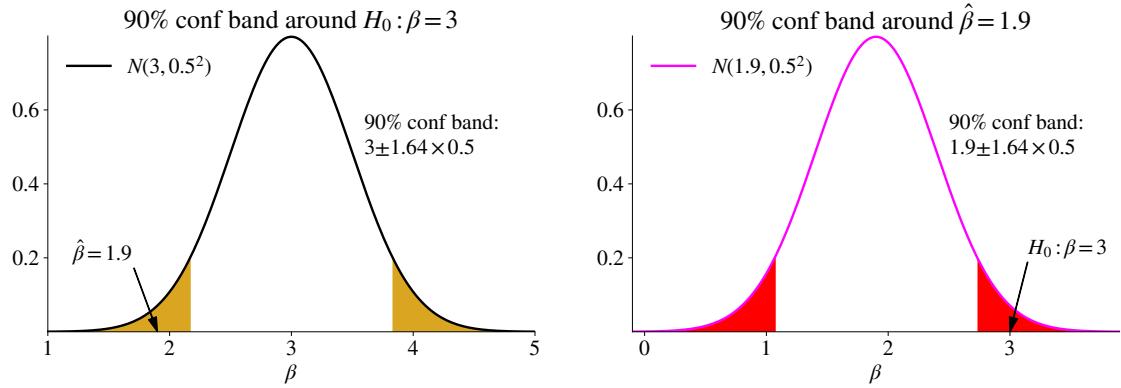


Figure 4.4: Confidence band around the null hypothesis or around the point estimate

$N(\beta, \text{Std}(\hat{\beta}))$, it must be the case that the t -stat is distributed as

$$t = (\hat{\beta} - q) / \text{Std}(\hat{\beta}) \sim N\left((\beta - q) / \text{Std}(\hat{\beta}), 1\right). \quad (4.8)$$

We can then calculate the power as the probability that $t \leq -1.64$ or $t \geq 1.64$, when t has the distribution on the RHS in (4.8). Clearly, the results depend on what the true value β really is. See Figure 4.5.

Example 4.8 If $\beta = 3.6$, $q = 3$ and $\text{Std}(\hat{\beta}) = 0.5$, then the power is 0.33.

4.4 Testing A Linear Combination

We can form a linear combination of the regressions coefficients and apply a t-test.

Let R be a $1 \times k$ (row) vector that defines our linear combination and suppose we want to test $R\beta = q$, where β is k -vector of coefficients. The estimate $\hat{\beta}$ could be from a linear regression, but could equally well be from some other method. We can test the hypothesis by noting that

$$\text{Var}(R\hat{\beta}) = R \text{Var}(\hat{\beta}) R', \quad (4.9)$$

is a scalar, although $\text{Var}(\hat{\beta})$ is a $k \times k$ matrix. Therefore, the t-test becomes

$$t = (R\hat{\beta} - q) / (R \text{Var}(\hat{\beta}) R')^{1/2} \sim N(0, 1). \quad (4.10)$$

Example 4.9 (Testing a difference) For simplicity, suppose we have only two coefficients and want to test the difference. Then, $R = [1 \ -1]$. Suppose (again for simplicity) that $\text{Var}(\hat{\beta}) = [\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}]$, where $\rho = \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. Clearly, $R \text{Var}(\hat{\beta}) R'$ equals $\text{Var}(\hat{\beta}_1) +$

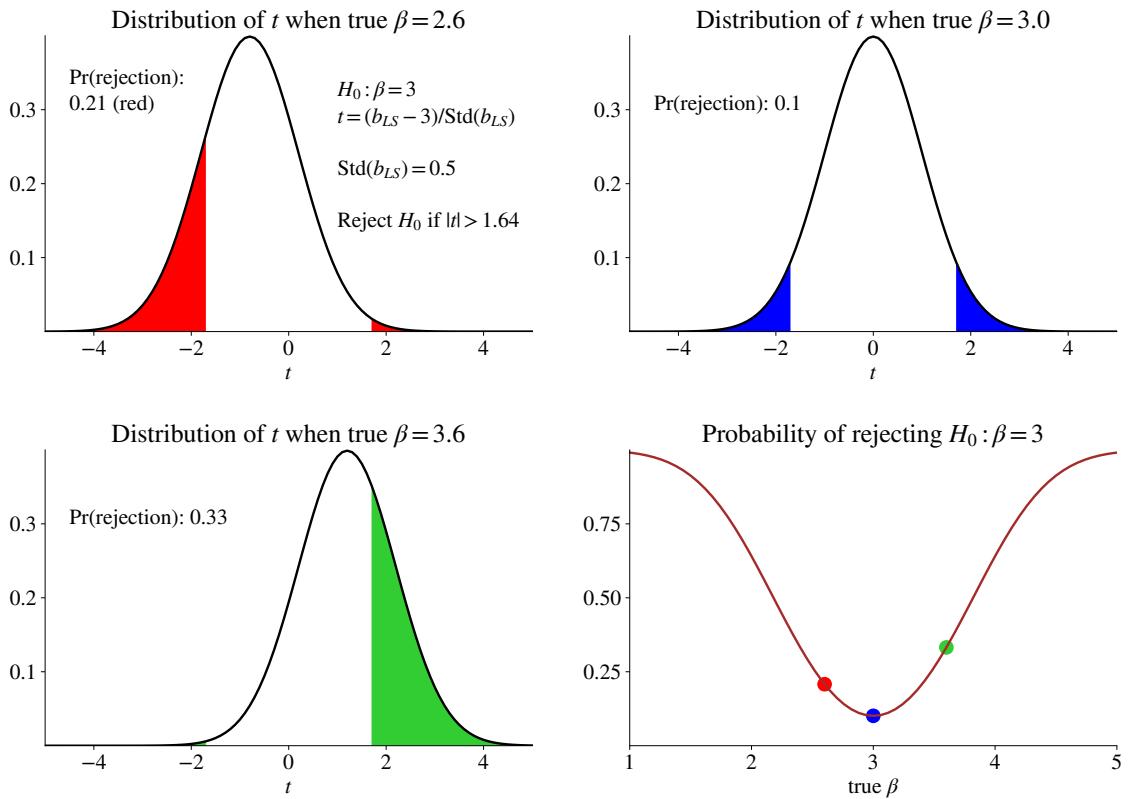


Figure 4.5: Power of t-test, assuming different true parameter values

$\text{Var}(\hat{\beta}_2) - 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ which here is $2(1 - \rho)$. A higher covariance means that $\hat{\beta}_1$ and $\hat{\beta}_2$ tend to move in the same direction so the difference has a small uncertainty. It is then easy to test the difference. The opposite is true when testing a sum with $R = [1 \ 1]$.

4.5 Joint Test of Several Coefficients

A joint test of several coefficients is different from testing the coefficients one at a time. For instance, suppose your economic hypothesis is that $\beta_1 = 1$ and $\beta_3 = 0$. You could clearly test each coefficient individually (by a t-test), but that may give conflicting results. In addition, it does not use the information in the sample as effectively as possible. Intuitively, a joint test is like exploiting the power of repeated sampling.

A joint test makes use of the following remark.

Remark 4.10 (Chi-square distribution) If v is a zero mean vector with n elements which

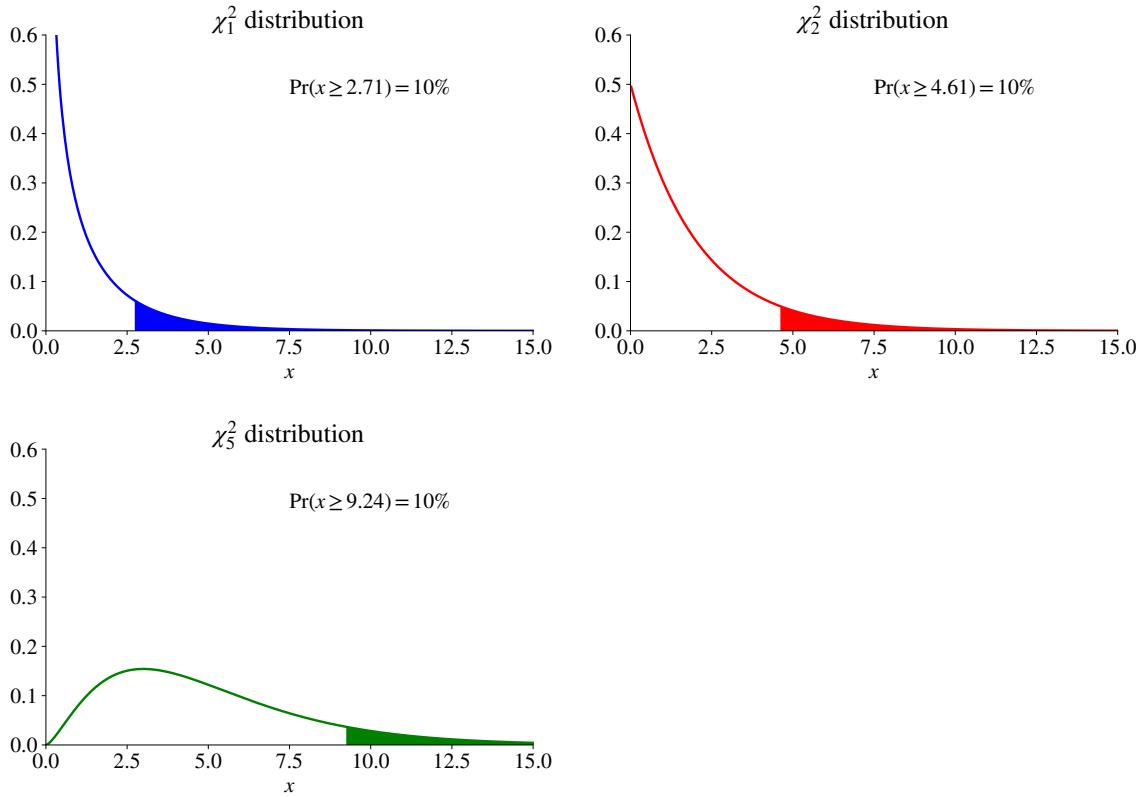


Figure 4.6: Density functions of χ^2 distributions with different degrees of freedom

are jointly normally distributed, $v \sim N(0, \Sigma)$, then

$$v' \Sigma^{-1} v \sim \chi_n^2.$$

As a special case, suppose the vector has just one element. In this case, the quadratic form can be written $v^2 / \text{Var}(v)$, which is the square of a t-statistic. It has a χ_1^2 distribution.

Example 4.11 (Quadratic form with a chi-square distribution) If the 2×1 vector v has the following normal distribution

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \right),$$

then the quadratic form

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}' \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1^2 + v_2^2/2$$

has a χ^2_2 distribution. In a more general example, the variables could be correlated.

The null hypothesis can be written on matrix form as

$$R\beta = q, \quad (4.11)$$

where R is a $J \times k$ matrix, β a k -vector and q is a J -vector.

Example 4.12 ((4.11) with $J = 2$) Suppose the model has three coefficients and the null hypothesis is

$$H_0 : \beta_1 = 1 \text{ and } \beta_3 = 0.$$

This can be written on the form of (4.11) as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Notice that the variance-covariance matrix of these linear combinations is

$$\text{Var}(R\hat{\beta}) = R \text{Var}(\hat{\beta}) R', \quad (4.12)$$

where $\text{Var}(\hat{\beta})$ denotes the covariance matrix of the coefficients. $\text{Var}(R\hat{\beta})$ is now a $J \times J$ matrix. As before, the estimate $\hat{\beta}$ could be from a linear regression, but could equally well be from some other method. Putting together these results we have the test static (a scalar)

$$(R\hat{\beta} - q)'(R \text{Var}(\hat{\beta}) R')^{-1}(R\hat{\beta} - q) \sim \chi^2_J. \quad (4.13)$$

This test statistic is compared to the critical values of a χ^2_J distribution. This is called a *Wald test*. (Alternatively, this test can be put in the form of an F statistic, which is a small sample refinement discussed below.)

Remark 4.13 ($J = 1$) When R has just one row, then the test statistic from (4.13) equals t^2 where t is from (4.10) and it follows a χ^2_1 distribution.

A particularly important case is the test of the joint hypothesis that all $k - 1$ slope coefficients in the regression (that is, excluding the intercept) are zero. The test statistic for this hypothesis is (assuming your regression also contains an intercept)

$$TR^2/(1 - R^2) \sim \chi^2_{k-1}, \quad (4.14)$$

where R^2 denotes the (scalar) goodness-of-fit and should not be confused with the R matrix used in formulating the null hypothesis (4.11). (A proof of (4.14) is given below.)

Empirical Example 4.14 (*Test of all slopes*) See Tables 4.1 and 4.2.

Example 4.15 (*Joint test*) Suppose $H_0: \beta_1 = 0$ and $\beta_3 = 0$; $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (2, 777, 3)$ and

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \text{Var}(\hat{\beta}) = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 33 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ so}$$

$$R \text{Var}(\hat{\beta}) R' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 33 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}.$$

(We assume $\text{Var}(\hat{\beta})$ is diagonal just because it makes it easier to invert.) Then, (4.13) is

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 777 \\ 3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)' \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 777 \\ 3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)$$

$$\begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 10,$$

which is higher than the 10% critical value of the χ^2_2 distribution, which is 4.61.

Remark 4.16 (*An alternative form of (4.13)) Define the standardised values $z = (R\hat{\beta} - q) / \text{Std}(R\hat{\beta} - q)$. Then, (4.13) can be also be written $z' \text{Corr}(z)^{-1} z$.

Remark 4.17 (*Power and size of a joint test**) Suppose $v \sim N(v_0, \Sigma)$, where v_0 might be non-zero. Then $v' \Sigma^{-1} v \sim \chi^2_n(\lambda)$ with $\lambda = v_0' \Sigma^{-1} v_0$ and where $\chi^2_n(\lambda)$ is a non-central chi-square distribution with non-centrality parameter λ . This distribution coincides with the traditional chi-square when $\lambda = 0$. Instead, if $R\beta - q = q_0$ (instead of zero), then the test static in (4.13) would have a $\chi^2_J(\lambda)$ distribution with $\lambda = q_0' [R \text{Var}(\hat{\beta}) R']^{-1} q_0$. We can then calculate the power of the test in (4.13) for different values of q_0 .

Proof (of (4.14)) Recall that the goodness-of-fit is $R^2 = \hat{v}(\hat{y}_t) / \hat{v}(y_t) = 1 - \hat{v}(\hat{u}_t) / \hat{v}(y_t)$, where $\hat{y}_t = x_t' \hat{\beta}$ and \hat{u}_t are the fitted value and residual respectively and $\hat{v}()$ is short hand notation for a sample variance. We therefore get $TR^2/(1 - R^2) =$

$T \hat{v}(\hat{y}_t) / \hat{v}(\hat{u}_t)$. To simplify the algebra, assume that both y_t and x_t are demeaned and that no intercept is used. We get the same results, but after more work, if we relax this assumption. In this case we can rewrite as $TR^2/(1 - R^2) = T\hat{\beta}'\hat{v}(x_t)\hat{\beta}/\hat{\sigma}^2$, where $\hat{\sigma}^2 = \hat{v}(\hat{u}_t)$. If the iid assumptions are correct, then the variance-covariance matrix of $\hat{\beta}$ is estimated as $\text{Var}(\hat{\beta}) = [T\hat{v}(x_t)]^{-1}\hat{\sigma}^2$, so we get

$$\begin{aligned} TR^2/(1 - R^2) &= \hat{\beta}'T\hat{v}(x_t)/\sigma^2\hat{\beta} \\ &= \hat{\beta}'\text{Var}(\hat{\beta})^{-1}\hat{\beta}. \end{aligned}$$

This has the same form as (4.13) with the matrix $R = I$ and the vector $q = \mathbf{0}$ and where J equals to the number of slope coefficients. \square

Remark 4.18 (*F-test**) *The joint test can also be cast in terms of the F distribution. Divide (4.13) by J and replace $\text{Var}(\hat{\beta})$ by the estimated covariance matrix, for instance, $\hat{V} = \hat{\sigma}^2(\sum_{t=1}^T x_t x_t')^{-1}$ where we use \hat{V} to denote the estimate, but where we (as in reality) have to estimate the variance of the residuals by the sample variance of the fitted residuals, $\hat{\sigma}^2$. This gives*

$$(R\hat{\beta} - q)'(R\hat{V}R')^{-1}(R\hat{\beta} - q)/J \sim F_{J,T-k}.$$

For instance, the test of the joint hypothesis that all $k-1$ slope coefficients in the regression (that is, excluding the intercept) are zero can be written (assuming your regression also contains an intercept) $[R^2/(k-1)]/[(1-R^2)/(T-k)] \sim F_{k-1,T-k}$. The F-test puts fairly strict requirements of the regression model (normal distribution is small samples and independence of residuals and regressors), and is therefore not used much in these notes.

4.6 Confidence Bands for a Forecast*

Suppose we have estimated the linear model $y_t = x_t'\beta + u_t$. For a given (known) vector x_s , our *forecast* of y_s is

$$E(y_s|x_s) = x_s'\hat{\beta}. \quad (4.15)$$

For a given x_s , this is just a linear combination of the estimated coefficients, so the result in (4.12) holds, but with x_s' replacing R

$$\text{Var}[E(y_s|x_s)] = x_s' \text{Var}(\hat{\beta}) x_s. \quad (4.16)$$

Instead, if we want the uncertainty about the *forecast error*

$$y_s - \mathbb{E}(y_s|x_s) = x'_s(\beta - \hat{\beta}) + u_s, \quad (4.17)$$

then we have to add the uncertainty of u_s

$$\text{Var}[y_s - \mathbb{E}(y_s|x_s)] = x'_s \text{Var}(\hat{\beta})x_s + \sigma^2. \quad (4.18)$$

(To show this last result, notice that x_s is not random and that u_s is not correlated with $\hat{\beta}$, at least not if the latter is estimated from a sample that does not contain observation s .)

4.7 Testing Nonlinear Hypotheses (Delta Method)

Suppose we want the asymptotic distribution of a function of some parameters β

$$\gamma = f(\beta), \quad (4.19)$$

where $f(.)$ has continuous first derivatives, where γ could be a scalar or a vector. We have estimates $(\hat{\beta})$ which satisfy

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, W). \quad (4.20)$$

As before, the estimate $\hat{\beta}$ could be from a linear regression, but could equally well be from some other method.

Example 4.19 (*Testing a Sharpe ratio I*) Stack the mean and variance in the vector $\beta = [\mu, \sigma^2]$. The Sharpe ratio is calculated as a function of β , $f(\beta) = \mu / (\sigma^2)^{1/2}$.

We can construct a test by noticing that a first-order Taylor approximation is

$$f(\hat{\beta}) - f(\beta) \approx D(\hat{\beta} - \beta), \quad (4.21)$$

where D is the matrix of partial derivatives (the Jacobian)

$$D = \begin{bmatrix} \frac{\partial f_1(\beta)}{\partial \beta_1} & \dots & \frac{\partial f_1(\beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q(\beta)}{\partial \beta_1} & \dots & \frac{\partial f_q(\beta)}{\partial \beta_k} \end{bmatrix}, \quad (4.22)$$

where $\partial f_i(\beta) / \partial \beta_j$ is the derivative of output i from the function $f()$ with respect to parameter j . The derivatives can sometimes be found analytically, otherwise numerical

differentiation can be used. In principle, the derivatives should be calculated at the true parameter values and using population moments of the data, but in practice we use the point estimates $\hat{\beta}$ and using the existing sample.

It follows that, as a first order approximation, the variance-covariance matrix of $f(\hat{\beta}) - f(\beta)$ is DWD' . In fact, asymptotically all higher-order terms can be disregarded. This means that, under that null hypothesis (that the true value is γ)

$$\sqrt{T}(f(\hat{\beta}) - \gamma) \xrightarrow{d} N(0, DWD'). \quad (4.23)$$

(An outline of a proof is given at the end of the section.) Now, a test can be done by using (4.23), similarly to the linear case. See below for some more details on the proof. This result could also be expressed in terms of the variance-covariance matrix of $\hat{\beta}$ (rather than of $\sqrt{T}\hat{\beta}$) as

$$\text{if } \hat{\beta} \xrightarrow{a} N(\beta, V), \text{ then } f(\hat{\beta}) \xrightarrow{a} N(\gamma, DWD'), \quad (4.24)$$

where \xrightarrow{a} means “is asymptotically distributed as.”

Example 4.20 (*Testing a Sharpe ratio II*) Continuing Example 4.19, notice that the derivatives are

$$D = \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix}.$$

If $\hat{\beta}$ is normally distributed, then (4.24) is straightforward to apply. Also, with the Sharpe ratios for two assets (i, j) as outputs of $f(\beta)$ and $\beta = (\mu_i, \sigma_i^2, \mu_j, \sigma_j^2)$ we get

$$D = \begin{bmatrix} \frac{1}{\sigma_i} & \frac{-\mu_i}{2\sigma_i^3} & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma_j} & \frac{-\mu_j}{2\sigma_j^3} \end{bmatrix}.$$

This can be used to test the difference of two Sharpe ratios.

Empirical Example 4.21 (*Test of Sharpe ratio*) See Table 4.4 for a test of the Sharpe ratio of the US equity market.

Example 4.22 (*Linear function*) When $f(\beta) = R\beta$, then the Jacobian is $D = R$, just like in (4.12).

Example 4.23 (*Testing a correlation of x_t and y_t*) Suppose you have estimated the variances of (x_t, y_t) and also their covariance. Stack the parameters in the vector

Equity returns	
SR	0.13
SR (ann.)	0.46
t-stat	3.17

Table 4.4: SR (with t-stat) for the US equity market. Monthly US data 1970:01-2024:12.
The annualised SR is $\text{SR}\sqrt{12}$

$\beta = [\sigma_{xx}, \sigma_{yy}, \sigma_{xy}]$. The correlation and the Jacobian is then

$$\rho(x, y) = f(\beta) = \frac{\sigma_{xy}}{\sigma_{xx}^{1/2} \sigma_{yy}^{1/2}}, \text{ so } D = \begin{bmatrix} -\frac{1}{2} \frac{\sigma_{xy}}{\sigma_{xx}^{3/2} \sigma_{yy}^{1/2}} & -\frac{1}{2} \frac{\sigma_{xy}}{\sigma_{xx}^{1/2} \sigma_{yy}^{3/2}} & \frac{1}{\sigma_{xx}^{1/2} \sigma_{yy}^{1/2}} \end{bmatrix}.$$

Delta Method Example: Confidence Bands around a Mean-Variance Frontier

A point on the mean-variance frontier at a given expected return is a non-linear function of the means and the second moment matrix of the returns of the investable assets. It is therefore straightforward to apply the delta method to calculate a confidence band around the estimate.

Empirical Example 4.24 (*MVF from industry portfolios*) Figure 4.7 (lower panel) shows GMM results. The uncertainty is lowest for the minimum variance portfolio. (This is related to the result that in a normal distribution, the uncertainty about an estimated variance is increasing in the true variance, $\text{Var}(\sqrt{T}\hat{\sigma}^2) = 2\sigma^4$.)

Delta Method Example: Testing the $1/N$ vs the Tangency Portfolio

It has been argued that (naive) $1/N$ diversification gives a portfolio performance which is no worse than an “optimal” portfolio (see, for instance, DeMiguel, Garlappi, and Uppal (2009)). One way to test this is to compare the Sharpe ratios of the tangency and equally weighted portfolios. Both are functions of the first and second moments of the returns of the investable assets, so a delta method approach can be applied.

Empirical Example 4.25 ($1/N$ vs. the tangency portfolio) See Figure 4.7.

Delta Method Example: Testing the Optimal Portfolio Weight

A mean-variance investor combines the risk-free asset with a mix (the tangency portfolio) of risky assets. The optimal portfolio weight on the latter is $w = E R_m^e / [k \text{Var}(R_m^e)]$, where R_m^e denotes the excess return of the tangency portfolio. If we have estimates and their covariance matrix of $E R_m^e$ and $\text{Var}(R_m^e)$, then it is straightforward to construct a confidence band around w .

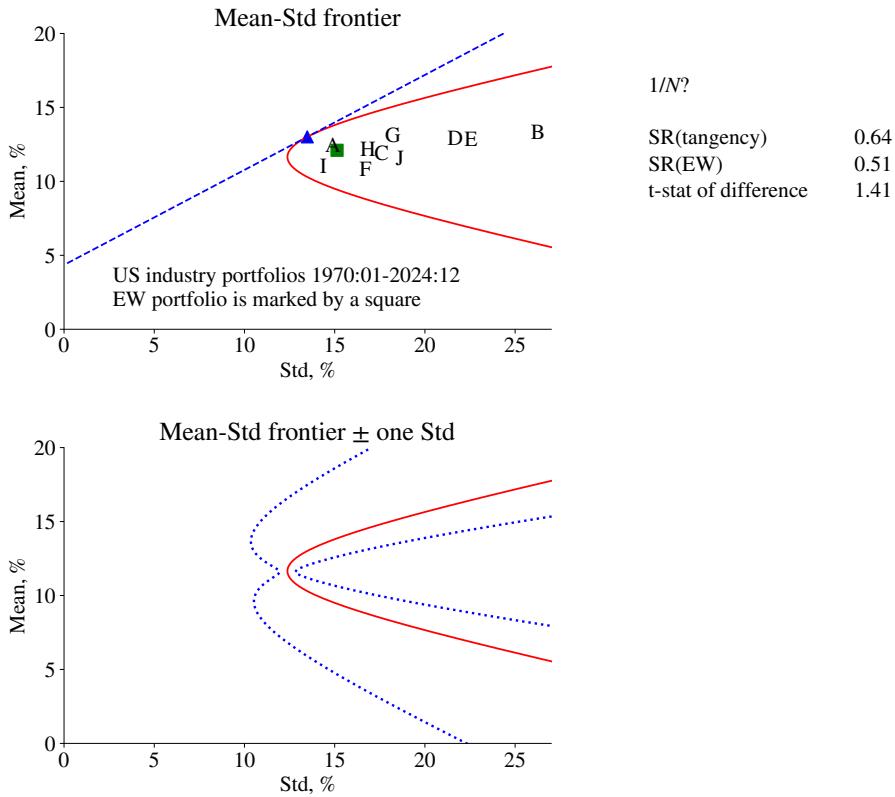


Figure 4.7: Mean-Variance frontier of US industry portfolios. Monthly returns are used in the calculations, but $100\sqrt{12}\text{Variance}$ is plotted against $100 * 12*\text{mean}$.

Empirical Example 4.26 (*Choosing between the risk-free and a single risky asset (S&P 500)*) See [Figure 4.8](#).

Proof (Sketch of a proof of (4.23), requiring some asymptotics*) By the mean value theorem we have

$$f(\hat{\beta}) = f(\beta) + \frac{\partial f(\beta^*)}{\partial \beta'} (\hat{\beta} - \beta),$$

where the derivatives are evaluated at β^* which is (weakly) between $\hat{\beta}$ and β . Premultiply by \sqrt{T} and rearrange as

$$\sqrt{T}[f(\hat{\beta}) - f(\beta)] = \frac{\partial f(\beta^*)}{\partial \beta'} \sqrt{T}(\hat{\beta} - \beta_0).$$

If $\hat{\beta}$ is consistent ($\text{plim } \hat{\beta} = \beta$) and $\partial f(\beta^*) / \partial \beta'$ is continuous, then, by Slutsky's theorem, the probability limit of the derivatives is $\partial g(\beta) / \partial \beta'$ (that is, evaluated at the true β —and thus a constant). If $\sqrt{T}(\hat{\beta} - \beta_0)$ is asymptotically normally distributed, then, by the continuous mapping theorem, this carries over to the left hand side. \square

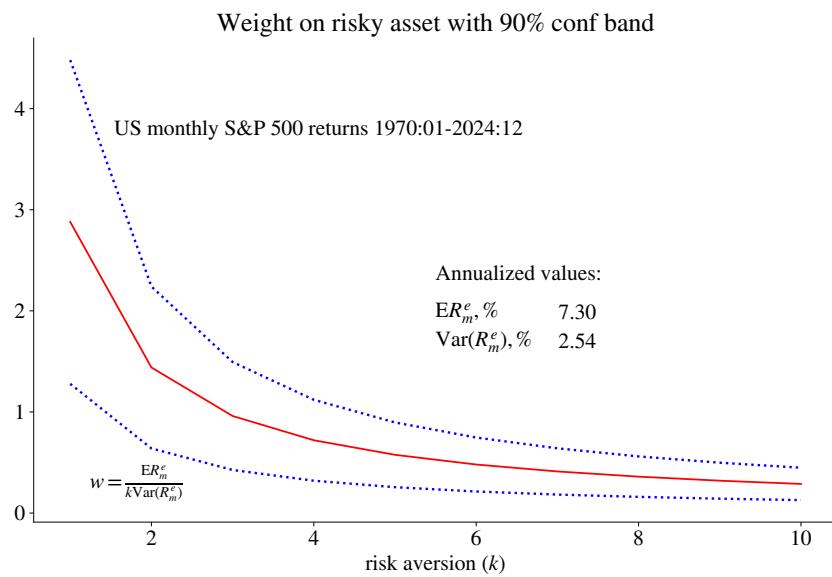


Figure 4.8: Portfolio choice for different risk aversions, with confidence band

Further Reading

Verbeek (2017) 2, Greene (2018) 4-5 and parts of 9 and 20 and Hansen (2022a) 9 add many results.

Chapter 5

Least Squares: Non-iid Residuals

5.1 Heteroskedasticity

Suppose we have a regression model

$$y_t = x'_t b + u_t, \text{ where } E u_t = 0 \text{ and } \text{Cov}(x_{it}, u_t) = 0. \quad (5.1)$$

In the standard case we assume that u_t is iid (independently and identically distributed), which rules out variation in the volatility of the residual (heteroskedasticity).

If the residuals actually are heteroskedastic, least squares (LS) is nevertheless a useful estimator: its consistency is not affected by the heteroskedasticity. However, the traditional (assuming iid errors) expression for the standard errors of the coefficients is often not correct. This is illustrated in Tables 5.1 – 5.2, which show results from simulations. Notice, however, that the simulations suggest that heteroskedasticity that is unrelated to the regressors does not cause any problems with the standard errors. The constant is a special case of this (discussed more in detail below).

In contrast, when there is a positive relation between x_t^2 and σ_t^2 , then the traditional standard errors underestimate the true uncertainty. The intuition is that much of the precision (low variance of the estimates) of OLS comes from data points with extreme values of the regressors: think of a scatter plot and notice that the slope depends a lot on fitting the data points with very low and very high values of the regressor. This nice property is destroyed if the data points with extreme values of the regressor also have lots of noise (high variance of the residual). See Figure 5.1 for an illustration.

To test for heteroskedasticity, we can use *White's test of heteroskedasticity*. The test assumes that the fitted residuals come from consistent estimates and allows for stochastic regressors. The null hypothesis is homoskedasticity, and the alternative hypothesis is

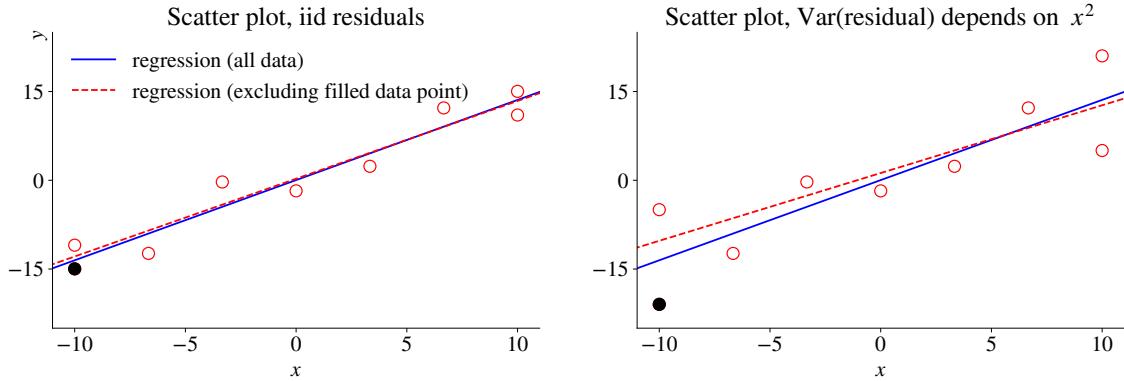


Figure 5.1: Effect of heteroskedasticity on uncertainty about regression line

the kind of heteroskedasticity that is related to the levels, squares, and cross products of the regressors. The reason for this specification is that if the squared residuals are uncorrelated with the regressors, then the covariance matrix can be estimated with the traditional approach (see the simulations mentioned above).

To implement White's test, let w_t be a vector of the squares and cross products of the regressors, but ensure that a constant is among them. The test is then to run a regression of squared fitted residuals on w_t

$$\hat{u}_t^2 = w_t' \gamma + v_t, \quad (5.2)$$

and to test if all the slope coefficients (not the intercept) in γ are zero. This can be done by using the fact that $TR^2/(1 - R^2) \sim \chi_p^2$, where p is the number of slope coefficients in γ . (Some authors prefer to use TR^2 instead, but the difference is likely to be small.) There are actually several alternative versions of this test: (a) using only linear terms, (also called the Breusch-Pagan test); (b) using only linear and quadratic terms, not the cross products; (c) using only a subset of the regressors.

Example 5.1 (White's test) If the regressors include $(1, x_{1t}, x_{2t})$ then w_t in (5.2) is the vector $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$.

Remark 5.2 (*Duplicate variables in w_t .) If x_t contains a dummy variable, then its square will be the same. You can still use the same test statistic (provided the estimation algorithm does not crash), but p should be the number of linearly independent variables in w_t minus 1.)

	$\gamma = 0$		$\gamma = 1$	
	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$
Simulated	7.1	13.4	13.5	19.4
OLS formula	7.0	13.3	13.3	19.2
White's	7.0	13.3	13.3	19.2
Bootstrap	7.1	13.3	13.4	19.2
Bootstrap 2	7.0	13.3	13.3	19.1
Jackknife	7.1	13.4	13.4	19.4
FGLS	7.2	11.6	13.7	18.2

Table 5.1: Standard error of OLS intercept (%) under heteroskedasticity (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1 + \gamma|z_t| + \alpha|x_t|)^2$, where z_t is iid $N(0, 1)$ and independent of x_t . Sample length: 200. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

There are two ways to handle heteroskedasticity in the residuals. First, we could use some other estimation method than LS that incorporates the structure of the heteroskedasticity. For instance, combining the regression model (5.1) with an ARCH structure of the residuals, and estimating with maximum likelihood (MLE). As a by-product we get correct standard errors, provided the assumed distribution (in the likelihood function) is appropriate. Second, we could stick to OLS, but use another expression for the variance-covariance matrix of the coefficients: a heteroskedasticity consistent method, among which “*White's covariance matrix*” is the most common. (There is also a third possible solution: using GLS, but that is often a non-robust approach.)

To understand the construction of White's covariance matrix, recall that the variance of $\hat{\beta}$ is found from

$$\hat{\beta} = \beta + S_{xx}^{-1} (x_1 u_1 + x_2 u_2 + \dots x_T u_T), \quad (5.3)$$

where $S_{xx} = \sum_{t=1}^T x_t x_t'$. If we assume that the residuals (or more precisely, $x_t u_t$ and $x_s u_s$) are uncorrelated with each other and that regressors are fixed, then the variance-covariance matrix of $\hat{\beta}$ is

$$\begin{aligned} \text{Var}(\hat{\beta}) &= S_{xx}^{-1} (x_1 x_1' \sigma_1^2 + x_2 x_2' \sigma_2^2 + \dots x_T x_T' \sigma_T^2) S_{xx}^{-1} \\ &= S_{xx}^{-1} S S_{xx}^{-1}, \text{ where } S = \sum_{t=1}^T x_t x_t' \sigma_t^2. \end{aligned} \quad (5.4)$$

(Notice that S_{xx} and S denote very different things.) A similar expression can be derived

	$\gamma = 0$		$\gamma = 1$	
	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$
Simulated	7.1	19.0	13.5	24.8
OLS formula	7.1	13.3	13.4	19.2
White's	7.0	18.5	13.3	24.3
Bootstrap	7.1	18.5	13.4	24.3
Bootstrap 2	7.0	18.5	13.3	24.3
Jackknife	7.1	18.9	13.5	24.8
FGLS	7.5	17.3	14.0	24.1

Table 5.2: Standard error of OLS slope (%) under heteroskedasticity (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1 + \gamma|z_t| + \alpha|x_t|)^2$, where z_t is iid $N(0, 1)$ and independent of x_t . Sample length: 200. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

for stochastic regressors, and it would be valid in large samples.

Expression (5.4) cannot be simplified further since σ_t is not constant, and it is related to $x_t x_t'$. The idea of White's method is to estimate S by using \hat{u}_t^2 instead of σ_t^2

$$\hat{S} = \sum_{t=1}^T x_t x_t' \hat{u}_t^2. \quad (5.5)$$

Remark 5.3 (*An interpretation of (5.5)*) Recall that $x_t \hat{u}_t$ is a k -vector of zero mean variables, so \hat{S} is $T \widehat{\text{Var}}(x_t \hat{u}_t)$, that is, T times the (sample) variance-covariance matrix of the vector $x_t \hat{u}_t$. For instance, with a constant and one more regressor so $x_t = [1, z_t]$, we get

$$\widehat{\text{Var}}(x_t u_t) = \widehat{\text{Var}}\left(\begin{bmatrix} u_t \\ z_t u_t \end{bmatrix}\right) = \begin{bmatrix} \widehat{\text{Var}}(u_t) & \widehat{\text{Cov}}(u_t, z_t u_t) \\ \widehat{\text{Cov}}(z_t u_t, u_t) & \widehat{\text{Var}}(z_t u_t) \end{bmatrix}.$$

White's covariance matrix should be applied when White's test (5.2) indicates problems, otherwise perhaps not. While White's covariance estimator provides safety against heteroskedasticity, it also comes at a cost: estimating the S matrix as in (5.5) risks introducing more noise

Remark 5.4 (*Matrix form of (5.5)) With x_t' in row t of the $T \times k$ matrix X , $\hat{S} = X' \text{diag}(\hat{u}_t^2) X$, where $\text{diag}(\hat{u}_t^2)$ is a $T \times T$ matrix with \hat{u}_t^2 along the principal diagonal and zeros elsewhere. (However, this is not an efficient way of calculating \hat{S} .)

	HiTec	Utils
constant	-0.05 (-0.41)	0.23 (1.70)
market return	1.24 (39.02)	0.52 (14.57)
R^2	0.76	0.33
Autocorr	0.32	0.93
White	0.05	0.00
All slopes	0.00	0.00
obs	660	660

Table 5.3: CAPM regressions, monthly returns, %, US data 1970:01-2024:12. Numbers in parentheses are t-stats. Autocorr is the p-value for no autocorrelation; White is the p-value for homoskedasticity; All slopes is the p-value for all slope coefficients being zero.

	HiTec	Utils
R_m^e	0.07 (0.40)	-0.22 (-1.45)
$R_m^e R_m^e$	0.05 (2.48)	0.08 (4.26)
White	0.05	0.00

Table 5.4: Regression of \hat{u}_{it}^2 on regressors and squares from CAPM regressions ($R^e = \alpha + \beta R_m^e + u$), monthly returns, US data 1970:01-2024:12. t-stats are in parentheses. White is the p-value for the null hypothesis that all slopes are 0 (White's test).

Empirical Example 5.5 (*Test and effect of heteroskedasticity*) See Tables 5.3 –5.5. The evidence shows significant heteroskedasticity. This affects the t-stats of the slope coefficient to some extent, but not the intercept. In contrast, Figure 5.2 shows an example (from daily data) where the effect is considerable.

Remark 5.6 (*Standard OLS vs White's variance**) For simplicity, consider the case of only one regressor, and use asymptotic expressions. Recall that $\text{Cov}(x_t^2, \sigma_t^2) = \mathbb{E} x_t^2 \sigma_t^2 - \mathbb{E} x_t^2 \mathbb{E} \sigma_t^2$. It follows that if x_t^2 is not correlated with σ_t^2 , then the S term in (5.4), divided by T, can be thought of as $\mathbb{E} \sigma_t^2 \mathbb{E} x_t^2$, that is, it involves the average variance, typically estimated as $\sum_{t=1}^T u_t^2 / T$. This is the same as for the traditional OLS expression. In contrast, if σ_t^2 and x_t^2 are positively correlated, then $\mathbb{E} x_t^2 \sigma_t^2 > \mathbb{E} x_t^2 \mathbb{E} \sigma_t^2$, which means that the standard OLS expression underestimates the uncertainty.

	HiTec	Utils
intercept, iid	-0.41	1.74
intercept, White	-0.41	1.70
intercept, Newey-West 1 lag	-0.42	1.70
slope, iid	45.34	18.12
slope, White	39.02	14.57
slope, Newey-West 1 lag	35.56	14.68

Table 5.5: t-stats from different methods (iid, White's, Newey-West), for CAPM regressions ($R^e = \alpha + \beta R_m^e + u$), monthly returns, US data 1970:01-2024:12.

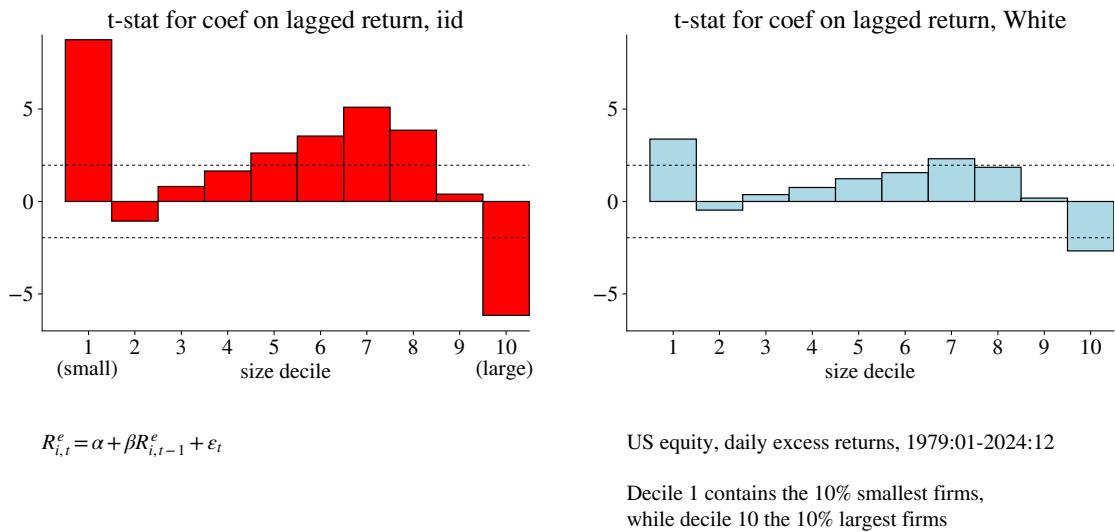


Figure 5.2: t-stat for AR(1) coefficient on daily returns

Remark 5.7 (Jackknife covariance matrix) The jackknife covariance matrix is: (a) reestimate the model T times, each time with one observation excluded (first exclude observation 1, then put it back and instead exclude observation 2,...), to get a $T \times k$ matrix of estimates; (b) calculate the covariance matrix of the k series; (c) multiply the result by $T - 1$. This typically leads to results that are similar to those from (5.4)–(5.5).

Remark 5.8 (GLS*) With heteroskedasticity and/or autocorrelation, OLS is still consistent and we can adjust the covariance matrix of the coefficients. As an alternative, Generalized Least Squares (GLS) transforms the regression equation so $y_t^* = x_t^{*\prime} \beta + \varepsilon_t^*$, have iid residuals. OLS can then be applied and the traditional expression of the variance-covariance matrix applies. For instance, with heteroskedasticity, the transformation is

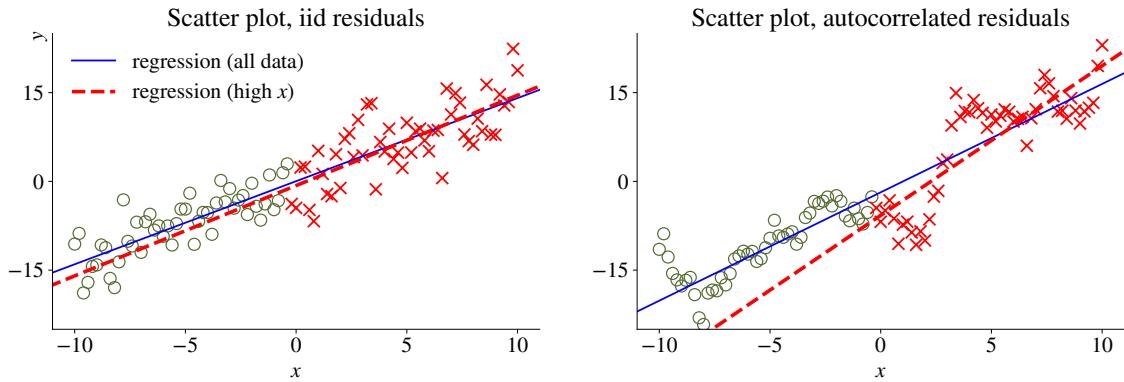


Figure 5.3: Effect of autocorrelation on uncertainty about regression line

$y_t/\sigma_t = (x_t/\sigma_t)'\beta + \varepsilon_t/\sigma_t$. (Yes, also the constant is divided by σ_t .) Notice that ε_t/σ_t has a constant variance (equal to one). In practice we don't know σ_t , so we first estimate it. The method is then called "feasible" GLS, FGLS. A commonly applied approach is the following (a) let $\hat{\varepsilon}_t$ be the residual from the OLS regression; (b) regress $\ln(\hat{\varepsilon}_t^2)$ on the regressors and all the squares (and cross-products of them); (c) let z_t be the fitted values from the regression in (b) and set $\sigma_t = \sqrt{\exp(z_t)}$. This may improve the efficiency, but can be unstable if we model estimates the heteroskedasticity wrongly.

5.2 Autocorrelation

Autocorrelation of the residuals ($\text{Cov}(u_t u_{t-s}) \neq 0$) is also a violation of the iid assumptions underlying the standard expressions for the variance of $\hat{\beta}$. Again, LS is typically still consistent (exception: when the lagged dependent variable is a regressor), but the variances are again wrong.

The typical effect of positively autocorrelated residuals is to increase the uncertainty about the OLS estimates, above what is indicated by the traditional standard errors. The reason is that positively autocorrelated residuals imply long swings around the true regression line for extended periods of time, causing imprecise estimates. See Figure 5.3 for an illustration.

There are several tests of autocorrelation, which all assume that the fitted residuals come from consistent estimates. The null hypothesis is no autocorrelation. First, estimate the autocorrelations of the fitted residuals as

$$\rho_s = \text{Corr}(\hat{u}_t, \hat{u}_{t-s}), s = 1, \dots, L. \quad (5.6)$$

Second, test autocorrelation s by using the fact that $\sqrt{T}\hat{\rho}_s$ has a standard normal distribution (in large samples, under the hypothesis of no autocorrelation at all)

$$\sqrt{T}\hat{\rho}_s \sim N(0, 1). \quad (5.7)$$

To extend (5.7) to several lags, use the Box-Pierce test

$$Q_L = T \sum_{s=1}^L \hat{\rho}_s^2 \sim \chi_L^2. \quad (5.8)$$

Proof (of (5.8)) If (5.7) is true for all lags, then (in a large sample) $T\hat{\rho}_s^2 \sim \chi_1^2$ and $\hat{\rho}_s^2$ and $\hat{\rho}_\tau^2$ are independent, so Q is a sum of independent χ_1^2 variables. \square

An alternative for testing the first autocorrelation coefficient is the Durbin-Watson. The test statistic is (approximately)

$$DW \approx 2 - 2\hat{\rho}_1, \quad (5.9)$$

and the null hypothesis is rejected in favour of positive autocorrelation if $DW < 1.5$ or so, depending on sample size and the number of regressors.

All these tests should also be applied to *each of the elements in $x_t u_t$* (instead of just u_t), since it is actually the autocorrelations of these cross terms that matter most for the uncertainty of the slope coefficients (see the discussion below).

If there *is autocorrelation*, then we can choose to estimate a fully specified model (including how the autocorrelation is generated) by MLE or we can stick to OLS but apply an autocorrelation consistent covariance matrix. (A third approach, GLS, is discussed in a remark below.)

Notice that for OLS, the variance-covariance matrix of $\hat{\beta}$ is (assuming fixed regressors)

$$\text{Var}(\hat{\beta}) = S_{xx}^{-1} S S_{xx}^{-1}, \text{ where } S = \text{Var}(x_1 u_1 + x_2 u_2 + \dots + x_T u_T). \quad (5.10)$$

The S matrix ($k \times k$) in the middle needs to account for correlation across time periods. A similar expression can be derived for stochastic regressors, which would be valid in large samples.

It is clear from (5.10) that what really counts is not so much the autocorrelation in u_t in itself, but the autocorrelation of $x_t u_t$. If this is positive, then the standard expression underestimates the true variance of the estimated coefficients (and vice versa). For instance, the autocorrelation of $x_t u_t$ is likely to be positive when both the residual and the regressor are positively autocorrelated. (Notice that a constant is extremely positively

autocorrelated, so autocorrelation of the residual along is enough to cause problems with the intercept.) In contrast, when the regressor has no autocorrelation, then the product does not either. This is illustrated in Tables 5.6 – 5.7.

	$\rho = 0.0$	$\rho = 0.75$
Simulated	5.9	23.1
OLS formula	5.8	8.6
Newey-West	5.7	18.5
VARHAC	5.7	22.2
Bootstrapped	5.5	19.6
FGLS	5.9	23.2

Table 5.6: Standard error of OLS intercept (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t$, ξ_t is iid $N(0)$. NW uses 10 lags. VARHAC uses 10 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300.

	$\kappa = 0.0$		$\kappa = 0.75$	
	$\rho = 0.0$	$\rho = 0.75$	$\rho = 0.0$	$\rho = 0.75$
Simulated	5.8	8.8	3.9	10.9
OLS formula	5.8	8.6	3.9	5.8
Newey-West	5.6	8.3	3.7	9.4
VARHAC	5.6	8.4	3.7	10.3
Bootstrapped	5.8	8.5	3.8	10.1
FGLS	5.8	4.7	3.9	5.9

Table 5.7: Standard error of OLS slope (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t$, ξ_t is iid $N(0)$. $x_t = \kappa x_{t-1} + \eta_t$, η_t is iid $N(0)$. NW uses 10 lags. VARHAC uses 10 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300.

The next example provides some hints of how to construct an autocorrelation-robust estimator of the S matrix in (5.10).

Example 5.9 (S with $T = 3$) For simplicity, let $T = 3$, assume that covariances beyond the first lag are zero ($\sigma_{13} = \sigma_{31} = 0$) to get

$$S = x_1 x'_1 \sigma_1^2 + x_2 x'_2 \sigma_2^2 + x_3 x'_3 \sigma_3^2$$

$$(x_2 x'_1 + x_1 x'_2) \sigma_{1,2} + (x_3 x'_2 + x_2 x'_3) \sigma_{2,3}$$

where each term is a $k \times k$ matrix and $\sigma_{t,t-s} = \text{Cov}(u_t, u_{t-s})$. Estimate $\sigma_{t,t-s}$ by $\hat{u}_t \hat{u}_{t-s}$.

Example 5.9 suggests that, under the assumption that all covariances beyond the first lag are zero, S could be estimated by

$$\hat{S} = \Gamma_0 + (\Gamma_1 + \Gamma'_1), \text{ where } \Gamma_s = \sum_{t=s+1}^T x_t x'_{t-s} \hat{u}_t \hat{u}_{t-s} \quad (5.11)$$

The Newey-West approach extends this to m (instead of 1) lags and also introduces weights on Γ_s which are declining in the lag s . This ensures that the \hat{S} matrix remains invertible (to show this is somewhat involved). Their approach is thus

$$\hat{S} = \Gamma_0 + \sum_{s=1}^m [1 - s/(m+1)](\Gamma_s + \Gamma'_s), \text{ where} \quad (5.12)$$

$$\Gamma_s = \sum_{t=s+1}^T x_t x'_{t-s} \hat{u}_t \hat{u}_{t-s}. \quad (5.13)$$

Remark 5.10 (An interpretation of (5.13)) Notice that $x_t \hat{u}_t$ is a k -vector of zero mean variables. Therefore, $\Gamma_0 = T \widehat{\text{Var}}(x_t \hat{u}_t)$, that is, T times the (sample) variance-covariance matrix of the vector $x_t \hat{u}_t$. Similarly, $\Gamma_s = T \widehat{\text{Cov}}(x_t \hat{u}_t, x_{t-s} \hat{u}_{t-s})$.

Example 5.11 (\hat{S} with $m = 2$) The calculations in (5.12)–(5.13) are like

$$\hat{S} = \Gamma_0 + \frac{2}{3}(\Gamma_1 + \Gamma'_1) + \frac{1}{3}(\Gamma_2 + \Gamma'_2), \text{ where}$$

$$\Gamma_0 = \sum_{t=1}^T x_t x'_t u_t u_t$$

$$\Gamma_1 = \sum_{t=2}^T x_t x'_{t-1} u_t u_{t-1}$$

$$\Gamma_2 = \sum_{t=3}^T x_t x'_{t-2} u_t u_{t-2}$$

The weights $1 - s/(m+1)$ in (5.12) are close to 1 for small lags (s values), but decline linearly (tent shaped weights) to zero. This suggests that m should be somewhat larger than the last lag with significant autocorrelation. However, a common rule of thumb is to use $m = \text{floor}(0.75T^{1/3})$, where $\text{floor}()$ means rounding down to nearest integer (and alternative rule is $m = \text{floor}(4(T/100)^{2/9})$). Alternatively, study the autocorrelations of $x_t u_t$ and set m equal (or slightly higher) than the last lag. Notice that by setting $m = 0$, the result is the same as in White's approach. Hence, Newey-West estimator handles also heteroskedasticity.

The Newey-West approach should be applied when the tests of the residuals indicate autocorrelation, otherwise probably not. The method involves estimating lots of parameters in the S matrix, and this can in itself introduce noise.

Remark 5.12 (*Scaling of S) Equation (5.10) says that $S = \text{Var}(\sum_{t=1}^T g_t)$, where $g_t = x_t u_t$. Therefore, $S/T = \text{Var}(\sqrt{T}\bar{g})$ (which equals $\text{Var}(g_t)$ in case g_t is iid) and $S/T^2 = \text{Var}(\bar{g})$, where \bar{g} is the time average ($\sum_{t=1}^T g_t/T$).

Remark 5.13 (Hansen-Hodrick) The Hansen-Hodrick approach instead uses flat weights, that is, disregards the $s/(m+1)$ term in 5.12. This is appropriate when we know that the autocorrelation structure is of MA style, for instance, with overlapping returns (daily data on weekly returns, say).

Remark 5.14 (VARHAC*) The VARHAC estimator of the covariance matrix (see Andrews and Monahan (1992)) is to first fit a VAR(p) to $z_t = x_t \hat{u}_t$

$$z_t = A_0 + \sum_{i=1}^p A_i z_{t-i} + \varepsilon_t$$

and then calculate $D = I - \sum_{i=1}^p A_i$. Then, $\hat{S} = D^{-1} \hat{S}^\varepsilon D^{-1}$, where \hat{S}^ε is Newey-West estimate applied to $\hat{\varepsilon}_t$ only.

	HiTec	Utils
u	-0.04 (-0.99)	0.00 (0.09)
$R_m^e u$	0.19 (4.88)	-0.02 (-0.44)

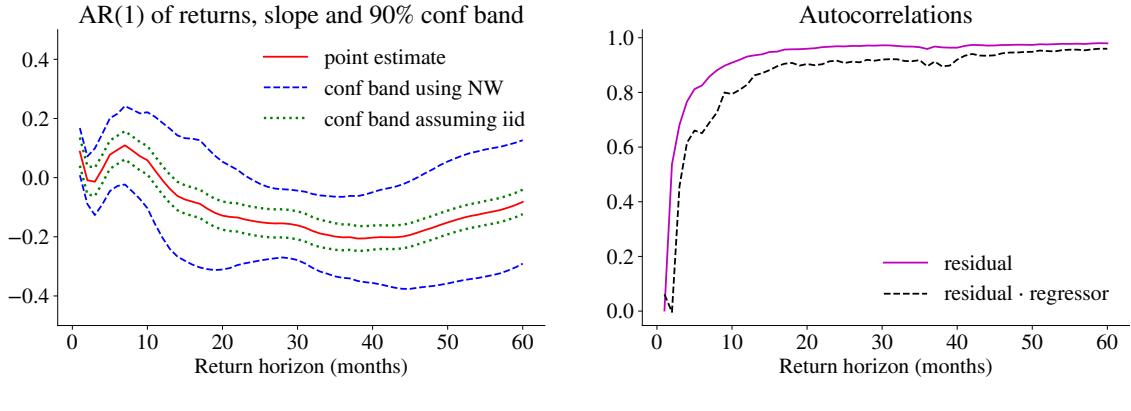
Table 5.8: First autocorrelation of $x_t u_{it}$ from CAPM regressions ($R^e = \alpha + \beta R_m^e + u$), monthly returns, US data 1970:01-2024:12. t-stats are in parentheses.

Empirical Example 5.15 (Test and effect of autocorrelation) See Tables 5.3, 5.5, and 5.8. There is some evidence of autocorrelation (of $x_t u_t$), but the effect on the t-stats is small. In contrast, see Figures 5.4 for a case where the autocorrelation has a large effect.

Remark 5.16 (GLS*) With first-order autocorrelation, ($\varepsilon_t = \rho \varepsilon_{t-1} + v_t$, where v_t is iid), we can implement FGLS by doing a “quasi-difference” of the regression equation

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})' \beta + (\varepsilon_t - \rho \varepsilon_{t-1}).$$

This new residual, $\varepsilon_t - \rho \varepsilon_{t-1}$, is iid. In practice we don't know ρ , so we first estimate it. This is the Cochrane-Orcutt approach, which sometimes iterate over $\hat{\beta}$ and $\hat{\rho}$.



US stocks, monthly excess log returns, 1926:01-2024:12, overlapping data

Figure 5.4: Slope coefficient, LS vs Newey-West standard errors

5.3 Cross-Sectional Correlations (“Clustering”)

In cross-sectional regressions, it is often found that there are clusters of observations with correlated residuals, for instance, for firms within a sector. This has an effect that is similar to the autocorrelation discussed before, except that we now have to be more explicit with indicating which observations that might be correlated and which not. (In a cross-sectional regression, the ordering of observations typically has no particular meaning.)

We again have to estimate the S matrix in (5.10), but this time by defining C different “clusters” within which residuals can be correlated (we have to know which cross-sectional units i that belong to which cluster). In contrast, we assume that there is no correlation across clusters.

Following the usual convention, we use i (instead of t) to indicate an observation and let there be N (not T) observations. We will also introduce the notation G_c for the sum of $x_i u_i$ inside cluster c . We can then write

$$\sum_{i=1}^N x_i u_i = \sum_{c=1}^C G_c, \text{ where } G_c = \sum_{i \in \text{cluster } c} x_i u_i. \quad (5.14)$$

Since there is no correlation across clusters, the S matrix in (5.10) can then be written

$$S = \text{Var}(\sum_{i=1}^N x_i u_i) = \sum_{c=1}^C \text{Var}(G_c). \quad (5.15)$$

See the next example for an illustration.

Example 5.17 (*Cluster method on $N = 4$*) To save space, let $g_i = x_i u_i$ be a scalar and

let $v()$ denote a variance and $\gamma(,)$ a covariance. Assume that individuals 1 and 2 form cluster 1 and that individuals 3 and 4 form cluster 2, and disregard correlations across clusters. This means setting the covariances across clusters to zero

$$\begin{aligned} \text{Var}(\sum_{i=1}^N g_i) &= v(g_1) + v(g_2) + v(g_3) + v(g_4) \\ &\quad + 2\gamma(g_1, g_2) + \underbrace{2\gamma(g_1, g_3)}_0 + \underbrace{2\gamma(g_1, g_4)}_0 + \underbrace{2\gamma(g_2, g_3)}_0 + \underbrace{2\gamma(g_2, g_4)}_0 + 2\gamma(g_3, g_4). \end{aligned}$$

(In case g_i is a vector, $2\gamma(g_1, g_2)$ should be replaced by $\gamma(g_1, g_2) + \gamma(g_2, g_1)$.) Rewrite to get S

$$S = \text{Var}(g_1 + g_2) + \text{Var}(g_3 + g_4).$$

Notice that $\text{Var}(g_1 + g_2)$ can be estimated by $(g_1 + g_2)^2$ since the latter equals $g_1^2 + g_2^2 + 2g_1g_2$ which provides a (rough) estimate of $v(g_1) + v(g_2) + 2\gamma(g_1, g_2)$.

We can estimate S in (5.10) as

$$\hat{S} = \sum_{c=1}^C G_c G'_c. \quad (5.16)$$

In this approach, $G_c G'_c$ can be thought of as a (crude) estimate of $\text{Var}(G_c)$ since g_i (and thus G_c) are zero mean variables. Summing across clusters creates an estimate of S . It is often argued that scaling \hat{S} by $C/(C - 1)$ improves the small properties.

The iid case is when each i is her/his own cluster. In contrast, we cannot allow everyone to be in the same cluster, since this would give $g^c = 0$.

Example 5.18 (*Importance of correlations*) For simplicity, let g_i be a scalar, assume an equal number of members in each cluster (N/C) and that the correlation within a cluster is ρ . It is then straightforward to show that $\text{Var}(\bar{g}) = [1 + \rho(N/C - 1)]\sigma^2/N$. This is the same as with iid data but with an effective sample size of $N^* = N/[1 + \rho(N/C - 1)]$ instead of N . For instance, with $(N, C, \sigma^2) = (500, 10, 100)$

ρ	$\text{Var}(\bar{g})$	Effective sample size
0	0.2	500
0.1	1.2	85
0.25	2.6	38

Remark 5.19 (*Jackknife covariance matrix with clustering*) The jackknife covariance matrix is similar to that in Remark 5.7, except that we now (1) exclude one cluster (not

observation) each time to get a $C \times k$, matrix of estimates; (b) calculate the covariance matrix of the K series; (c) multiply the result by $C - 1$.

Further Reading

See also Verbeek (2017) 4 and Hansen (2022a) 7 for further details.

Chapter 6

A System of OLS Regressions

6.1 A System of Two OLS Regressions

Consider regressions for two different dependent variables (y_{1t} and y_{2t}), such as the returns on two different assets, on the same set of regressors (x_t)

$$y_{1t} = x_t' \beta_1 + u_{1t} \quad (6.1)$$

$$y_{2t} = x_t' \beta_2 + u_{2t}, \quad (6.2)$$

where β_1 is the vector of regression coefficients for y_{1t} and β_2 for y_{2t} . When the regressors are the same, as is the case here, this is often called SURE (Seemingly Unrelated Regression Equations).

It is straightforward to show (see below) that if the *residuals are iid and independent of all regressors*, but we allow for $\text{Cov}(u_{1t}, u_{2t}) \neq 0$, then the variance-covariance matrix for the vector of the two sets of coefficients is

$$\text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) = \begin{bmatrix} \sigma_{11} S_{xx}^{-1} & \sigma_{12} S_{xx}^{-1} \\ \sigma_{21} S_{xx}^{-1} & \sigma_{22} S_{xx}^{-1} \end{bmatrix}, \quad (6.3)$$

where $\sigma_{ij} = \text{Cov}(u_{it}, u_{jt})$ and where $S_{xx} = \sum_{t=1}^T x_t x_t'$.

Remark 6.1 (Background to (6.3)) As discussed in earlier chapters, the variance-covariance matrix (6.3) can be motivated by either (a) assuming that x_t are fixed regressors or (b) as an estimate of the variance-covariance matrix motivated by asymptotic results.

More generally, when the residuals are heteroskedastic or autocorrelated,

$$\text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) = \begin{bmatrix} S_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & S_{xx}^{-1} \end{bmatrix} \Omega \begin{bmatrix} S_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & S_{xx}^{-1} \end{bmatrix}, \text{ where} \quad (6.4)$$

$$\Omega = \text{Var} \left(\sum_{t=1}^T \begin{bmatrix} x_t u_{1t} \\ x_t u_{2t} \end{bmatrix} \right). \quad (6.5)$$

The Ω matrix (which is $2k \times 2k$ if there are k regressors in x_t) could be estimated with the methods of White or Newey-West. This is essentially the same as in the case of just one regression, once $(x_t u_{1t}, x_t u_{2t})$ is stacked into a single vector. In practice, we estimate (6.4) by plugging in $S_{xx} = \sum_{t=1}^T x_t x_t'$ and an estimate of Ω .

Once we have the variance-covariance matrix in (6.4), it is straightforward to test across equations, for instance, that the slope coefficients on a regressor are the same in all regressions or that all intercepts are zero.

Example 6.2 (*Testing the intercepts*) Suppose there are only two regressors and that the first one is the constant. The intercepts (here called α) are then picked out by

$$\hat{\alpha} = R \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \text{ where } R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

and their variance-covariance matrix is

$$\text{Var}(\hat{\alpha}) = R \text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) R'.$$

We can then test all the intercepts by a Wald test, $\hat{\alpha} \text{Var}(\hat{\alpha})^{-1} \hat{\alpha}$ which follows a χ^2_2 distribution.

Extensions to more than two regression equations are straightforward and discussed below.

Proof (of (6.3)–(6.5)) Similarly to the single-equation OLS, we can write

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + S_{xx}^{-1} \sum_{t=1}^T x_t u_{1t} \\ \hat{\beta}_2 &= \beta_2 + S_{xx}^{-1} \sum_{t=1}^T x_t u_{2t} \end{aligned}$$

The variance (matrix) of $\hat{\beta}_1$ or of $\hat{\beta}_2$ follows the same pattern as for single-equation OLS. In contrast, the covariance (matrix) is

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = S_{xx}^{-1} \text{Cov}(\sum_{t=1}^T x_t u_{1t}, \sum_{t=1}^T x_t u_{2t}) S_{xx}^{-1}.$$

Together, this gives (6.4)–(6.5). If the residuals are iid and independent of the regressors, then $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ simplifies to $\sigma_{12} S_{xx}^{-1}$ which gives (6.3). \square

6.1.1 Calendar Time Regressions

To investigate how the performance (alpha) or exposure (betas) of different investors/funds are related to investor/fund characteristics, we often use the *calendar time* (CalTime) approach. First define M ($M = 2$ in this section) discrete investor groups (for instance, age 18–30, 31–40, etc) and calculate their respective average excess returns (\bar{R}_{jt}^e for group j)

$$\bar{R}_{jt}^e = \sum_{i \in \text{Group}_j} R_{it}^e / N_j, \quad (6.6)$$

where N_j is the number of individuals in group j .

Then, we run a factor model (6.1)–(6.2), where $y_{jt} = \bar{R}_{jt}^e$ and where x_t typically includes a constant and various return factors (for instance, excess returns on equity and bonds). By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the “alpha”) is higher for the M th group than for the first group.

6.2 A System of n OLS Regressions

Remark 6.3 (Kronecker product) Let \otimes denote the Kronecker product, that is, if A and B are matrices, then

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

For instance, with

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \text{ and } B = \begin{bmatrix} 10 & 11 \end{bmatrix}, \text{ we get } A \otimes B = \begin{bmatrix} 10 & 11 & 30 & 33 \\ 20 & 22 & 40 & 44 \end{bmatrix}.$$

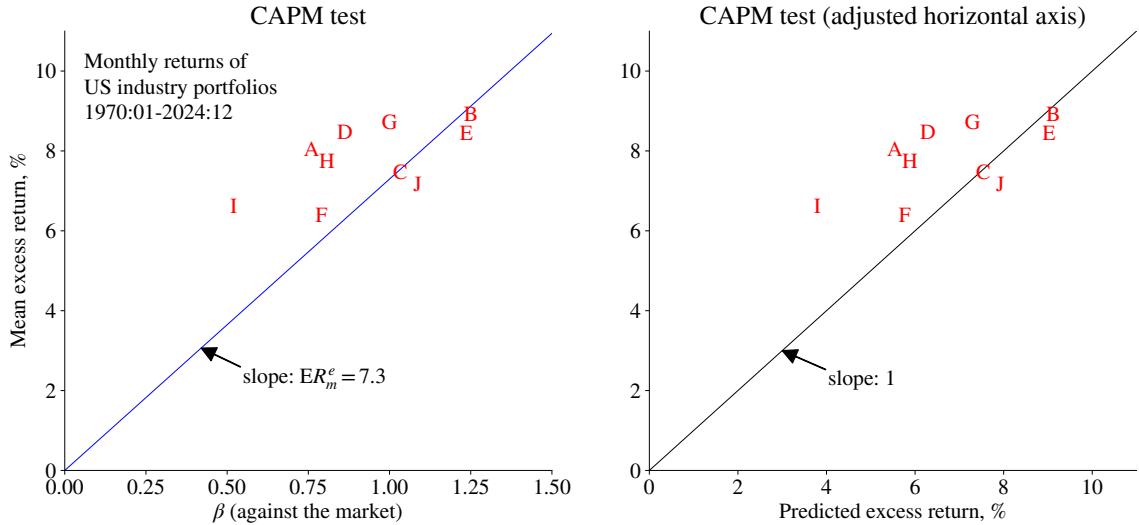
Let $\hat{\beta}$ be the vector consisting of $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$ stacked in order (equation 1 first, then equation 2, etc). With iid residuals the variance-covariance matrix (instead of (6.3)) is

$$\text{Var}(\hat{\beta}) = \Sigma \otimes S_{xx}^{-1}, \quad (6.7)$$

where $\Sigma = \text{Var}(u_t)$ is the $n \times n$ variance-covariance matrix of the n residuals.

Similarly, with non-iid residuals we get (instead of (6.4))

$$\text{Var}(\hat{\beta}) = (I_n \otimes S_{xx}^{-1}) \Omega (I_n \otimes S_{xx}^{-1}), \quad (6.8)$$



	α (ann.)	t-stat	σ (ann.)
A (NoDur)	2.44	2.06	8.73
B (Durbl)	-0.25	-0.11	17.26
C (Manuf)	-0.12	-0.14	6.43
D (Enrgy)	2.13	0.93	16.84
E (HiTec)	-0.63	-0.41	11.17
F (Telcm)	0.57	0.38	10.99
G (Shops)	1.38	1.13	8.97
H (Hlth)	1.81	1.22	10.94
I (Utils)	2.77	1.74	11.75
J (Other)	-0.80	-0.85	6.97

$$\begin{aligned} \text{CAPM: } R_i^e &= \alpha_i + \beta_i R_m^e + e_i \\ \text{Predicted excess return: } \beta_i R_m^e \\ \text{10% crit. value (Bonferroni): } 2.58 \\ \text{Test if all } \alpha_i = 0: \\ \text{Wald stat} &\quad 10.55 \\ \text{5% crit val} &\quad 18.31 \\ \text{p-value} &\quad 0.39 \end{aligned}$$

Figure 6.1: CAPM regressions on US industry indices

where I_n is the $n \times n$ identity matrix. Let u_t be the vector of the n residuals in period t $(u_{1t}, u_{2t}, \dots, u_{nt})$. Then, Ω is (instead of (6.5))

$$\Omega = \text{Var} \left(\sum_{t=1}^T u_t \otimes x_t \right), \quad (6.9)$$

which is an $nk \times nk$ matrix which can be estimated by the methods of White or Newey-West, applied to $(x_t u_{1t}, x_t u_{2t}, \dots, x_t u_{nt})$ stacked into a single vector.

Empirical Example 6.4 (CAPM on industry portfolios) Figure 6.1 shows results for the intercepts from regressing US industry portfolios on the market. The joint test is for whether all intercepts are zero.

Further Reading

See also Wooldridge (2010) 7.3, Greene (2018) 10 and Hansen (2022a) 11.

Chapter 7

Testing CAPM and Multifactor Models

7.1 Market Model

Let $R_{it}^e = R_{it} - R_{ft}$ be the excess return on asset i over the risk-free asset, and let R_{mt}^e be the excess return on the market portfolio. The basic implication of CAPM is that the expected excess return of an asset ($E R_{it}^e$) is linearly related to the expected excess return on the market portfolio ($E R_{mt}^e$) according to

$$E R_{it}^e = \beta_i E R_{mt}^e, \text{ where } \beta_i = \text{Cov}(R_i^e, R_m^e) / \text{Var}(R_m^e). \quad (7.1)$$

To test this, consider the regression

$$R_{it}^e = \alpha_i + b_i R_{mt}^e + \varepsilon_{it}. \quad (7.2)$$

Take expectations of the regression (assuming we know the coefficients) to get

$$E R_{it}^e = \alpha_i + b_i E R_{mt}^e. \quad (7.3)$$

Notice that the LS estimate of b_i is the sample analogue to β_i in (7.1). It is then clear that CAPM implies that the intercept (α_i) of the regression should be zero, which is also what empirical tests of CAPM focus on.

This test of CAPM can be given two interpretations. If we assume that R_{mt}^e is the correct benchmark (the tangency portfolio for which (7.1) is true by definition), then it is a test of whether asset R_{it}^e is correctly priced. This is typically the perspective in performance analysis of mutual funds. Alternatively, if we assume that R_{it}^e is correctly priced, then it is a test of the mean-variance efficiency of R_{mt}^e . That is, we test if the market portfolio is the correct “pricing factor” of all the test assets. This is the perspective of CAPM tests.

The test of the null hypothesis that $\alpha_i = 0$ uses the fact that, under fairly mild conditions, the t-statistic has an asymptotically normal distribution, that is

$$\hat{\alpha}_i / \text{Std}(\hat{\alpha}_i) \xrightarrow{d} N(0, 1) \text{ under } H_0 : \alpha_i = 0. \quad (7.4)$$

We get $\text{Std}(\hat{\alpha}_i)$ from the OLS regression (possibly with an adjustment due to autocorrelation and/or heteroskedasticity).

The test assets are typically portfolios of firms with similar characteristics, for instance, small size or having their main operations in the retail industry. There are two main reasons for testing the model on such portfolios: individual stocks are very volatile and firms can change substantially over time (so the beta changes), whereas the portfolios can be constructed to represent fairly constant characteristics. For instance, a portfolio of small firms could include the firms in the lowest size decile over the previous year (and thus being rebalanced annually). Moreover, it is of interest to see how the deviations from CAPM are related to firm characteristics (size, industry, etc), since that may suggest how the model should be amended. The empirical results from such tests vary with the test assets used.

Empirical Example 7.1 (CAPM on industry portfolios) Figure 7.1 shows results for US industry portfolios. Here CAPM works reasonably well.

In Figure 7.1, the results are presented in two different ways:

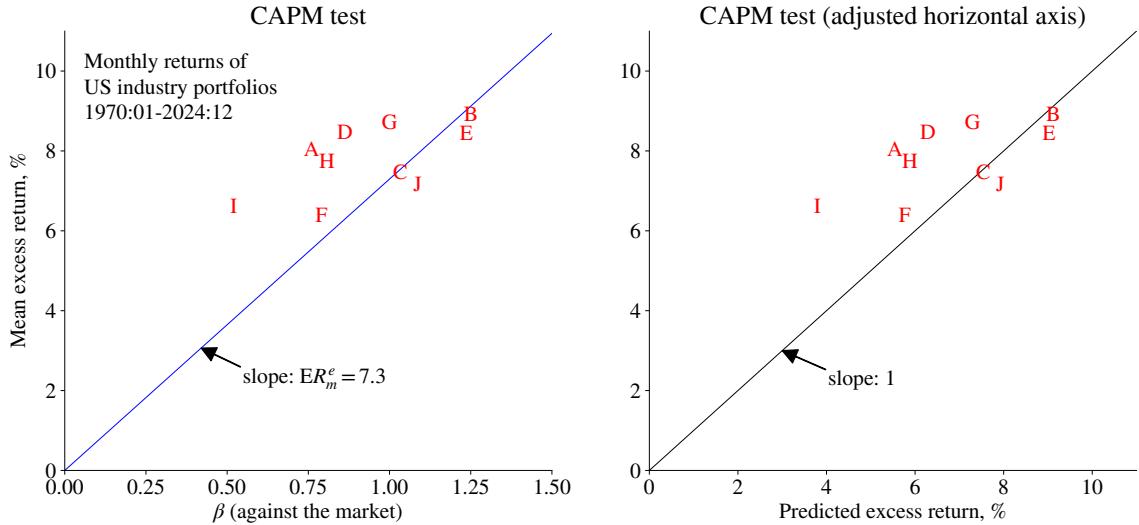
	<u>horizontal axis</u>	<u>vertical axis</u>
1 : β_i	\bar{R}_{it}^e	
2 : $\beta_i \bar{R}_{mt}^e$	\bar{R}_{it}^e	

(7.5)

where \bar{R}_{it}^e is the (time series) average excess return of asset i . In the first approach, CAPM says that all data points (different assets, i) should cluster around a straight line with a slope equal to the average market excess return, \bar{R}_{mt}^e . In the second approach, CAPM says that all data points should cluster around a line with a slope of one. In either case, the vertical distance to the line is α_i (which should be zero according to CAPM).

7.1.1 Key Properties of the CAPM Test

It is typically found that these tests require a substantial deviation from CAPM and/or a long sample to get good power. The basic reason for this is that asset returns are very volatile. For instance, suppose that the standard OLS assumptions (iid residuals that are



	α (ann.)	t -stat	σ (ann.)
A (NoDur)	2.44	2.06	8.73
B (Durbl)	-0.25	-0.11	17.26
C (Manuf)	-0.12	-0.14	6.43
D (Enrgy)	2.13	0.93	16.84
E (HiTec)	-0.63	-0.41	11.17
F (Telcm)	0.57	0.38	10.99
G (Shops)	1.38	1.13	8.97
H (Hlth)	1.81	1.22	10.94
I (Utils)	2.77	1.74	11.75
J (Other)	-0.80	-0.85	6.97

$$\begin{aligned} \text{CAPM: } R_i^e &= \alpha_i + \beta_i R_m^e + e_i \\ \text{Predicted excess return: } \beta_i R_m^e \\ \text{10% crit. value (Bonferroni): } 2.58 \\ \text{Test if all } \alpha_i = 0: \\ \text{Wald stat} &\quad 10.55 \\ \text{5% crit val} &\quad 18.31 \\ \text{p-value} &\quad 0.39 \end{aligned}$$

Figure 7.1: CAPM regressions on US industry indices

independent of the market return) are correct. Then, it is straightforward to show that the variance of Jensen's alpha is

$$\text{Var}(\hat{\alpha}_i) = (1 + SR_m^2)\sigma^2/T, \quad (7.6)$$

where σ^2 is the variance of the residual in (7.2) and SR_m is the Sharpe ratio of the market portfolio. We see that the uncertainty about the alpha is high when the residual is volatile and when the sample is short, but also when the Sharpe ratio of the market is high. Note that a large market Sharpe ratio means that the market asks for a high compensation for taking on risk. A lot of uncertainty about how risky asset i is then translates in a large uncertainty about what the risk-adjusted return should be.

Example 7.2 Suppose we have monthly data with $\hat{\alpha}_i = 0.2\%$ (2.4% per year), $\sigma = 3\%$ ($\approx 10\%$ per year) and a market Sharpe ratio of 0.15 (≈ 0.5 per year). These values

correspond well to US CAPM regressions for industry portfolios, see Figure 7.1. A significance level of 10% requires a t-statistic (7.4) of at least 1.64, so

$$\frac{0.2}{\sqrt{1 + 0.15^2} \sqrt{3} / \sqrt{T}} \geq 1.64 \text{ or } T \geq 626.$$

We need a sample of at least 626 months (52 years). With a sample of only 26 years (312 months), the alpha needs to be almost 0.3% per month (3.6% per year) or the standard deviation of the residual no more than 2% (7% per year). Notice that accumulating a 0.3% return over 25 years means almost 2.5 times the initial value.

Proof (*Proof of (7.6)) The chapter on OLS demonstrated that the variance of the intercept (α_i) can be written

$$\frac{\sigma^2}{T} \frac{\widehat{\text{Var}}(z) + \bar{z}^2}{\widehat{\text{Var}}(z)}$$

where z is the non-constant regressor, here R_{mt}^e . Simplify and notice that $\bar{z}^2 / \widehat{\text{Var}}(z)$ corresponds to the squared Sharpe ratio of the market return. \square

7.1.2 Interpretation of the CAPM Test*

Instead of a t-test, we can use the equivalent chi-square test

$$(t\text{-stat})^2 = \hat{\alpha}_i^2 / \text{Var}(\hat{\alpha}_i) \xrightarrow{d} \chi_1^2 \text{ under } H_0: \alpha_i = 0. \quad (7.7)$$

It is quite straightforward to use the properties of minimum-variance frontiers (see Gibbons, Ross, and Shanken (1989), and also MacKinlay (1995)) to show that the test statistic in (7.7) can be written

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} = \frac{SR_c^2 - SR_m^2}{(1 + SR_m^2)/T}, \quad (7.8)$$

where SR_m is the Sharpe ratio of the market portfolio and SR_c is the Sharpe ratio of the tangency portfolio when investment in both the market return and asset i is possible. (Recall that the tangency portfolio is the portfolio with the highest possible Sharpe ratio.) If the market portfolio has the same (squared) Sharpe ratio as the tangency portfolio of the mean-variance frontier of R_{it} and R_{mt} (so the market portfolio is mean-variance efficient also when we take R_{it} into account) then the test statistic, $\hat{\alpha}_i^2 / \text{Var}(\hat{\alpha}_i)$, is zero and CAPM is not rejected.

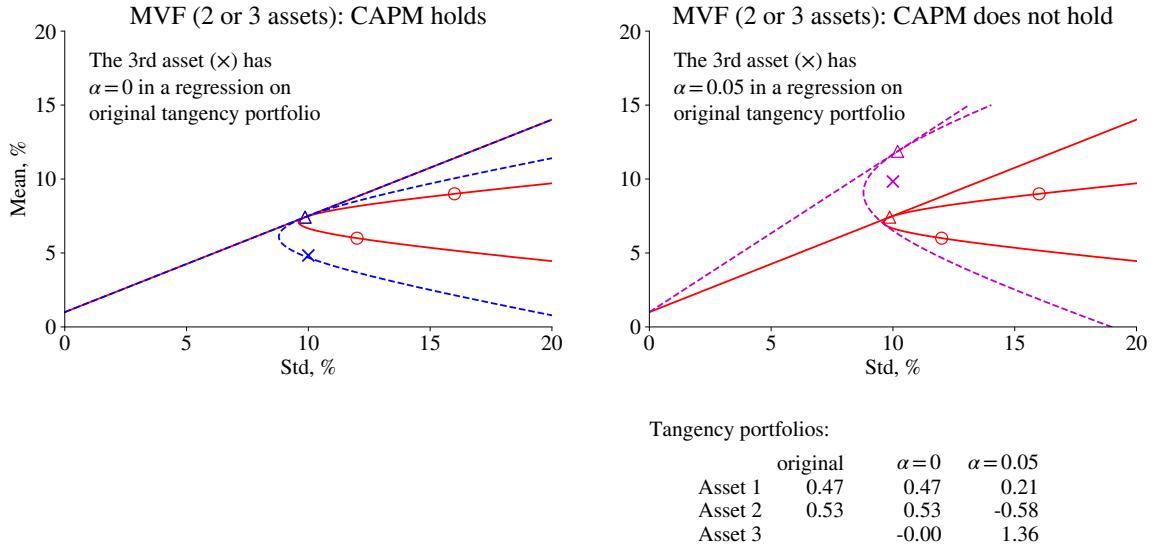


Figure 7.2: Effect on MV frontier of adding assets

Proof (*Proof of (7.8)) From the CAPM regression (7.2) we have

$$\text{Var} \begin{pmatrix} R_{it}^e \\ R_{mt}^e \end{pmatrix} = \begin{bmatrix} \beta_i^2 \sigma_m^2 + \text{Var}(\varepsilon_{it}) & \beta_i \sigma_m^2 \\ \beta_i \sigma_m^2 & \sigma_m^2 \end{bmatrix}, \text{ and } \begin{bmatrix} \mu_i^e \\ \mu_m^e \end{bmatrix} = \begin{bmatrix} \alpha_i + \beta_i \mu_m^e \\ \mu_m^e \end{bmatrix}.$$

Suppose we use this information to construct a mean-variance frontier for both R_{it} and R_{mt} , and we find the tangency portfolio, with excess return R_{ct}^e . It is straightforward to show that the square of the Sharpe ratio of the tangency portfolio is $\mu^e' \Sigma^{-1} \mu^e$, where μ^e is the vector of expected excess returns and Σ is the covariance matrix. By using the covariance matrix and mean vector above, we get that the squared Sharpe ratio for the tangency portfolio, $\mu^e' \Sigma^{-1} \mu^e$, (using both R_{it} and R_{mt}) is

$$\left(\frac{\mu_c^e}{\sigma_c} \right)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + \left(\frac{\mu_m^e}{\sigma_m} \right)^2 \text{ or } (SR_c)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + (SR_m)^2.$$

Combine this with (7.6) which shows that $\text{Var}(\hat{\alpha}_i) = [1 + (SR_m)^2] \text{Var}(\varepsilon_{it})/T$. \square

This is illustrated in Figure 7.2 which shows the effect of adding an asset to the investment opportunity set. In general, we would expect that adding an asset to the investment opportunity set would expand the mean-variance frontier (and it does) and that the tangency portfolio changes accordingly. However, the tangency portfolio is not changed by adding an asset with a zero alpha (intercept). The intuition is that such an asset has neutral performance compared to the market portfolio (obeys the beta representation), so investors should stick to the market portfolio.

7.1.3 Several Assets

In most cases, there are several (n) test assets, and we actually want to test if all the α_i (for $i = 1, 2, \dots, n$) are zero (otherwise CAPM is not correct). Ideally we then want to take into account the correlation of the different alphas. Such a system based approach is discussed in another chapter.

Alternatively, we can apply a Bonferroni correction of the individual t-stats: reject CAPM at the 10% significance level only if the largest t-stat (in absolute terms) exceeds the critical value at the $0.10/n$ significance level. For instance, with $n = 25$, the 10% critical value from a standard normal distribution would be 2.88 instead of 1.64.

Remark 7.3 *Fama and French (1993) and Fama and French (1996)) construct 25 stock portfolios according to two characteristics of the firm: the size (by market capitalization) and the book-value-to-market-value ratio (BE/ME). In June each year, they sort the stocks according to size and BE/ME. They then form a 5×5 matrix of portfolios, where portfolio ij belongs to the i th size quintile (quintiles divide sorted data into fifths of the sample) and the j th BE/ME quintile (a double-sort)*

$$\begin{bmatrix} \text{small size, low B/M} & \dots & \dots & \dots & \text{small size, high B/M} \\ \vdots & & \ddots & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \\ \text{large size, low B/M} & & & & \text{large size, high B/M} \end{bmatrix}$$

They run a traditional CAPM regression on each of the 25 portfolios (monthly data 1963–1991. It is found that there is almost no relation between the CAPM prediction and actual average returns.

Empirical Example 7.4 *Figure 7.1 reports also a joint test of the hypothesis that the alphas for all industry portfolios are zero. Also, Table 7.1 reports t-stats for the 25 FF portfolios. The Bonferroni adjusted critical values are 2.88 and 3.09 on the 10% and 5% significance levels.*

Empirical Example 7.5 *(Testing CAPM on FF portfolios) Figure 7.3 illustrates the CAPM pricing errors (alphas) for the 25 FF (US size/book-to-market) portfolios.*

	1	2	3	4	5
1	-3.45	-0.09	0.44	2.11	2.52
2	-2.33	0.53	1.43	2.30	1.84
3	-2.19	1.28	1.20	2.22	2.22
4	-0.86	0.39	1.21	2.06	1.54
5	0.24	1.23	1.17	0.11	0.97

Table 7.1: t-stats for α in CAPM, 25 FF portfolios 1970:01-2024:12. NW uses 1 lag. The Bonferroni adjusted 10% and 5% critical values are 2.88 and 3.09.

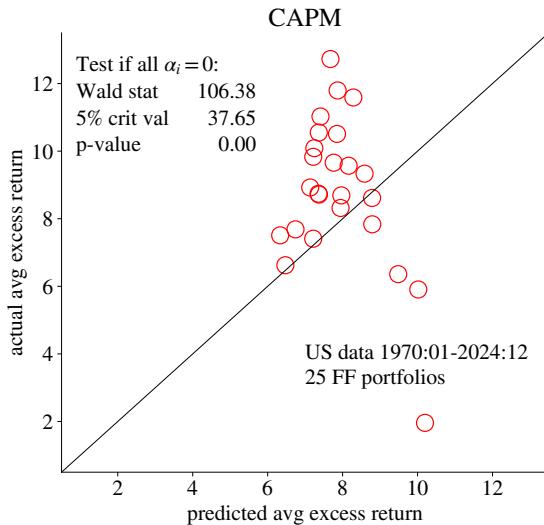


Figure 7.3: CAPM, FF portfolios

A quite different approach to study a cross-section of assets is to first estimate CAPM regressions (7.2) for each of the assets ($i = 1, 2, \dots, n$) and then the following cross-sectional regression

$$\bar{R}_i^e = \gamma + \lambda \hat{\beta}_i + u_i, \quad (7.9)$$

where \bar{R}_i^e is the (sample) average excess return on asset i . Notice that the estimated betas are used as regressors and that there are as many data points as there are assets (n). There are severe econometric problems with this regression equation since the regressor ($\hat{\beta}_i$) contains measurement errors (it is only an uncertain estimate), which tends to bias the slope coefficient (λ) towards zero. If we could overcome this bias, then the testable implications of CAPM is that $\gamma = 0$ and that λ equals the average market excess return. We also want (7.9) to have a high R^2 , since it should be unity in a very large sample (if

CAPM holds).

Remark 7.6 (*Representative results on mutual fund performance*) Mutual fund evaluations (estimated α_i) typically find (i) on average neutral performance (or less: trading costs&fees); (ii) large funds might be worse; (iii) perhaps better performance on less liquid (less efficient?) markets; and (iv) there is very little persistence in performance: α_i for one sample does not predict α_i for subsequent samples (except for bad funds).

7.2 Several Factors

In multifactor models, (7.2) is still valid—provided we reinterpret b_i and R_{mt}^e as vectors, so $b_i R_{mt}^e$ stands for $b_{io} R_{ot}^e + b_{ip} R_{pt}^e + \dots$

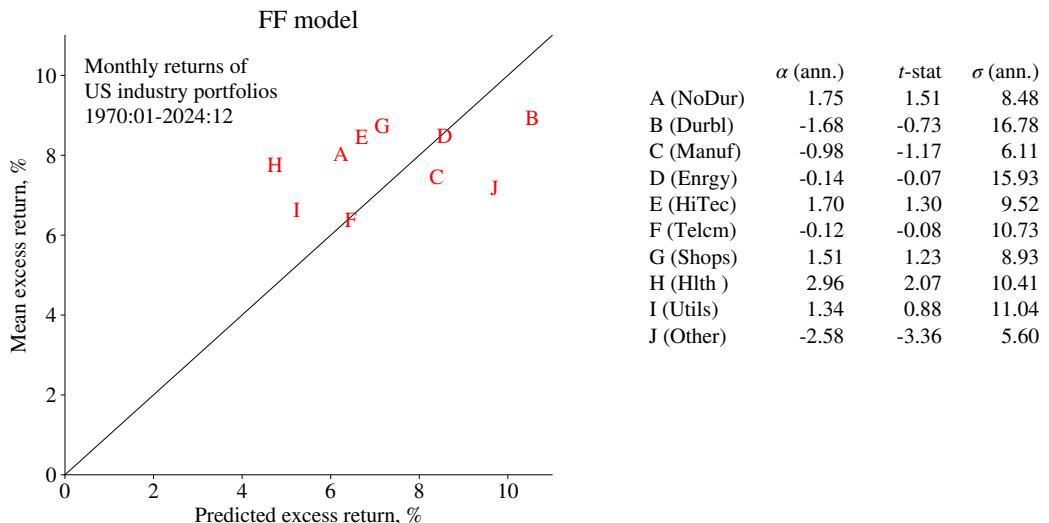
$$R_{it}^e = \alpha + b_{io} R_{ot}^e + b_{ip} R_{pt}^e + \dots + \varepsilon_{it}. \quad (7.10)$$

In this case, (7.2) is a multiple regression, but the test (7.4) still has the same form (the standard deviation of the intercept will be different, though).

Fama and French (1993) also try a multi-factor model. They find that a three-factor model fits the 25 stock portfolios fairly well. The three factors are: the market return, the return on a portfolio of small stocks minus the return on a portfolio of big stocks (SMB), and the return on a portfolio with high BE/ME minus the return on portfolio with low BE/ME (HML). This three-factor model is rejected at traditional significance levels, but it can still capture a fair amount of the variation of expected returns.

Remark 7.7 (*Returns on long-short portfolios**) Suppose you invest x USD into asset i , but finance that by short-selling asset j . (You sell enough of asset j to raise x USD.) The net investment is then zero, so there is no point in trying to calculate an overall return like “value today/investment yesterday - 1.” Instead, the convention is to calculate an excess return of your portfolio as $R_i - R_j$ (or equivalently, $R_i^e - R_j^e$). This excess return essentially says: if your exposure (how much you invested) is x , then you have earned $x(R_i - R_j)$. To make this excess return comparable with net returns, you add the risk-free rate: $R_i - R_j + R_f$, implicitly assuming that your portfolio includes a risk-free investment of the same size as your long-short exposure (x).

Chen, Roll, and Ross (1986) use a number of macro variables as factors, in addition to traditional market indices. They find that industrial production and inflation surprises are priced factors, whereas the market index may not be.



Fama-French model
 Factors: US market, SMB (size), and HML (book-to-market)
 Predicted excess return: $\beta_m R_m^e + \beta_{SMB} R_{SMB} + \beta_{HML} R_{HML}$

10% crit. value (Bonferroni): 2.58

Test if all $\alpha_i = 0$:
 Wald stat 23.28
 5% crit val 18.31
 p-value 0.01

Figure 7.4: Fama-French regressions on US industry indices

Empirical Example 7.8 (Testing a 3-factor model) Figure 7.4 shows results for the Fama-French 3-factor model on US industry portfolios and 7.5 on the 25 Fama-French portfolios. The improvement over CAPM is modest for the industry portfolios, but considerable for the 25 FF portfolios.

Further Reading

See Elton, Gruber, Brown, and Goetzmann (2014) 15, Campbell (2018) 3 and Cochrane (2005) 12 for more results on testing linear factor models.

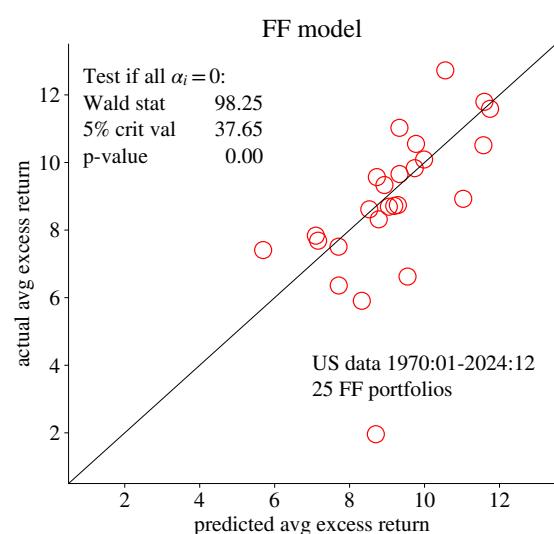


Figure 7.5: FF, FF portfolios

Chapter 8

Model Selection and Other Topics

8.1 Model Selection I

Excluding a relevant regressor will cause a bias of all coefficients, unless the included regressors are uncorrelated with the excluded regressor. In contrast, *including an irrelevant regressor* is not really dangerous, but is likely to decrease the precision.

To select the regressors, consider the following rules. Rule 1: use *economic theory*; rule 2: *avoid data mining* and mechanical searches for the right regressors; rule 3: maybe use a *general-to-specific approach*—start with a general regression and test restrictions,..., keep making it simpler until restrictions are rejected; rule 4: always *include a constant*; rule 5: avoid overfitting by “punishing” models with many parameters.

Remember that R^2 can never decrease by adding more regressors. To avoid overfitting, we could instead consider \bar{R}^2

$$\bar{R}^2 = 1 - (1 - R^2)(T - 1)/(T - k), \quad (8.1)$$

where T is the sample size and k is the number of regressors (including the constant). This measure includes trade-off between fit and the number of regressors (per data point). Notice that \bar{R}^2 can be negative (while $0 \leq R^2 \leq 1$). Clearly, the model must include a constant for R^2 (and therefore \bar{R}^2) to make sense. Alternatively, apply Akaike’s Information Criterion (AIC) or the Bayesian information criterion (BIC)

$$AIC = \ln \sigma^2 + 2k/T \quad (8.2)$$

$$BIC = \ln \sigma^2 + (k/T) \ln T, \quad (8.3)$$

where σ^2 is the variance of the fitted residuals. These measures also involve trade-offs between fit (low σ^2) and number of parameters (k , including the intercept). Select the

model with the *highest* \bar{R}^2 or *lowest* AIC or BIC.

Remark 8.1 (*Alternative expressions for AIC and BIC**) We can rewrite (8.2)–(8.3) as $AIC = \ln \text{Var}(y_t) + \ln(1 - R^2) + 2k/T$ and $BIC = \ln \text{Var}(y_t) + \ln(1 - R^2) + (k/T) \ln T$. This follows from using $R^2 = 1 - \sigma^2 / \text{Var}(y_t)$. Both expressions are decreasing in R^2 , but increasing in the number of regressors per data point (k/T). It therefore leads to a similar trade-off as in \bar{R}^2 .

Empirical Example 8.2 (*Empirical application of model selection*) See Table 8.1 for an empirical example showing a number of possible model specifications. The dependent variable is the monthly realized variance of S&P 500 returns (calculated from daily returns). The regressors considered are lags of the dependent variable, the VIX index and the S&P 500 returns. Similarly, see Table 8.2 for the best specification according to AIC. Notice that AIC tend to favour fairly large models with many regressors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RV_{t-1}	0.66 (8.91)						0.17 (2.14)
RV_{t-2}		0.45 (5.65)					-0.03 (-0.44)
VIX_{t-1}			0.93 (10.30)				0.96 (3.62)
VIX_{t-2}				0.66 (8.87)			-0.26 (-1.21)
R_{t-1}					-0.82 (-3.58)		0.00 (0.01)
R_{t-2}						-0.45 (-2.30)	-0.08 (-0.89)
constant	5.13 (4.94)	8.30 (6.62)	-2.71 (-1.82)	2.46 (1.91)	15.79 (18.83)	15.48 (18.08)	-0.51 (-0.43)
R^2	0.44	0.20	0.53	0.27	0.15	0.04	0.55
\bar{R}^2	0.43	0.20	0.53	0.26	0.14	0.04	0.55
obs	397	397	397	397	397	397	397

Table 8.1: Regression of monthly realized S&P 500 return volatility 1990:02-2024:12. Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

	(1)	(2)	(3)
RV _{t-1}	0.19 (2.18)		0.17 (2.00)
RV _{t-2}			
VIX _{t-1}	0.94 (5.51)	1.11 (6.22)	0.96 (5.32)
VIX _{t-2}	-0.27 (-2.01)	-0.24 (-1.73)	-0.29 (-2.04)
R _{t-1}			
R _{t-2}			-0.08 (-0.89)
constant	-0.61 (-0.68)	-1.80 (-1.53)	-0.33 (-0.41)
R ²	0.55	0.54	0.55
BIC	3.74	3.75	3.75
obs	397	397	397

Table 8.2: Regression of monthly realized S&P 500 return volatility 1990:02-2024:12. Ordered from best (1) according to BIC to third best (3). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

8.2 Model Selection II

Reference: Hastie, Tibshirani, and Friedman (2001) 3

In some cases, even good economic theory leaves us with too many potential regressors. This is often the case when developing forecasting models, where it also often noticed that models with many predictors tend to fail out of sample. At this point, it becomes crucial to apply a model selection technique, specifically a method that sets some regression coefficients to zero.

If there are k potential regressors, then there are 2^k different models. If the list of models is not too long, then we could compare AIC or BIC (8.2)–(8.3) across all possible models, see Table 8.2. Otherwise, we need some sort of sequential approach.

Example 8.3 (3 potential regressors) If the three potential regressors are 1, x_1 and x_2 , then the list of models has $2^3 - 1 = 7$ possibilities: 1, x_1 , x_2 , $(1, x_1)$, $(1, x_2)$, (x_1, x_2) , and $(1, x_1, x_2)$.

	(1)	(2)	(3)
RV _{t-1}		0.19 (2.18)	
RV _{t-2}			
VIX _{t-1}	0.93 (10.30)	1.11 (6.22)	0.94 (5.51)
VIX _{t-2}		-0.24 (-1.73)	-0.27 (-2.01)
R _{t-1}			
R _{t-2}			
constant	-2.71 (-1.82)	-1.80 (-1.53)	-0.61 (-0.68)
R ²	0.53	0.54	0.55
obs	397	397	397

Table 8.3: Best three regressions of monthly realized S&P 500 return volatility according to a forward step selection (based on t-stats), 1990:02-2024:12. Ordered from smallest model (1) to third smallest model (3). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

A *forward stepwise selection* is as follows

- (1) start with an intercept (8.4)
- (2) add the variable that improves the fit the most
- (3) repeat (2) until the fit does not improve much

To specify a stopping rule, first define the residual sum of squares (for a given vector of coefficients, β) as

$$RSS(\beta) = \sum_{t=1}^T (y_t - x'_t \beta)^2. \quad (8.5)$$

In step (2) we would then add the variable that gives the lowest RSS (when added to the previous selection). In step (3), it is often recommended that we stop adding regressors when

$$\frac{RSS(\hat{\beta}_{\text{old}}) - RSS(\hat{\beta}_{\text{new}})}{RSS(\hat{\beta}_{\text{new}})/(T - k - 1)} < c_{1,T-k-1}, \quad (8.6)$$

where k is the number of coefficients in $\hat{\beta}_{\text{old}}$ (including the intercept) so there are $k +$

1 coefficients in $\hat{\beta}_{\text{new}}$ and $c_{1,T-k-1}$ is the 90% or 95% critical value of an $F_{1,T-k-1}$ distribution. For instance, the 90% critical value of $F_{1,100}$ equals 2.76.

As an alternative to the *RSS* based rule in (8.5)–(8.6), we could instead use t-stats: in step (2) add the variable with the highest $|t\text{-stat}|$ (in a multiple regression together with the already included variables) and in step (3) stop adding variables when that $|t\text{-stat}|$ is lower than 1.64 (or 1.96).

Empirical Example 8.4 (*Forward stepwise selection*) Applying the forward step selection approach (based on t-stats) to the regression discussed in Example 8.2 gives a sequence of larger and larger models shown in Table 8.3.

	(1)	(2)	(3)
RV_{t-1}		0.15 (2.17)	0.15 (1.89)
RV_{t-2}			
VIX_{t-1}	0.93 (10.30)	0.76 (7.70)	0.72 (7.15)
VIX_{t-2}			
R_{t-1}			-0.21 (-2.10)
R_{t-2}			
constant	-2.71 (-1.82)	-1.85 (-1.43)	-0.73 (-0.68)
R^2	0.53	0.54	0.54
$\sum b_i $	0.73	0.75	0.81
obs	397	397	397

Table 8.4: Best three regressions of monthly realized S&P 500 return volatility where the model are selected by lasso, but then estimated with OLS, 1990:02-2024:12. Ordered from smallest model (1) to third smallest model (3). Numbers in parentheses are t-stats, based on Newey-West with 4 lags. The $\sum |b_i|$ is for regressions using standardized variables.

8.2.1 The Lasso, Ridge and Elastic Net Regressions*

An alternative approach to model selection is the *Lasso method*, which minimizes the sum of squared residuals (just like OLS), but with a penalty on $\sum_{i=1}^k |b_i|$. This introduces a

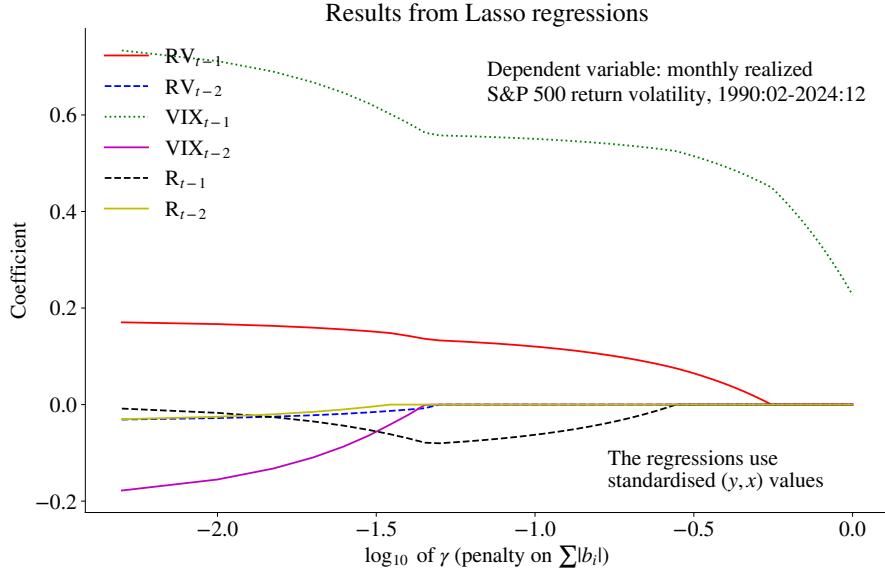


Figure 8.1: Lasso regressions

tradeoff between the fit and the magnitude of the coefficients: a higher penalty will lead to coefficients closer to zero. In short, the method solves the following optimization problem

$$\min_b \sum_{t=1}^T (y_t - \alpha - x'_t b)^2 + \gamma \sum_{i=1}^k |b_i|. \quad (8.7)$$

Having the same penalty γ for all coefficients makes perhaps most sense when the regressors have the same scale, for instance, unit standard deviation (and perhaps also zero mean). The *adaptive Lasso* instead uses weighted penalties, $\gamma \sum_{i=1}^k w_i |b_i|$, often with $w_i = 1/|b_i^{OLS}|$. This gives a larger effective penalty on variables whose OLS coefficients are small in absolute terms. It is sometimes an advantage to divide the first term in (8.7) by T , to make the interpretation of γ independent of the sample size, and sometimes also for numerical reasons.

Clearly, when $\gamma = 0$, then the lasso method reproduces the OLS estimates. For larger values of γ , the lasso will give smaller coefficients: some b_i will be zero and others tend to be closer to 0 than OLS would suggest, similar to other “shrinkage” methods like a ridge estimation (see the remark below) which has a penalty on $\sum_{i=1}^k b_i^2$. The difference is that the Lasso method will lead to some of the coefficients to be exactly zero, while the ridge regression will just reduce them towards zero. The remark below (on elastic net regression) discusses the case when the penalty term is $\gamma \sum_{i=1}^k |b_i - \beta_{i0}|$, where β_{i0} is a “target value.”

The lasso method can be used as a model selection technique by estimating a sequence of models with increasingly higher γ values. With a sufficiently high γ , only one coefficient is non-zero, which is for the most important variable. For a somewhat lower γ value, two coefficients are non-zero and so on. See Figure 8.1. Once the L (three, say) smallest specifications are found, we could re-estimate each of them with OLS. (This is the lars-OLS hybrid discussed in Efron, Hasti, Johnstone, and Tibshirani (2004).)

Empirical Example 8.5 (*Lasso regression*) Applying the Lasso approach to the regression discussed in Example 8.2 gives a sequence of smaller and smaller models. Figure 8.1 shows how the coefficients of the normalised variables change as penalty parameter is increased. Re-estimating the four smallest of those models with OLS gives the results in Table 8.4.

Remark 8.6 (*Application of the lasso/lars algorithms*) These algorithms often standardize x_t to have zero means and unit standard deviations, and y_t to have zero means (and perhaps unit standard deviation)

Remark 8.7 (*Ridge regression**) Let (y_t, x_t) be demeaned, so no intercept is needed, and let the target slope coefficients be the vector β_0 . The ridge regression then gives

$$\hat{b} = (\sum_{t=1}^T x_t x_t' + \lambda I)^{-1} (\sum_{t=1}^T x_t y_t + \lambda \beta_0).$$

This \hat{b} solves the problem $\min_b \sum_{t=1}^T (y_t - x_t' b)^2 + \lambda \sum_{i=1}^k (b_i - \beta_{i0})^2$. Notice that $\lambda = 0$ gives OLS, while $\lambda = \infty$ gives $\hat{b} = \beta_0$. In many applications, $\beta_{i0} = 0$.

Remark 8.8 (*Elastic net regression**) An elastic net regression is a mix of a Lasso regression and a ridge regression. It solves $\min_b \sum_{t=1}^T (y_t - \alpha - x_t' b)^2 + \gamma \sum_{i=1}^k |b_i - \beta_{i0}| + \lambda \sum_{i=1}^k (b_i - \beta_{i0})^2$. In many applications, $\beta_{i0} = 0$. Clearly, $\lambda = 0$ gives a lasso method with a target vector β_0 , while setting $\gamma = 0$ gives a ridge regression.

8.3 Weighted Least Squares

Weighted least squares (WLS) is used in several contexts: for implementations of generalised least squares (GLS) and also in non-parametric regressions. This brief section introduces the key results.

Suppose we want to put the weight w_t on observation (x_t, y_t) in a regression, $y_t = x'_t \beta + u_t$. That is, we want to minimize

$$\sum_{t=1}^T w_t (y_t - x'_t b)^2. \quad (8.8)$$

The first order conditions are

$$\mathbf{0}_{k \times 1} = \sum_{t=1}^T w_t x_t (y_t - x'_t \hat{\beta}), \quad (8.9)$$

so the solution is

$$\hat{\beta} = S_{wxx}^{-1} \sum_{t=1}^T w_t x_t y_t, \text{ where } S_{wxx} = \sum_{t=1}^T w_t x_t x'_t \quad (8.10)$$

It is straightforward to show that the variance-covariance matrix of $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = S_{wxx}^{-1} S S_{wxx}^{-1}, \text{ where } S = \text{Var}(\sum_{t=1}^T w_t x_t u_t). \quad (8.11)$$

We can estimate this by replacing u_t by the fitted residuals $\hat{u}_t = y_t - x'_t \hat{\beta}$ and apply the usual methods on the $T \times k$ matrix of “data” $w_t x_t u_t$. For instance, we get a heteroskedasticity consistent (like White’s) estimate by simply estimating a traditional $k \times k$ covariance matrix of $w_t x_t \hat{u}_t$.

Proof (of (8.11)*) Substitute for $y_t = x'_t \beta + u_t$ in (8.10) to get (after simplification)

$$\hat{\beta} = \beta + S_{wxx}^{-1} \sum_{t=1}^T w_t x_t u_t.$$

The result in (8.11) follows directly. \square

Remark 8.9 (*Alternative computation of WLS**) Let $\tilde{x}_t = \sqrt{w_t} x_t$ and $\tilde{y}_t = \sqrt{w_t} y_t$. Then, regressing \tilde{y}_t on \tilde{x}_t gives the same result as in (8.10). Notice also that (a) $\sum_{t=1}^T \tilde{x}_t \tilde{x}'_t$ is the same as S_{wxx} ; and (b) $\sum_{t=1}^T \tilde{x}_t \tilde{u}_t$ (where $\tilde{u}_t = \tilde{y}_t - \tilde{x}'_t \beta$) is the same as $\sum_{t=1}^T w_t x_t u_t$. This means that robust standard errors, for instance, White’s, from the regression of \tilde{y}_t on \tilde{x}_t gives the same variance-covariance matrix as in (8.11). WLS can thus be implemented as OLS on $(\tilde{y}_t, \tilde{x}_t)$, and this applies to both point estimates and the variance-covariance matrix.

8.4 Comparing Non-Nested Models

Consider two competing models

$$\text{Model A: } y_t = x_t' \beta + \varepsilon_t \quad (8.12)$$

$$\text{Model B: } y_t = z_t' \gamma + v_t. \quad (8.13)$$

For instance, these models could represent alternative economic theories of the same phenomenon. They are *non-nested* if z is not a subset of x at the same time as x is not a subset of z . Comparing the fit of these models starts with the usual criteria: R^2 , \bar{R}^2 , AIC, and BIC.

An alternative approach to compare the fit is to study *encompassing*. Model B is said to encompass model A if it can explain all that model A can (and more). This is clearly a good feature. To test this, run the regression

$$y_t = z_t' \gamma + x_{2t}' \delta_A + v_t, \quad (8.14)$$

where x_{2t} are those variables in x_t that are not also in z_t . Model B encompasses model A if $\delta_A = 0$ (test this restriction). Clearly, we can repeat this to see if A encompasses B .

8.5 Non-Linear Models

Regression analysis typically starts with a linear model—which may or may not be a good approximation.

Notice that models that are *non-linear in variables*

$$y_t = \alpha + \delta x_t^{3.4} + \varepsilon_t, \quad (8.15)$$

can be handled by OLS: just run OLS using $x_t^{3.4}$ as a regressor.

In contrast, models that are *non-linear in parameters*

$$y_t = \beta_1 + \beta_2 x_t^{\beta_3} + u_t \quad (8.16)$$

cannot be estimated by OLS, requiring an alternative approach, such as nonlinear least squares (NLS). This calls for using a numerical minimization routine to minimize the sum of squared residuals, $\sum_{t=1}^T u_t^2$. The NLS method can be seen as an application of GMM (discussed in another chapter).

To test the functional form (i.e., whether a linear specification is truly appropriate), estimate a nonlinear extension and test if the non-linear tests are significant. Alternatively, do a RESET test

$$y_t = x'_t \beta + \alpha_2 \hat{y}_t^2 + v_t, \quad (8.17)$$

where $\hat{y}_t = x'_t \hat{\gamma}$ (from a linear model, $y_t = x'_t \gamma + \varepsilon_t$). If the null hypothesis $\alpha_2 = 0$ cannot be rejected, then a linear model is good enough. Otherwise, we may need a non-linear specification.

A bin scatter plot can provide insights about the functional form. Suppose y_t depends on a vector of variables x_{1t} (which is possibly empty) and also a single variable x_{2t} . We want to investigate if y_t is linearly related to x_{2t} . To do that, create N non-overlapping bins for x_{2t} and let d_t be an N -vector of dummy variables indicating whether x_{2t} belongs to each of the bins. (Clearly, only one element in d_t is 1 and the rest are zero.) A common choice is to use the minimum, 10th percentile, 20th percentile,..., maximum of x_{2t} as bin boundaries (the minimum is clearly the lower boundary of the first bin and the maximum is the upper boundary of the last bin).

Example 8.10 (4 bins) With 4 bins, we would use the minimum, 0.25th percentile, the median, the 0.75th percentile and the maximum of x_{2t} . For instance, 0.05, 0.33, 1.0, 1.67 and 2.1. The first bin is then 0.05 to 0.33, the second is 0.33 to 1.0, and so on.

Then, run the regression

$$y_t = x'_{1t} \gamma + d'_t \beta + u_t \quad (8.18)$$

and report the estimates of the N -vector β and their standard errors (or confidence bands). If the elements in $\hat{\beta}$ follow a linear pattern when plotted against the mid values of the bins (not the bin number), then a linear model ($y_t = x'_{1t} \theta_1 + \theta_2 x_{2t} + u_t$) is well motivated. A piecewise linear regression

$$y_t = x'_{1t} \gamma + x_{2t} d'_t \lambda + u_t \quad (8.19)$$

is a related approach. It estimates a slope (λ_i) for each bin. If they are similar, then again a linear model is motivated. See Figure 8.2 for an example.

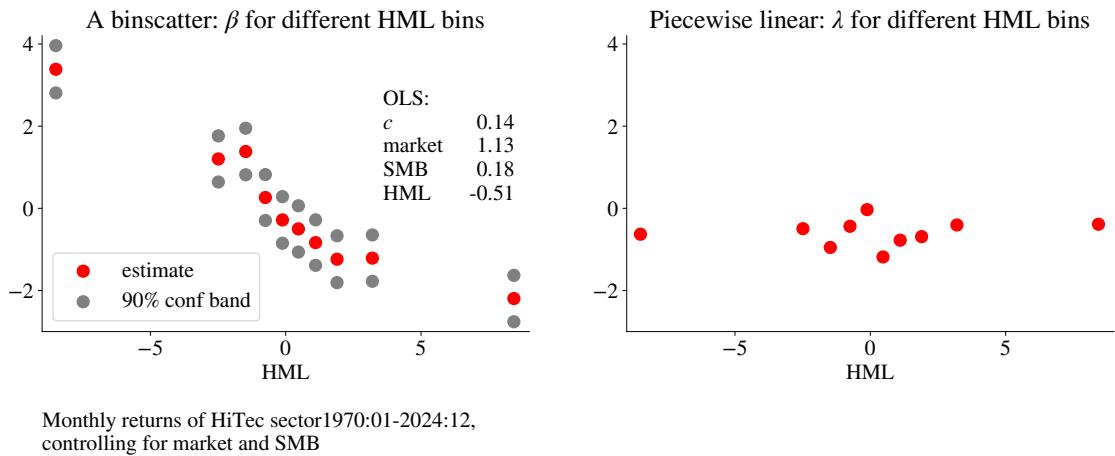


Figure 8.2: Binscatter and piecewise linear regression

8.6 Outliers

OLS is sensitive to extreme data points. The starting point (as always in empirical work) in detecting problems is to plot the data: time series plots and histograms, with the aim to study whether there are extreme data points.

Since the loss function defining OLS is quadratic, a few outliers can have a very large influence on the estimated coefficients. The loss function value will probably be lower if the coefficient is changed to pick up the outliers, even if this means that the errors for the other observations become larger (the sum of the square of many small errors can very well be less than the square of a single large error). See Figure 8.4.

There is of course nothing sacred about a quadratic loss function. Instead the sum of squared errors (as in OLS), one could, for instance, use a loss function in terms of the absolute value of the error $\sum_{t=1}^T |y_t - b_0 - b_1 x_t|$. This would produce the Least Absolute Deviation (LAD) estimator. It is typically less sensitive to outliers. Again see Figure 8.4. (A separate chapter will discuss LAD and other quantile regressions.)

However, LS is by far the most popular choice. There are two main reasons: (1) LS is very easy to compute and (2) it is fairly straightforward to construct (even robust) standard errors. (Also, OLS coincides with maximum likelihood when the errors are normally distributed, which means that they are the most precise estimates.)

It is a good idea to try to identify outliers from the regression results. First, estimate on the whole sample to get the estimates of the coefficients $\hat{\beta}$ and the fitted values \hat{y}_t . Second, estimate on the whole sample, except observation s , and record the estimate $\hat{\beta}^{(s)}$

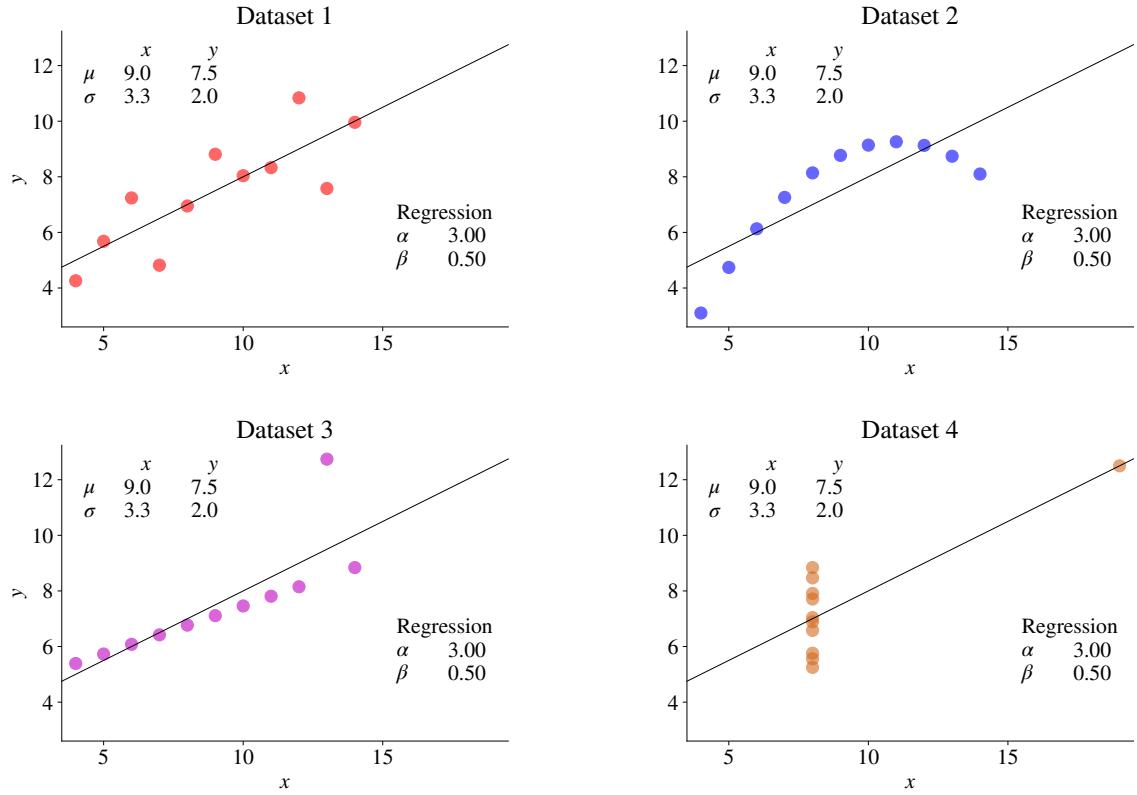


Figure 8.3: Model specification and descriptive statistics (Anscombe's quartet)

and the fitted value for period s (the one that was not used in the estimation) $\hat{y}_s^{(s)} = x_s' \hat{\beta}^{(s)}$. Repeat this for all data points (s) . Third, plot $\hat{\beta}^{(s)} - \hat{\beta}$, $\hat{y}_s^{(s)} - \hat{y}_s$ or $\hat{u}_s^{(s)} / \hat{\sigma}$. If these series make sudden jumps, then that data point is driving the results for the full sample. It then remains to determine whether this is good (a very informative data point) or bad (unrepresentative or even wrong data point). In particular, the *influence* of observation s (measured as $|\hat{y}_s^{(s)} - \hat{y}_s|$) is often plotted and analysed. See Figure 8.5 for an illustration.

Remark 8.11 (*An alternative way to calculate $\hat{\beta}^{(s)} - \hat{\beta}$ and $\hat{y}_s^{(s)} - \hat{y}_s$) Let $\hat{u}_t = y_t - x_t' \hat{\beta}$ be the residuals from the regression using the whole sample. It is then straightforward to show (see, for instance, Hansen (2022a) 3) that

$$\begin{aligned}\hat{\beta}^{(s)} - \hat{\beta} &= -S_{xx}^{-1} x_s \hat{u}_s / (1 - h_s), \text{ where} \\ S_{xx} &= \sum_{t=1}^T x_t x_t' \text{ and } h_s = x_s' S_{xx}^{-1} x_s.\end{aligned}$$

The h_s term is called the “leverage” since it will be a gauge of how important observation

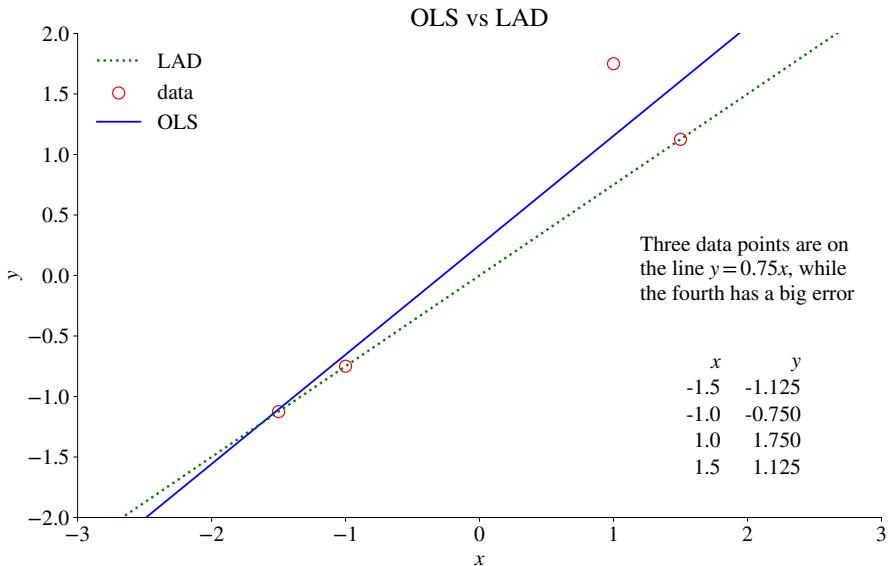


Figure 8.4: Data and regression line from OLS and LAD

s is. This is perhaps best seen by rewriting the difference in the fitted values as

$$\begin{aligned}\hat{y}_s^{(s)} - \hat{y}_s &= x_s'(\hat{\beta}^{(s)} - \hat{\beta}) \\ &= -x_s' S_{xx}^{-1} x_s \hat{u}_s / (1 - h_s) = -\hat{u}_s h_s / (1 - h_s).\end{aligned}$$

8.7 Estimation on Subsamples

To capture *time-variation in the regression coefficients*, it is fairly common to run the regression

$$y_t = x_t' \beta + \varepsilon_t \quad (8.20)$$

on a longer and longer data set (“recursive estimation”). In the standard recursive estimation, the first estimation is done on the sample $t = 1, 2, \dots, \tau$; while the second estimation is done on $t = 1, 2, \dots, \tau, \tau + 1$; and so forth until we use the entire sample $t = 1, \dots, T$. In the “backwards recursive estimate” we instead keep the end-point fixed and use more and more of old data. That is, the first sample could be $T - \tau, \dots, T$; the second $T - \tau - 1, \dots, T$; and so forth.

We could also apply an exponentially weighted moving average (EMA) estimator, which uses all data points since the beginning of the sample—but where recent observations carry larger weights. The weight for data in period t is λ^{T-t} where T is the latest

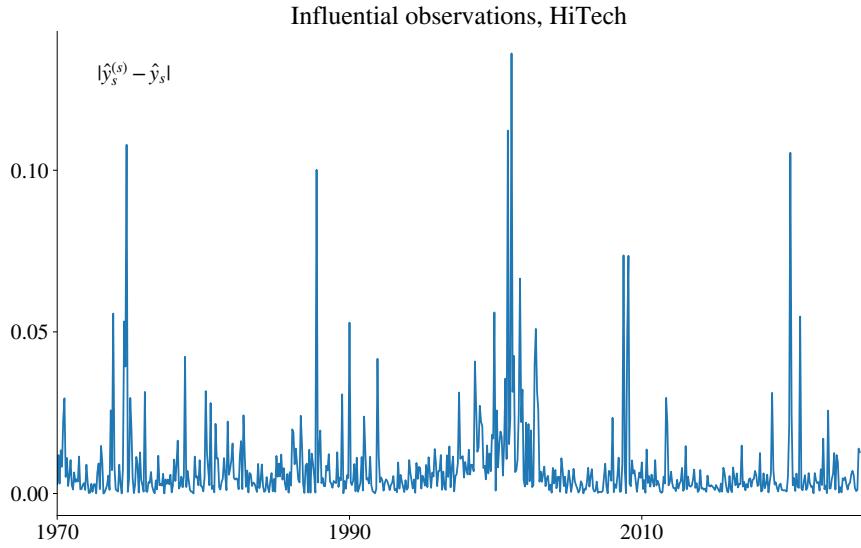


Figure 8.5: Influence of different observations

observation and $0 < \lambda < 1$, where a smaller value of λ means that old data carries low weights. In practice, this means that we define

$$\tilde{x}_t = x_t \lambda^{T-t} \text{ and } \tilde{y}_t = y_t \lambda^{T-t} \quad (8.21)$$

and then estimate

$$\tilde{y}_t = \tilde{x}'_t \beta + \varepsilon_t. \quad (8.22)$$

Notice that also the constant (in x_t) should be scaled in the same way. Again, see Figure 8.6 for an illustration.

Alternatively, a moving data window (“rolling samples”) could be used. In this case, the first sample is $t = 1, 2, \dots, \tau$; but the second is on $t = 2, \dots, \tau, \tau + 1$. This means that we drop one observation at the start of the sample and add one at the end. This approach could also be reversed: use all the data except that in the window (this is similar to the approach for detecting outliers, except that we exclude a window, not a single data point).

Empirical Example 8.12 (*Time-varying betas of the HiTech sector*) See Figure 8.6.

Empirical Example 8.13 (*Range of historical betas (different subsamples)*) See 8.7 for an illustration.

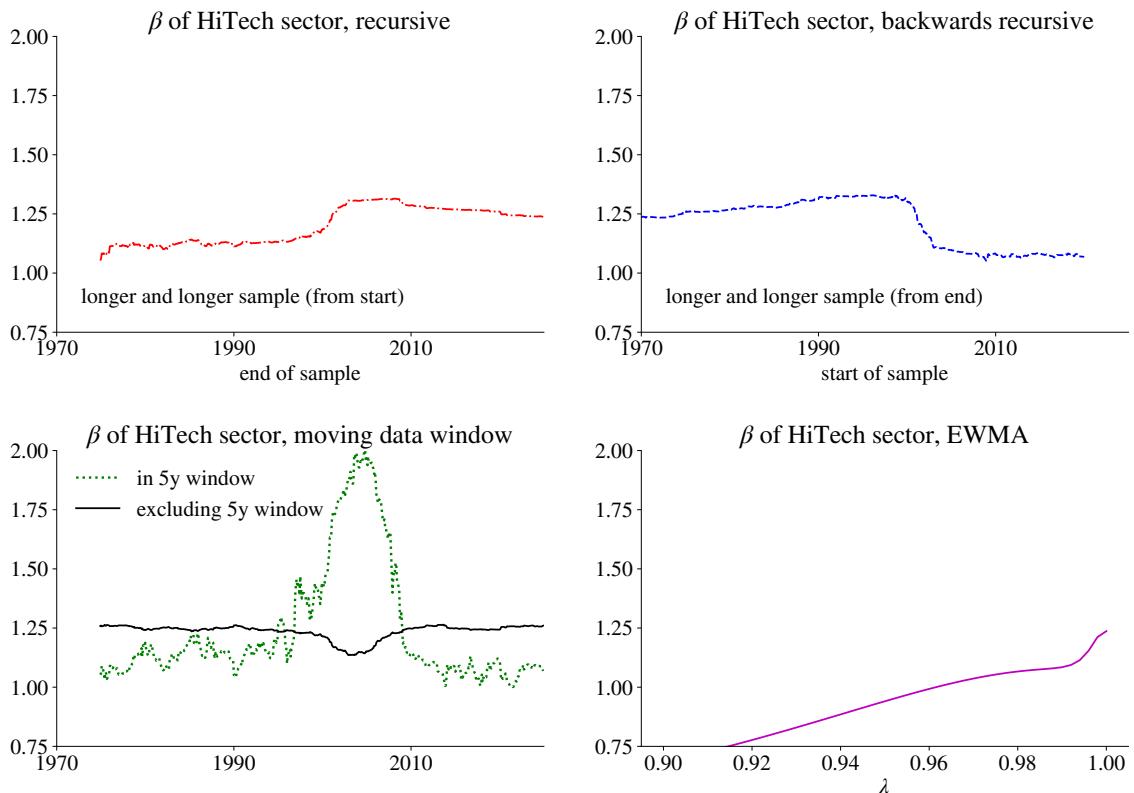


Figure 8.6: Betas of US industry portfolios

Estimation on subsamples is not only a way of getting a more recent/modern estimate, but also a way to gauge the historical range and volatility in the betas—which may be important for putting some discipline on judgemental forecasts.

From the estimations on subsamples (irrespective of method), it might be informative to study plots of (a) residuals with confidence band (0 ± 2 standard errors) or standardized residuals with confidence band (0 ± 2) and (b) coefficients with confidence band (± 2 standard errors). In these plots, the standard errors are typically from the subsamples.

A more formal way of testing for a one-time permanent change of a regression coefficient, also called a *structural break*, is to add a dummy for a subsample and interact it with the those regressors that we suspect have structural breaks (denoted z_t , which is a

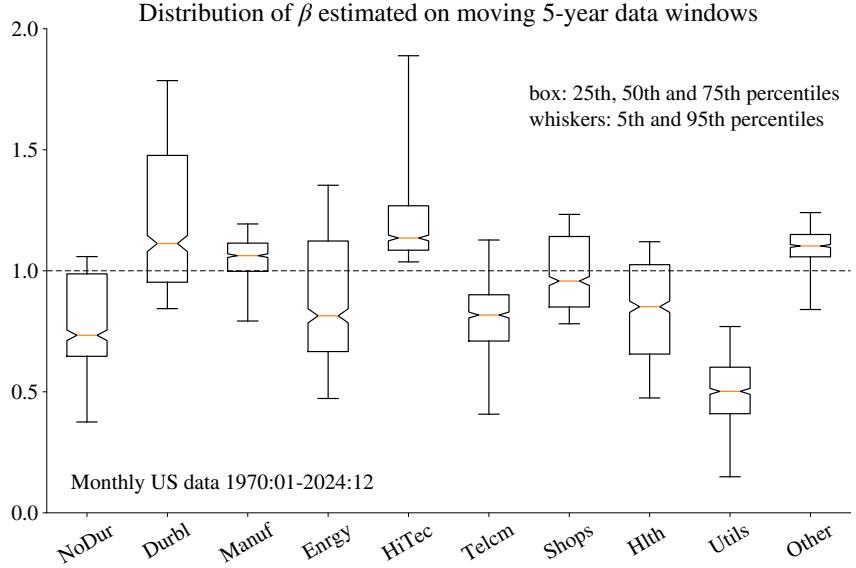


Figure 8.7: Distribution of betas of US industry portfolios (estimated on 5-year data windows)

subset of x_t)

$$y_t = x_t' \beta + g_t z_t' \delta + \varepsilon_t, \text{ where} \quad (8.23)$$

$$g_t = \begin{cases} 1 & \text{for some subsample} \\ 0 & \text{else} \end{cases} \quad (8.24)$$

and test $\delta = \mathbf{0}$ (this is a “Chow test”). Notice that δ measures the change of the coefficients from one sub sample to another, since the elements in z_t are also included in x_t . In principle, the subsample should be defined a priori and δ tested by a t-test (for a single coefficient) or a χ^2 -test (for several coefficients), but plotting the results for several subsamples is useful as a descriptive device. See Figure 8.8 for an illustration.

Remark 8.14 (CUSUM test*) The recursive CUSUM test (see, for instance, Enders (2004)) of parameter stability is as follows. First, consider a regression on the sample of observation 1 to $s - 1$ and use the estimated coefficients (denoted $\hat{\beta}_{s-1}$) to calculate a forecast and a scaled forecast error for s as

$$\hat{y}_s = x_s' \hat{\beta}_{s-1} \text{ and } v_s = (y_s - \hat{y}_s) / (1 + x_s' S_{xx,s-1}^{-1} x_s')^{1/2},$$

where $S_{xx,s-1} = \sum_{t=1}^{s-1} x_t x_t'$. Do this calculation for $s = k + 1$ (where k is the number of

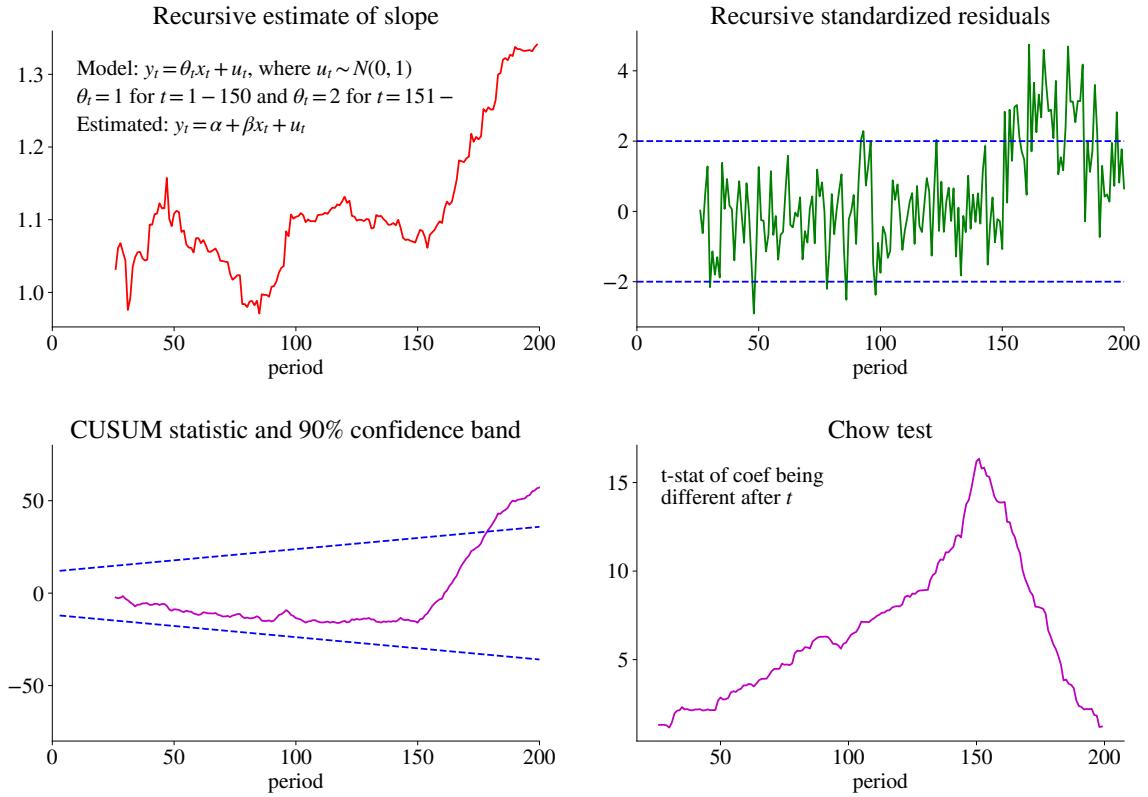


Figure 8.8: Example of a structural break

regressors), then for $s = k + 2$ etc. Second, estimate the standard deviation (denoted σ_v below) of the v_s series. Third, calculate the sequence of accumulated standardised errors

$$W_t = \sum_{s=k+1}^t v_s / \sigma_v, \text{ for } t = k + 1, \dots, T.$$

For instance, $W_{k+1} = v_{k+1} / \sigma_v$ and $W_{k+2} = (v_{k+1} + v_{k+2}) / \sigma_v$. Fourth, plot W_t (as a function of t) along with confidence interval: $\pm \lambda (\sqrt{T-k} + 2(t-k) / \sqrt{T-k})$, with $\lambda = 0.948$ for a 95% confidence band and 0.85 for a 90% confidence interval. Reject stability if any observation W_t is outside. See Figure 8.8 for an example.

Empirical Example 8.15 (CUSUM test of a CAPM regression) See Figure 8.9.

Further Reading

See Verbeek (2017) 3, Greene (2018) 4-5, Hansen (2022a) 3 for more details.

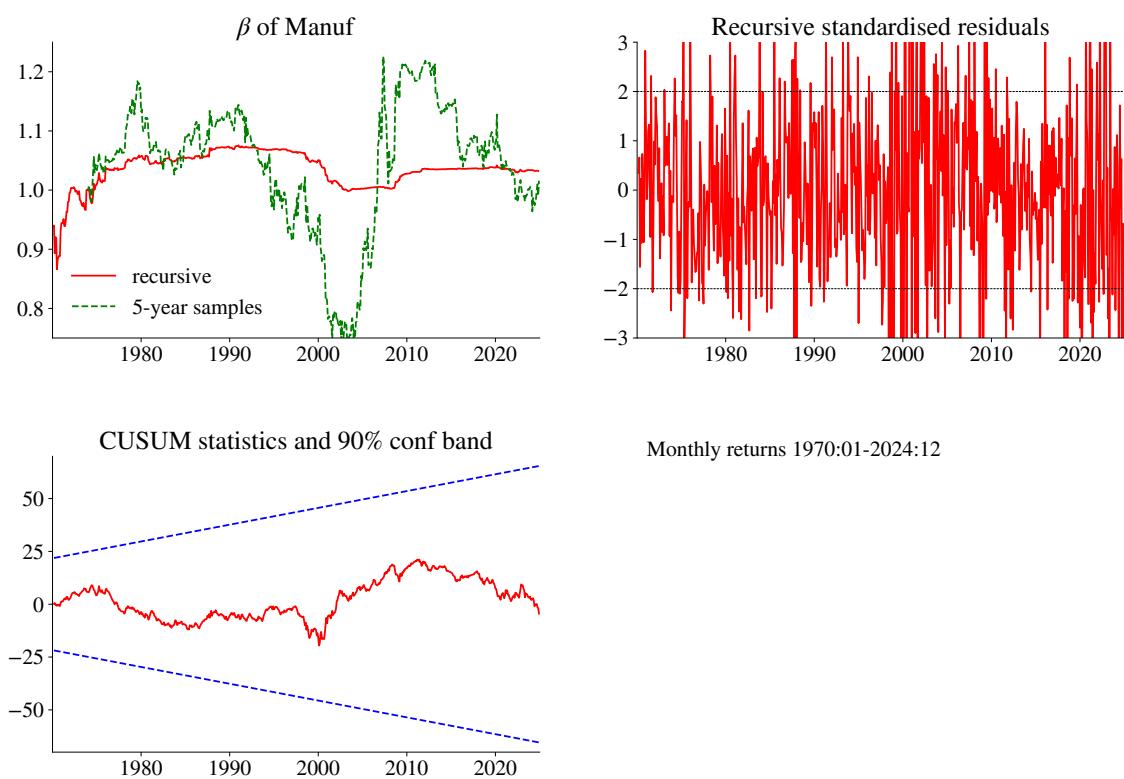


Figure 8.9: Stability test of a CAPM regression

Chapter 9

Asymptotic Results on OLS

9.1 Motivation of Asymptotics

There are several problems when the standard assumptions about linear regressions are wrong. First, unbiasedness ($E \hat{\beta} = \beta$) relies on the assumption that the regressors are fixed (or that we take expectations conditional on $\{x_1, \dots, x_T\}$). If this assumption does not hold, then OLS might be biased in finite samples, see Figure 9.1. Second, the result that $\hat{\beta}$ is normally distributed relies on the assumption that residuals are normally distributed. Otherwise it is not true (in a finite sample), see Figure 9.2.

There are two main approaches to addressing these issues: *(a)* conducting computer (Monte Carlo or bootstrap) simulations; *(b)* deriving results for $T \rightarrow \infty$ (“asymptotic properties”). The asymptotic properties are often more general than simulations, and often provide good approximations for large samples. However, they are less accurate in small samples, in which case simulations might prove more useful.

The asymptotic properties, which are the focus of this chapter, are based on the fact that most estimators are sample averages, and that sample averages often have nice large sample properties. In particular, the *law of large numbers* (LLN) and the *central limit theorem* (CLT) are powerful tools, see Figure 9.9 for an illustration.

9.2 Consistency

Reference: Greene (2018) 4.4; Hamilton (1994) 8.2; Davidson (2000) 3, Verbeek (2017) 2 and 5

This section discusses whether OLS comes closer to the true values as the sample size increases, so it is *consistent*. If not, we may have to consider other estimation methods, for instance, instrumental variables or maximum likelihood.

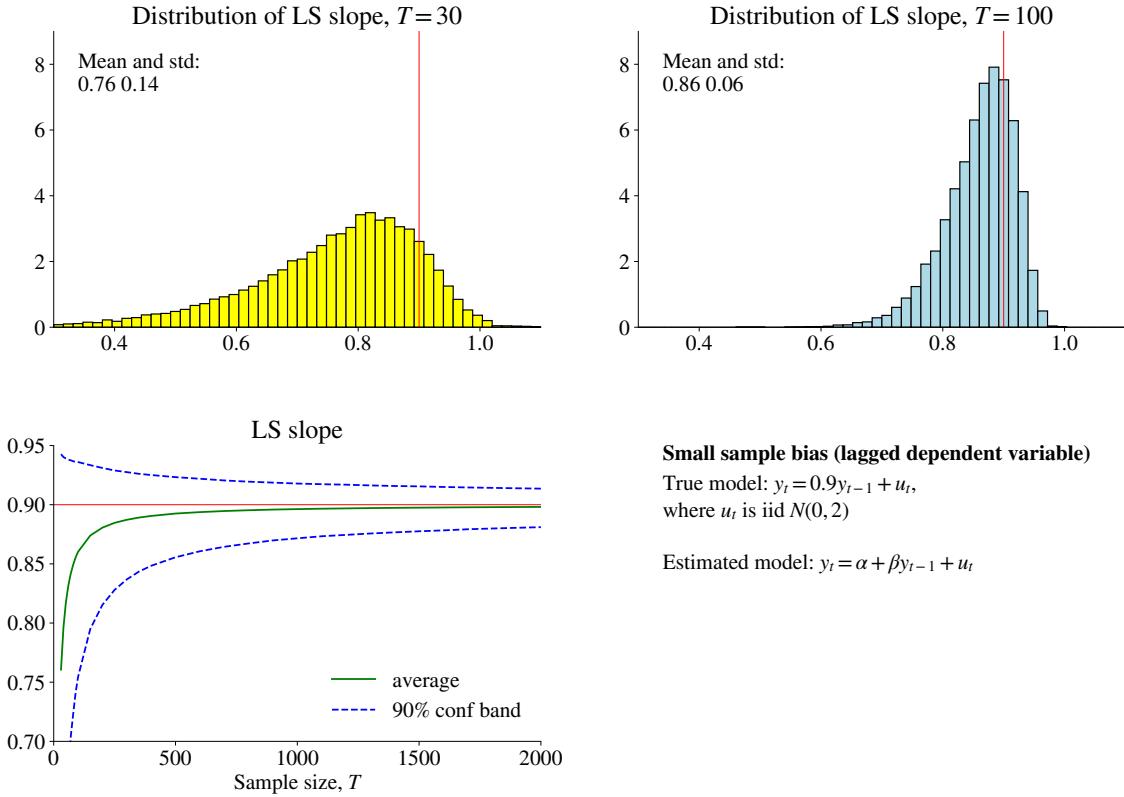


Figure 9.1: Distribution of LS estimator of autoregressive parameter

9.2.1 Probability Limits and the Law of Large Numbers

We need some basic facts about statistics (probability limits) for the discussion of consistency. These are summarized in the following remarks.

Remark 9.1 (*Convergence in probability*) \hat{b} converges in probability to b if (loosely speaking) $\hat{b} \rightarrow b$ as $T \rightarrow \infty$. Notation: $\text{plim } \hat{b} = b$ where plim stands for the probability limit. (Sometimes, $\hat{b} \rightarrow^p b$ is used.) In the expression, $\hat{b} \rightarrow b$ should be understood as that the probability that $|\hat{b} - b| < \text{any positive number}$ goes to 1.

Remark 9.2 (*Probability limits of a product and of a function*) If $\text{plim } \hat{a} = a$ and $\text{plim } \hat{b} = b$, then $\text{plim } \hat{a}\hat{b} = ab$. (In contrast, this does not hold for expectations: $E\hat{a}\hat{b} \neq E\hat{a}E\hat{b}$ unless \hat{a} and \hat{b} are uncorrelated.) More generally, Slutsky's theorem says that if $g()$ is a continuous function, then $\text{plim } g(\hat{a}) = g(\text{plim } \hat{a})$. For instance, $\text{plim } 1/\bar{x} = 1/\text{plim } \bar{x}$.

Remark 9.3 (*Law of large numbers, simple version*) A LLN says that the sample average converges to the population mean as the sample size increases (to infinity). Clearly, this

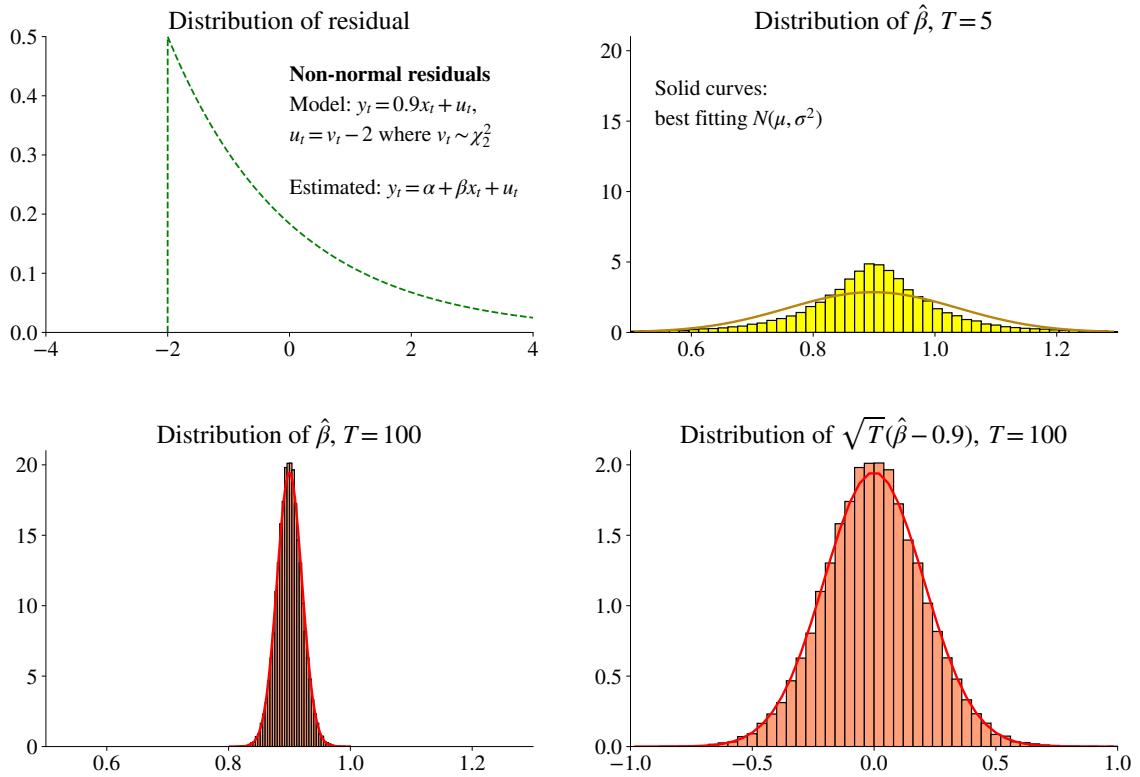


Figure 9.2: Results from a Monte Carlo experiment with thick-tailed errors.

means that the sample average is a consistent estimator of the population mean. Notation:
 $\text{plim}(\bar{x}) = E(x)$.

9.2.2 Consistency of OLS

Remark 9.4 (Consistency) Consistency means that the estimate $\hat{\beta}$ converges in probability to the true value as the sample size increases (to infinity).

The OLS estimate of a slope coefficient is (after dividing and multiplying by T)

$$\hat{\beta} = \beta + \underbrace{\left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)^{-1}}_{\rightarrow \Sigma_{xx}^{-1}} \underbrace{\frac{1}{T} \sum_{t=1}^T x_t u_t}_{\rightarrow E(xu)} \quad (9.1)$$

where u_t are the residuals we could calculate if we knew the true slope coefficient (denoted β), that is, the true residuals. The symbols below the equation indicate what the different terms converge to (according to a LLN) as the sample size increases. In particular, the

inverse is a continuous function so the first term converges to the inverse of the second moment matrix of x_t ($E x_t x_t'$) which is denoted Σ_{xx} . Notice that Σ_{xx} indicates a matrix, not a sum like $\Sigma_{t=1}^T$ does. This clearly assumes that x_t is such that the expectation is well defined, which we assume. Also, the two terms form a product so we can apply the rule that the probability limit is the product of the two (individual) probability limits.

In short, the probability limit is

$$\text{plim } \hat{\beta} = \beta + \Sigma_{xx}^{-1} E(x_t u_t), \quad (9.2)$$

where Σ_{xx}^{-1} is (asymptotically) a matrix of constants: there is nothing random about it. Clearly, for the estimate $\hat{\beta}$ to converge to the true values (β), it is necessary that $E(x_t u_t) = 0$. If $E u_t = 0$ (which is a basic assumption in most regression analysis), then $E(x_t u_t) = \text{Cov}(x_t, u_t)$, so consistency of $\hat{\beta}$ requires the regressors and the (true) residuals to be uncorrelated.

Some important observations for the following discussion. *First*, we can not (easily) test whether $E(x_t u_t) = 0$, since OLS constructs $\hat{\beta}$ and the fitted residuals \hat{u}_t in a way that ensures that the sample average of $x_t \hat{u}_t$ is zero.

Second, the standard regression assumption that u_t and x_t are independent implies $E(x_t u_t) = 0$, which in turn suggests that OLS is consistent. However, this assumption may not always hold.

Third, OLS can be biased (in a small sample), but still be consistent. This means that OLS could be systematically wrong in any small sample, but the problem vanishes in large samples. See Figure 9.1. In these figures, $\text{Cov}(u_{t-1}, x_t) \neq 0$ so OLS is biased since x_t is not independent of *all* residuals, but $\text{Cov}(u_t, x_t) = 0$ so it is consistent since x_t is not correlated with the *contemporaneous* residual.

Fourth, there are cases when $E(x_t u_t) = 0$ does *not* make sense. Then OLS is inconsistent, which will be discussed later in this chapter. See Figure 9.1 for an example of where OLS is consistent, and Figure 9.3 when it is not.

Fifth, we have so far assumed that the limits of $\Sigma_{t=1}^T x_t x_t'$ and $\Sigma_{t=1}^T x_t u_t$ are well defined and finite. This is not always the case, and something we also discuss later in this chapter (in the section on “spurious regressions”).

What should be done if OLS is (likely to be) inconsistent? If possible one should change the model or data. Otherwise consider alternative estimation methods (for instance, IV/2SLS, GMM, or MLE), which will be discussed in later chapters.

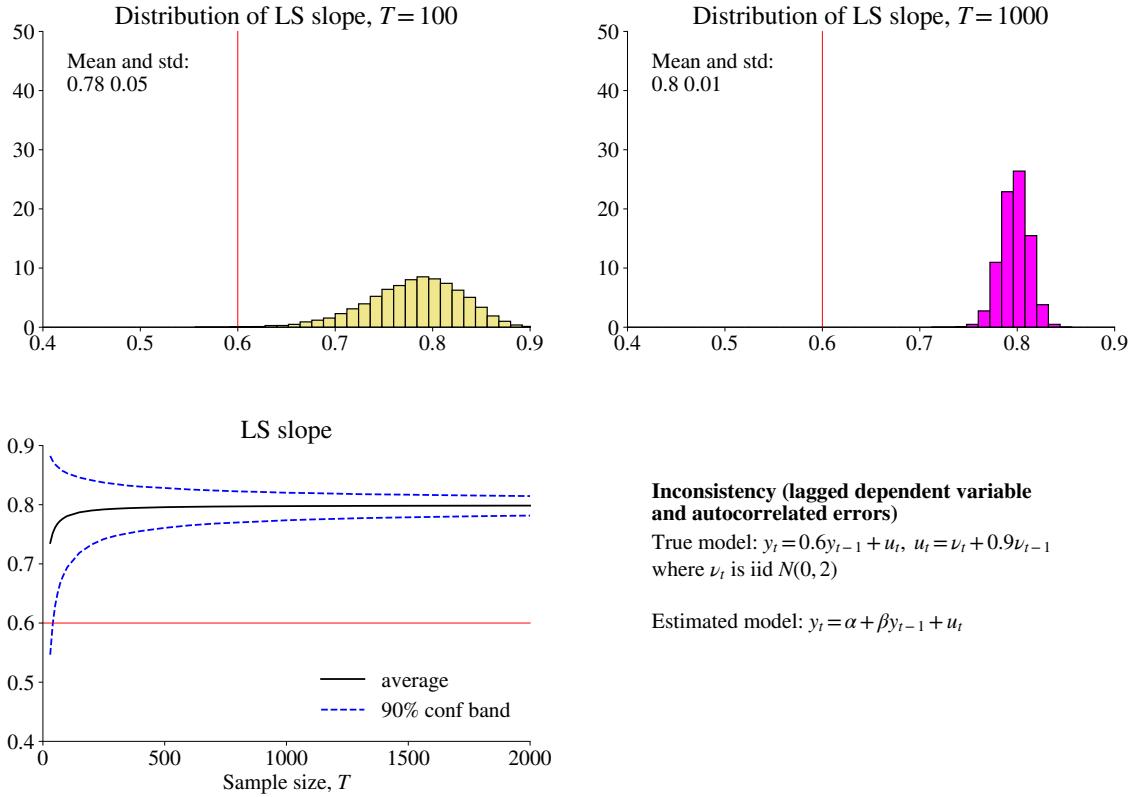


Figure 9.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

9.3 When OLS Is Inconsistent

This section discusses some typical cases where OLS is inconsistent, that is, $E(x_t, u_t) \neq 0$. These cases include (i) omitted variables; (ii) autocorrelated errors combined with lagged dependent variable; (iii) measurement errors in regressors; and (iv) endogenous regressors.

9.3.1 Omitted Variables

Reference: Greene (2018) 4.3

Omitted variables can make the coefficients on the included variables inconsistent. Suppose the true regression model is

$$y_t = x_t' \beta + h_t' \gamma + \varepsilon_t, \quad (9.3)$$

where $E(x_t \varepsilon_t) = 0$ and $E(h_t \varepsilon_t) = 0$.

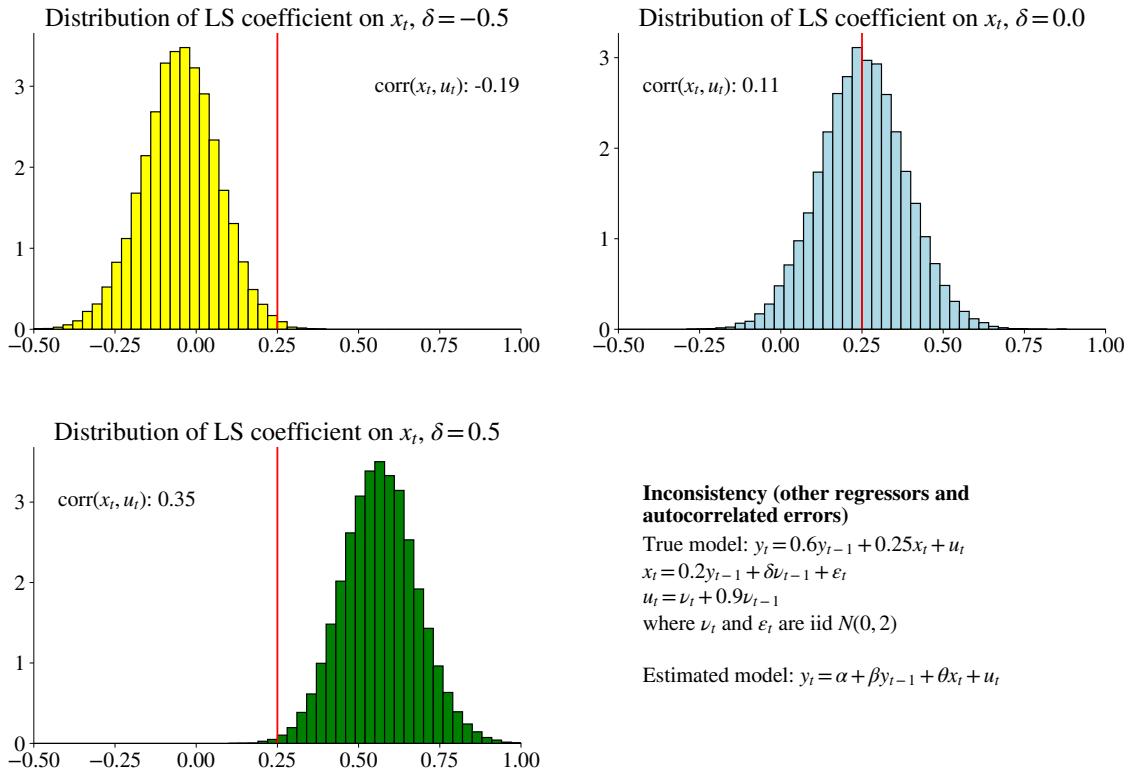


Figure 9.4: Results from a Monte Carlo experiment of LS estimation of the coefficient on x_t when the errors are autocorrelated.

Suppose we omit (exclude) the variables in h_t and instead estimate

$$y_t = x_t' b + u_t. \quad (9.4)$$

This means that the residual from in the regression (9.4) is $u_t = h_t' \gamma + \varepsilon_t$, that is, it incorporates the effect of both the omitted variables and the “true” residual, ε_t . Therefore, if the omitted variables are correlated with the included variables ($\text{cov}(h_t, x_t) \neq 0$), then $\text{cov}(x_t, u_t) \neq 0$ and $\text{plim } \hat{b}$ from (9.4) will differ from the true β in (9.3).

In fact, the probability limit of \hat{b} is

$$\text{plim } \hat{b} = \beta + [\theta_1 \dots \theta_L] \gamma, \quad (9.5)$$

where θ_i is the probability limit of a vector of coefficients obtained by regressing h_{it} on x_t . (See below for a proof.) This analysis shows that \hat{b} incorporates how x_t comoves with h_t . In case they are uncorrelated ($\theta_i = 0$), then omitting the h_t variables does not affect \hat{b} . However, if they are correlated, then \hat{b} are inconsistent (and biased) in the sense of being

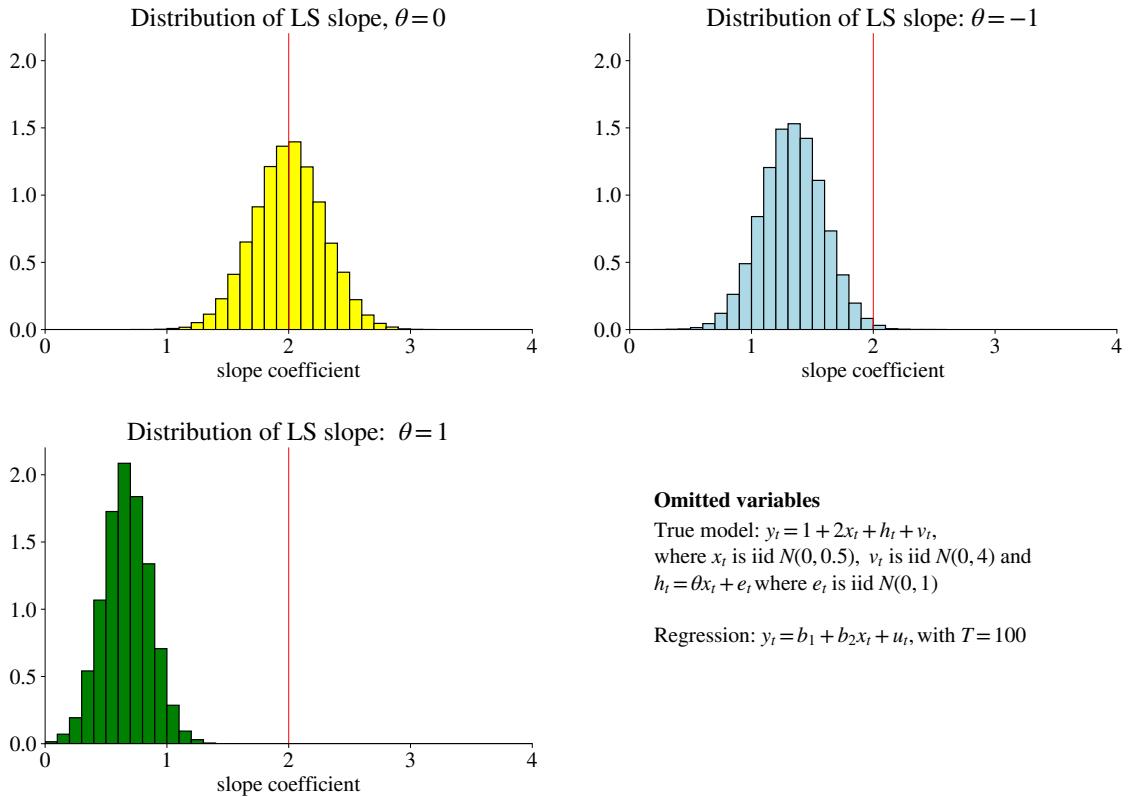


Figure 9.5: Effect of omitted variables

systematically different from β in (9.3). See Figure 9.5 for simulation results.

Actually, the relation in (9.5) holds in each sample, not just as a probability limit, if all coefficients estimated from that sample (a proof is found below).

However, notice the following about \hat{b} from (9.4). *First*, \hat{b} is actually the right number to use if we want to predict: “given x_t , what is the best guess of y_t ?” The reason is that \hat{b} factors in also how x_t predicts h_t (which affects y_t). *Second*, \hat{b} is not right number to use if we want to understand an economic mechanism: “if we increase x_{it} , by one unit (but holding all other variables constant), what is the likely effect on y_t ?” The reason is that we here need a consistent estimate of β . (Sometimes, economic theory is useful in guessing the sign of the bias.)

Proof (of (9.5)*) Recall that the OLS estimates are $\hat{b} = \beta + S_{xx}^{-1} \Sigma_{t=1}^T x_t u_t$, where $S_{xx} = \Sigma_{t=1}^T x_t x_t'$. Since $u_t = h_t' \gamma + \varepsilon_t$, we can write this as $\hat{b} = \beta + S_{xx}^{-1} \Sigma_{t=1}^T x_t h_t' \gamma + S_{xx}^{-1} \Sigma_{t=1}^T x_t \varepsilon_t$. The probability limit of the last term is zero (see (9.3)), while the middle term can be written $[\hat{\theta}_1 \dots \hat{\theta}_L] \gamma$ where $\hat{\theta}_i$ is the (column) vector of coefficients obtained

by regressing h_{it} on x_t , $\hat{\theta}_i = S_{xx}^{-1} \Sigma_{t=1}^T x_t h_{it}$. Take the probability limit and combine to get (9.5). \square

Proof (of that (9.5) hold in every sample, based on sample moments*) For notational convenience, let x_t and h_t be scalars and assume all variables have zero means so we can disregard intercepts. (1) Regress y_t on x_t to get $\hat{b} = S_{xx}^{-1} S_{xy}$. (2) Regress h_t on x_t to get $\hat{\theta} = S_{xx}^{-1} S_{xh}$. (3) Regress y_t on (x_t, h_t) (as in (9.3)) to get

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \frac{1}{S_{xx} S_{hh} - S_{xh}^2} \begin{bmatrix} S_{hh} & -S_{xh} \\ -S_{xh} & S_{xx} \end{bmatrix} \begin{bmatrix} S_{xy} \\ S_{hy} \end{bmatrix},$$

where we use the standard expression for the inverse of a 2×2 matrix. Combine to calculate $\hat{\beta} + \hat{\theta}\hat{\gamma}$ and note that it equals \hat{b} . \square

9.3.2 Autocorrelated Errors Combined with Lagged Dependent Variable

The combination of autocorrelated errors and having a lagged dependent variable as regressor can also cause inconsistency. As an example of how autocorrelated errors combined with a lagged dependent variable as regressor leads to inconsistent OLS estimates, consider

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t, \text{ where} \quad (9.6)$$

$$u_t = v_t + \theta v_{t-1}, v_t \text{ iid.} \quad (9.7)$$

As a special case, $\beta_2 = 0$ gives an ARMA(1,1) model. See Figure 9.3 for simulation results.

The issue is that y_{t-1} is correlated with the lagged shock (v_{t-1}) and hence with the OLS residuals u_t : $\text{Cov}(y_{t-1}, u_t) \neq 0$. This is a common problem in dynamic models.

Remark 9.5 (*Autocorrelated errors and other regressors**) *Autocorrelated errors may affect the coefficient estimates also on other regressors (x_t), but only if x_t is directly related to the residual. It is not enough that they are correlated with y_{t-1} . This might seem puzzling, since a correlation of a regressor with y_{t-1} will lead to a correlation with the residual, so $E(x_t u_t) \neq 0$. However, the bias in the coefficient on y_{t-1} will effectively create another residual that will be uncorrelated with x_t . (This could probably be proved by using the Frisch-Waugh theorem.) See Figure 9.4.*

Proof (of inconsistency of OLS estimate of an ARMA model*) Consider the case in (9.6)–(9.7) but where $\beta_2 = 0$ so the regression is an AR(1) but the errors follow an MA(1) process. In the limit, the OLS estimate is $\hat{\beta}_3 = \text{Cov}(y_t, y_{t-1}) / \text{Var}(y_{t-1})$. Using (9.6) to

replace y_t gives $\hat{\beta}_3 = \beta_3 + \text{Cov}(v_t + \theta v_{t-1}, y_{t-1}) / \text{Var}(y_{t-1})$. The 2nd term is non-zero if $\theta \neq 0$. \square

9.3.3 Measurement Errors in a Regressor

A measurement error in a regressor can make OLS inconsistent, whereas a measurement error in the dependent variable typically does not (it is just larger residuals). As an example, consider a true regression model like

$$y_t = \beta_1 + \beta_2 w_t + v_t. \quad (9.8)$$

However, we estimate with a proxy x_t for the regressor w_t

$$y_t = b_1 + b_2 x_t + u_t, \text{ with} \quad (9.9)$$

$$x_t = w_t + e_t, \quad (9.10)$$

where e_t is a measurement error. This is a common problem in micro data, including corporate finance. In this case, $u_t = -\beta_2 e_t + v_t$ (solve for $w_t = x_t - e_t$, use in (9.8)), so $\text{Cov}(x_t, u_t) \neq 0$ since x_t depends on the measurement error e_t directly, see (9.10). Therefore, OLS is inconsistent for estimating β_2 . See Figure 9.6 for simulation results.

In fact,

$$\text{plim } \hat{b}_2 = \beta_2 \left(1 - \sigma_e^2 / (\sigma_w^2 + \sigma_e^2)\right). \quad (9.11)$$

(See below for a proof.) Notice that $\hat{b}_2 \rightarrow 0$ if the measurement error dominates ($\sigma_e^2 \rightarrow \infty$), since y_t is not related to the measurement error. In contrast, $\hat{b}_2 \rightarrow \beta_2$ as measurement vanishes ($\sigma_e^2 \rightarrow 0$): no measurement error. Measurement errors will thus bias the coefficient towards zero. Any significant coefficient can therefore be seen as a conservative estimate.

Proof (of (9.11)) To simplify, assume that x_t has a zero mean. From (9.2), we then have $\text{plim } \hat{b}_2 = \beta_2 + \Sigma_{xx}^{-1} \text{E}(x_t u_t)$. Here, $\Sigma_{xx}^{-1} = 1 / \text{Var}(x_t)$, but notice from (9.10) that $\text{Var}(x_t) = \sigma_w^2 + \sigma_e^2$ if w_t and e_t are uncorrelated. We also have $\text{E}(x_t u_t) = \text{Cov}(x_t, u_t)$, which from the definition of x_t in (9.10) and of $u_t = -\beta_2 e_t + v_t$ gives $\text{Cov}(x_t, u_t) = \text{Cov}(w_t + e_t, -\beta_2 e_t + v_t) = -\beta_2 \sigma_e^2$. Together we get $\text{plim } \hat{b}_2 = \beta_2 + \Sigma_{xx}^{-1} \text{E}(x_t u_t) = \beta_2 - \beta_2 \sigma_e^2 / (\sigma_w^2 + \sigma_e^2)$, which is (9.11). \square

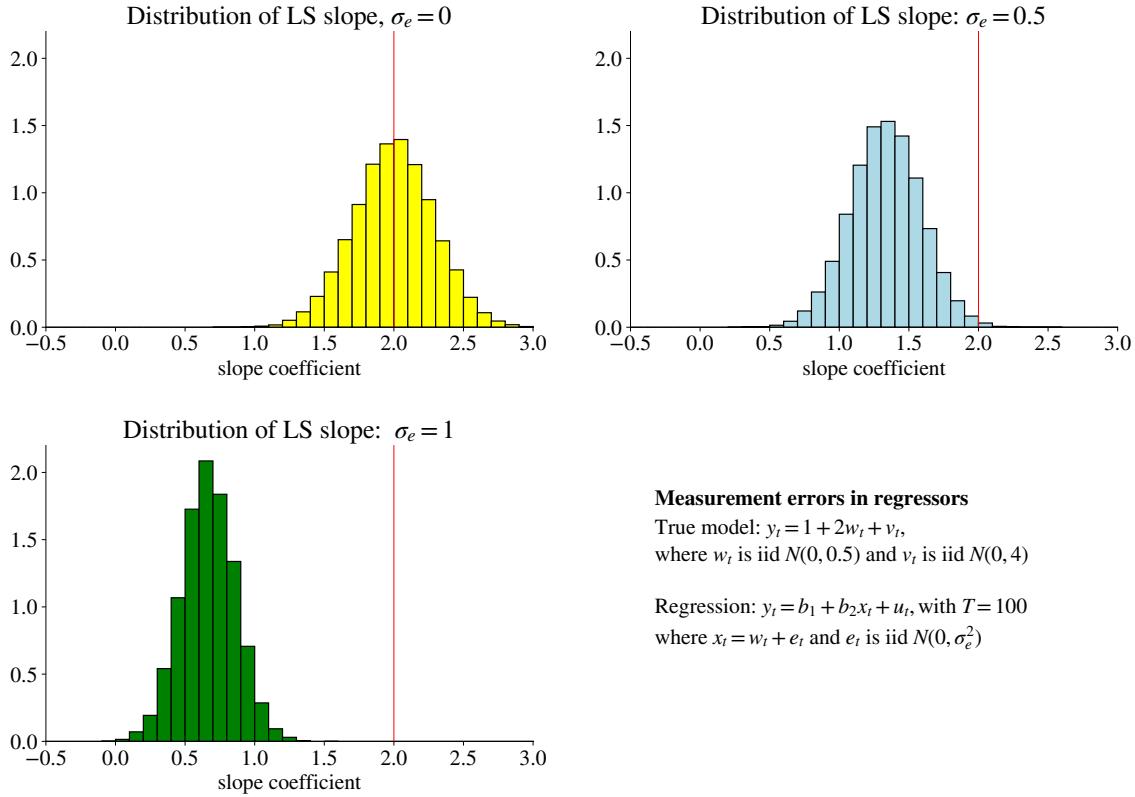


Figure 9.6: Effect of measurement error in regressor

9.3.4 Endogenous Regressors (System of Simultaneous Equations)

When the regressor is partially driven by the dependent variable (endogenous regressor, reverse causality), then we again have problems with the consistency of OLS. As an example, consider a simple model for supply and demand on a market. Supply is

$$q_t = \gamma p_t + \varepsilon_{st}, \quad \gamma > 0, \quad (9.12)$$

and demand is

$$q_t = \beta p_t + \alpha A_t + \varepsilon_{dt}, \quad \beta < 0, \quad (9.13)$$

where A_t is an observable demand shock (perhaps income).

Suppose we try to estimate the supply equation (9.12) by LS

$$q_t = bp_t + u_t,$$

However, p_t is correlated with u_t (since $\varepsilon_{st} \rightarrow q_t \rightarrow p_t$), implying that $\text{cov}(p_t, u_t) \neq 0$,

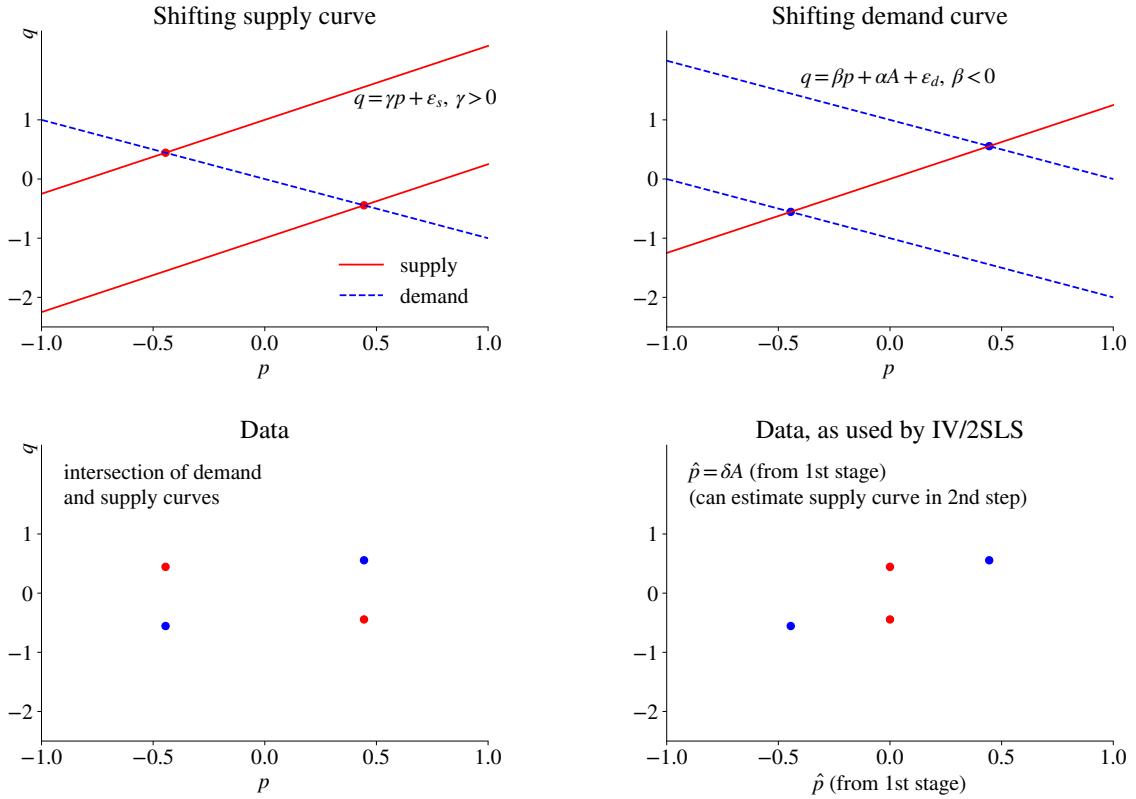


Figure 9.7: Illustration of demand and supply curves

so we cannot hope that LS will be consistent, see Appendix 9.6 for details. Figure 9.7 provides an illustration (disregard the IV/2SLS subfigure for now).

It is clear that the OLS estimate \hat{b} will be a mixture of the true γ and β values (and other things). Figure 9.8 illustrates this by a Monte Carlo simulation.

9.4 Asymptotic Normality

Reference: Greene (2018) 4.4; Hamilton (1994) 8.2; Davidson (2000) 3

This section discusses the distribution of OLS in large samples.

9.4.1 Central Limit Theorems

We need some basic facts about the central limit theorem (CLT) for the discussion of asymptotic normality, summarized in the remarks below.

Remark 9.6 (*Convergence in distribution*) Let \hat{z} be a random variable which depends on

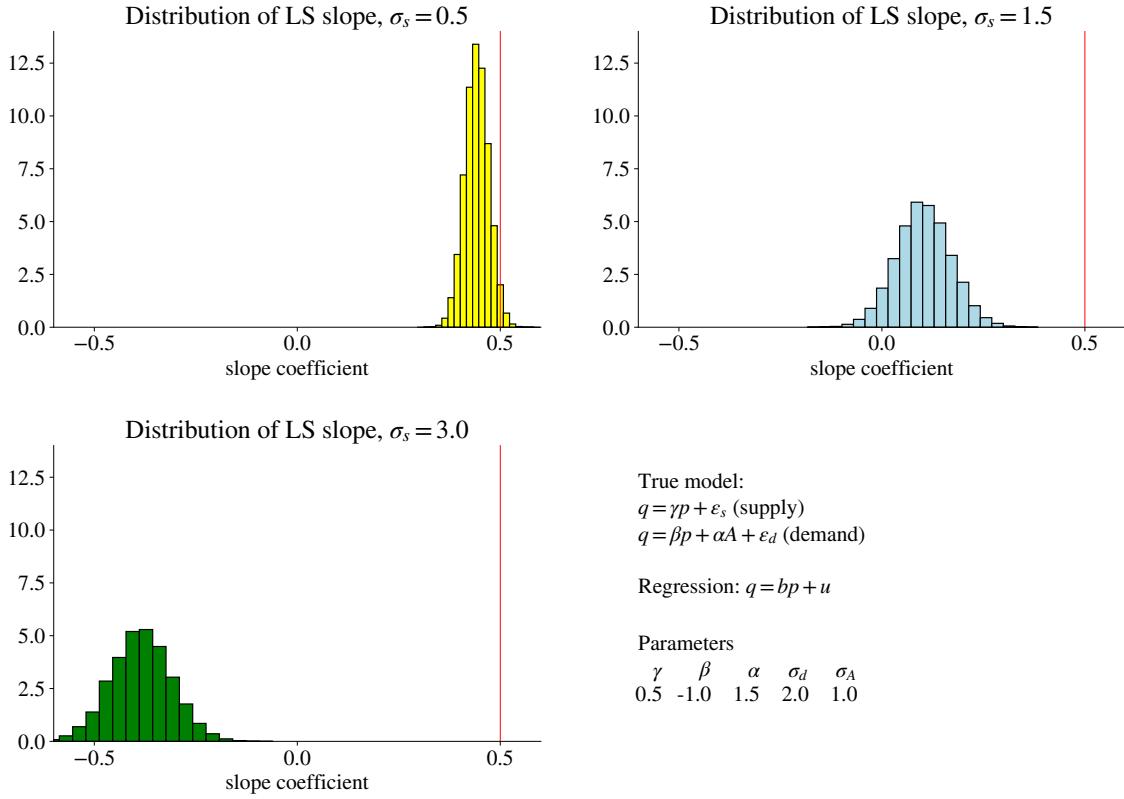


Figure 9.8: Distribution of LS estimator of supply elasticity

the sample size T and let Z be another random variable that does not. If the cdf of \hat{z} is the same as the cdf of z as $T \rightarrow \infty$, then \hat{z} converges in distribution to the random variable Z . Notation $\hat{z} \xrightarrow{d} Z$.

Remark 9.7 (Central limit theorem, simple version) A CLT says that $\sqrt{T}\bar{x} \xrightarrow{d} N()$, that is, becomes normally distributed as T goes to infinity. This holds for many random variables, although exceptions exist. Notice that the distribution of \bar{x} converges to a spike as T increases (this is a LLN), but the distribution of $\sqrt{T}\bar{x}$ converges to a normal distribution. See Figure 9.9.

Remark 9.8 (Continuous mapping theorem.) Let the random variables \hat{z} and \hat{q} and the non-random a_T be such that $\hat{z} \xrightarrow{d} Z$, $\text{plim } \hat{q} = Q$ (a finite and positive definite matrix) and $a_T \rightarrow a$ (a traditional limit). Also, let $g(z, y, a)$ be a continuous function. Then $g(\hat{z}, \hat{q}, a_T) \xrightarrow{d} g(Z, Q, a)$.

Example 9.9 For instance, the sequences in Remark 9.8 could be $\hat{z} = \sqrt{T} \sum_{t=1}^T x_t / T$

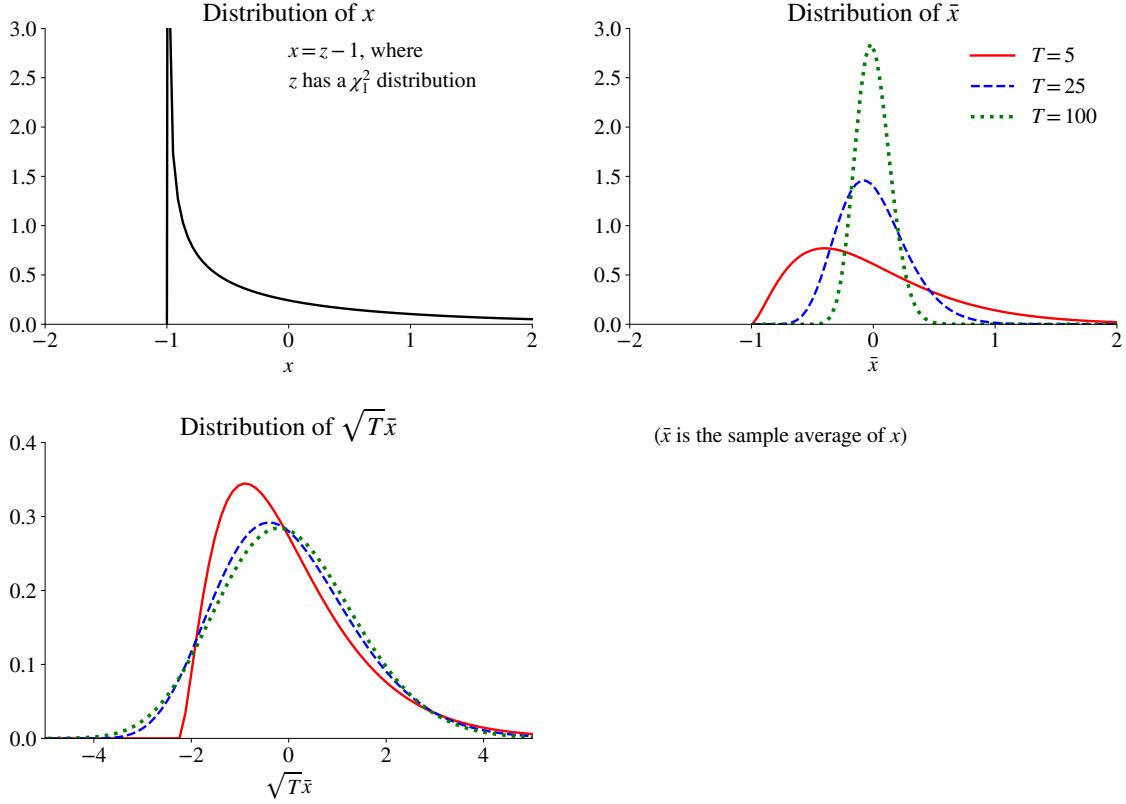


Figure 9.9: Distribution of sample averages

(the scaled sample average of an iid zero mean random variable x_t) which obeys a CLT; $\hat{q} = (\sum_{t=1}^T x_t^2 / T)^{1/2}$ (the sample standard deviation); and $a_T = \sum_{t=1}^T 0.7^t$ (which converges to $2\frac{1}{3}$). Then, $a_T \hat{\beta} / \hat{q}$ would converge to a $N(0, 5\frac{4}{9})$ variable. (To see this, notice that with iid data, $\text{Std}(\bar{x}) = \sigma / \sqrt{T}$, where σ is the standard deviation of x , so $\text{Std}(\sqrt{T}\bar{x}) = \sigma$.)

9.4.2 Asymptotic Normality of OLS

Subtract β from both sides of (9.1), and multiply both sides by \sqrt{T} to get

$$\sqrt{T}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)^{-1}}_{\rightarrow \Sigma_{xx}^{-1}} \underbrace{\sqrt{T} \frac{1}{T} \sum_{t=1}^T x_t u_t}_{\sqrt{T} \times \text{sample average}}, \quad (9.14)$$

where Σ_{xx} is the probability limit of $\sum_{t=1}^T x_t x_t' / T$. (Again, notice that Σ_{xx} denotes a matrix, not a sum like $\sum_{t=1}^T$ does.) The first term converges (by a LLN) to the matrix

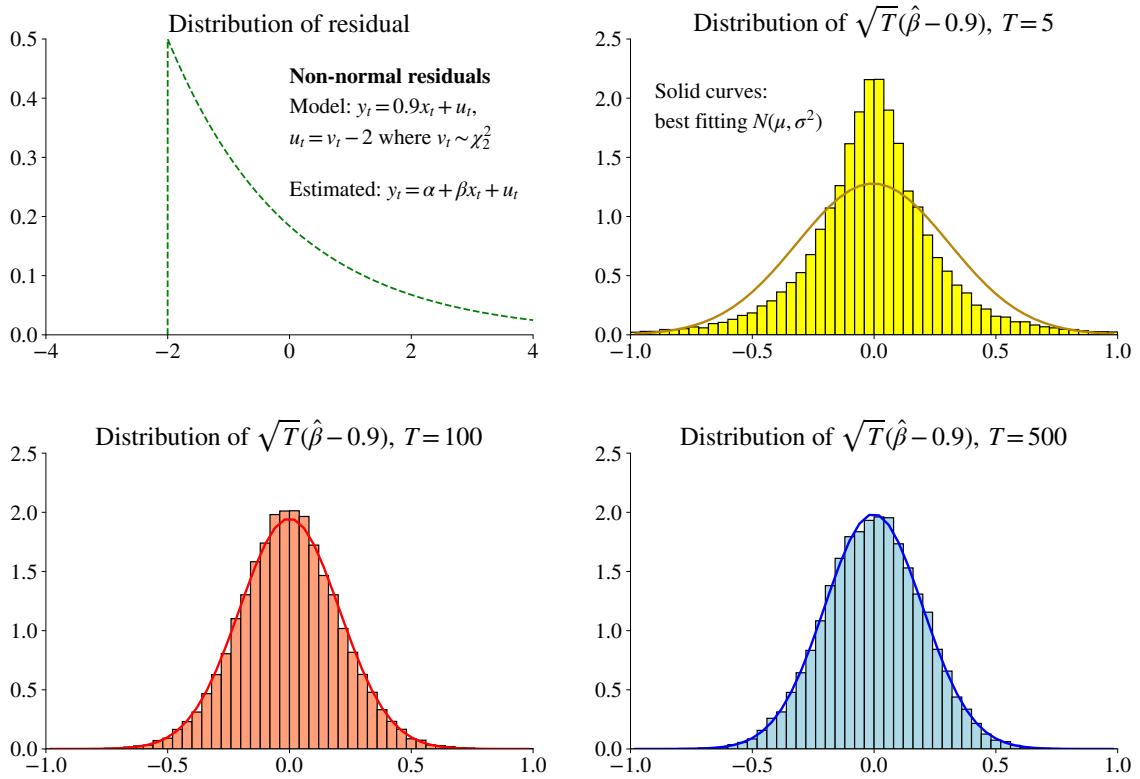


Figure 9.10: Results from a Monte Carlo experiment with thick-tailed errors.

of constants, Σ_{xx}^{-1} , assuming the x_t is such that the limit is well defined. (This is the $\text{plim } \hat{q} = Q$ part in Remark 9.8.) We will later discuss the case when this assumption is wrong. The second term is like $\sqrt{T} \times \text{sample average (of } x_t u_t)$, which (by a CLT) will (typically) converge in distribution to a normally distributed variable. (This is the $\hat{z} \xrightarrow{d} Z$ part in Remark 9.8.) According to Remarks 9.7 and 9.8, we should therefore expect $\sqrt{T} \hat{\beta}$ to be normally distributed in *large* samples—even if the residual doesn’t have a normal distribution. See Figure 9.10 for an example.

According to Remark 9.8 and (9.14) $\sqrt{T}(\hat{\beta} - \beta)$ converges in distribution to Σ_{xx}^{-1} times a normally distributed variable (vector). If OLS is consistent, then the normal distribution has a zero mean ($E x_t u_t = 0$). Let Σ denote the variance-covariance matrix of the last term in (9.14)

$$\Sigma = \text{Var} \left(\sqrt{T} \sum_{t=1}^T x_t u_t / T \right), \quad (9.15)$$

which is the same as \sqrt{T} times the sample average of $x_t u_t$. Together, we then have

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma_{xx}^{-1} \Sigma \Sigma_{xx}^{-1}) \quad (9.16)$$

We sometimes rewrite this as

$$\hat{\beta} \stackrel{a}{\sim} N(\beta, \Sigma_{xx}^{-1} \Sigma \Sigma_{xx}^{-1} / T), \quad (9.17)$$

where $\stackrel{a}{\sim}$ means “is asymptotically distributed as.” (The transformation to get this is (a) divide the LHS of (9.16) by \sqrt{T} and thus the covariance matrix by T ; (b) add β to the LHS and thus shift the mean from 0 to β .)

If we estimate by $\hat{\Sigma}_{xx} = S_{xx}/T$ where $S_{xx} = \sum_{t=1}^T x_t x'_t$ as in previous chapters and $\hat{\Sigma} = \hat{S}/T$ where \hat{S} is an estimate of the variance-covariance matrix of $\sum_{t=1}^T x_t u_t$ as in previous chapters, then after cancelling T terms, an estimate of the variance-covariance term in (9.17) is

$$\widehat{\text{Var}}(\hat{\beta}) = \left(\frac{S_{xx}}{T}\right)^{-1} \frac{\hat{S}}{T} \left(\frac{S_{xx}}{T}\right)^{-1} \frac{1}{T} = S_{xx}^{-1} \hat{S} S_{xx}^{-1}. \quad (9.18)$$

Notice that this has the same form as an estimate of the variance-covariance matrix derived under the assumption of fixed regressors. Clearly, how to estimate S depends on whether there is heteroskedasticity (White’s method, perhaps) and/or autocorrelation (Newey-West’s method, perhaps).

9.5 Spurious Regressions

This section contains an informal discussion of what happens when the data has trends and/or unit root behaviour. Loosely speaking, a unit root process is such that the effect of a shock never vanishes, like for a random walk. This means, for instance, that the limits of $\sum_{t=1}^T x_t x'_t$ and $\sum_{t=1}^T x_t u_t$, previously used in (9.1)–(9.2), may not be well defined. Rather, they may be infinite or random variables also in the limit. This can lead to very strange properties for OLS. A precise formal discussion of this is rather involved and beyond the scope of these notes, so the focus in this section is on providing some intuition.

Example 9.10 (*Random walk and $\sum_{t=1}^T x_t^2 / T$) If $x_t = x_{t-1} + \varepsilon_t$, where ε_s is iid with zero mean and variance σ^2 , then $E x_t^2 = t\sigma^2$. We thus get $E \sum_{t=1}^T x_t^2 / T = \sum_{t=1}^T t\sigma^2 / T = \sigma^2(T+1)/2$, since $\sum_{t=1}^T t = T(T+1)/2$. This goes to infinity as T does. This property is shared by other unit root processes.

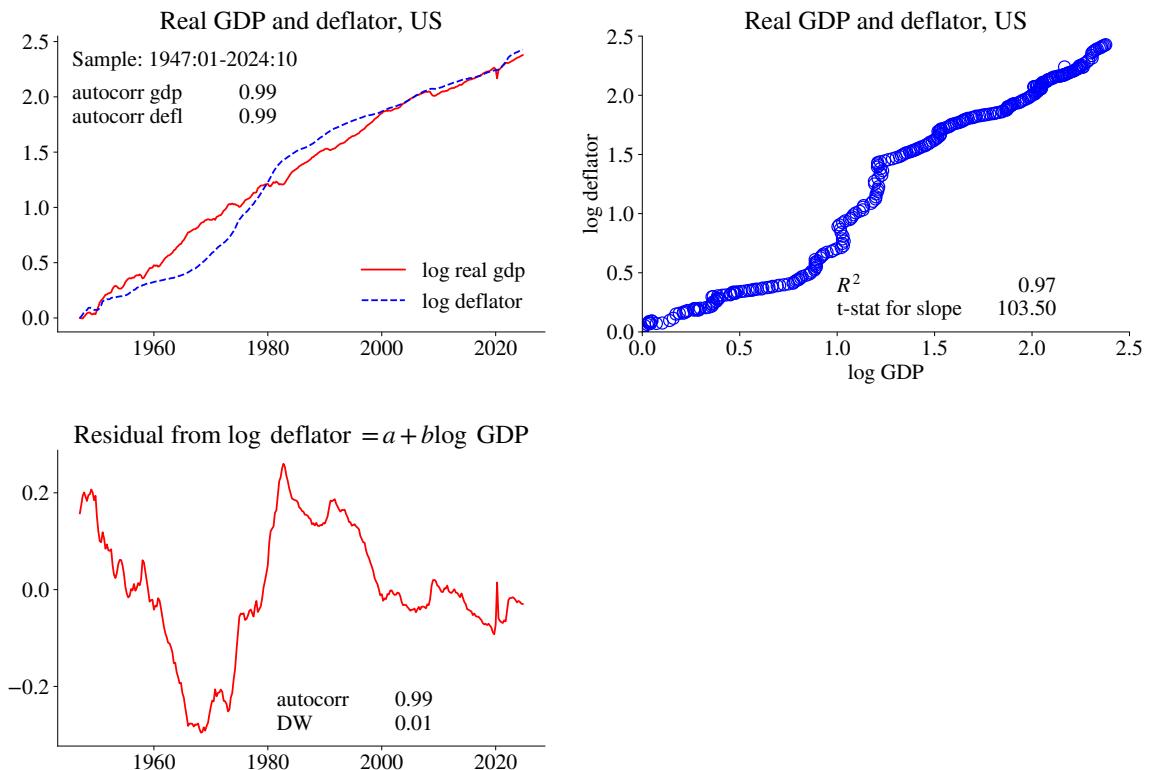


Figure 9.11: Example of a spurious regression

When y_t and x_t both have trending behaviour, but where an explicit trend is missing from the set of regressors, then a regressor may look significant just because it is a proxy for the missing trend (a missing variable bias). The same holds for non-stationary processes, even if they have no deterministic trends. The reason is that the innovations (shocks) accumulate and the series therefore tend to be trending in small samples. Asymptotic results are typically of *little use* here, since the non-stationarity means that the asymptotic results are degenerate (for instance, infinite variance). A warning sign of a spurious regression is when $R^2 > DW$ (Durbin-Watson) statistic.

Empirical Example 9.11 (*Regressing the price level on GDP*) See Figure 9.11 for results from regressing the U.S. price level (GDP deflator) on output (GDP level). The results indicate a very significant regression slope, but extreme autocorrelation. This is likely to be a spurious regression. Also, economics would suggest that nominal (price level) and real variables (output) variables are driven by completely different factors.

Figure 9.12 illustrates a Monte Carlo simulation, which suggests that (with a unit root)

a larger sample size might not reduce the uncertainty, here indicated by confidence bands. Figure 9.13 shows how the distribution of the t-stat (of a true null hypothesis) has a very wide distribution—far from being $N(0, 1)$. This holds also when Newey-West standard errors are used. Finally, Figure 9.14 show that the R^2 gets a strange distribution and that the autocorrelation of the residuals is typically very high.

For trend-stationary data, this problem is easily solved by detrending with a linear trend (before estimating or just adding a trend variable to the set of regressors).

However, this is usually a poor method for a unit root process. A first difference is needed instead. For instance, a first difference of the random walk with drift is

$$\begin{aligned}\Delta y_t &= y_t - y_{t-1} \\ &= \mu + \varepsilon_t,\end{aligned}\tag{9.19}$$

which is white noise (any finite difference, like $y_t - y_{t-s}$, will give a stationary series), so we could proceed by applying standard econometric tools to Δy_t . See Figure 9.15 for an illustration.

Further Reading

See Verbeek (2017) 2, Greene (2018) 4, Hansen (2022a) 6-7 for more details.

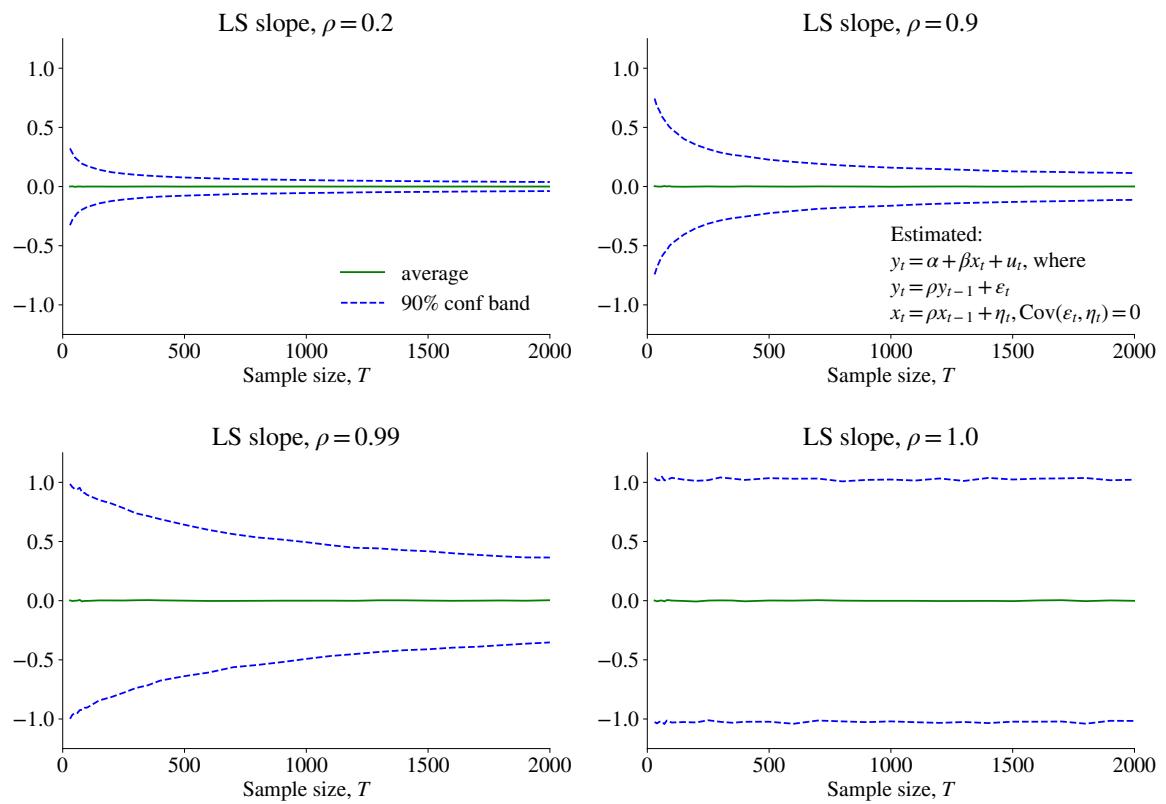


Figure 9.12: Distribution of slope coefficient when y_t and x_t are independent AR(1) processes

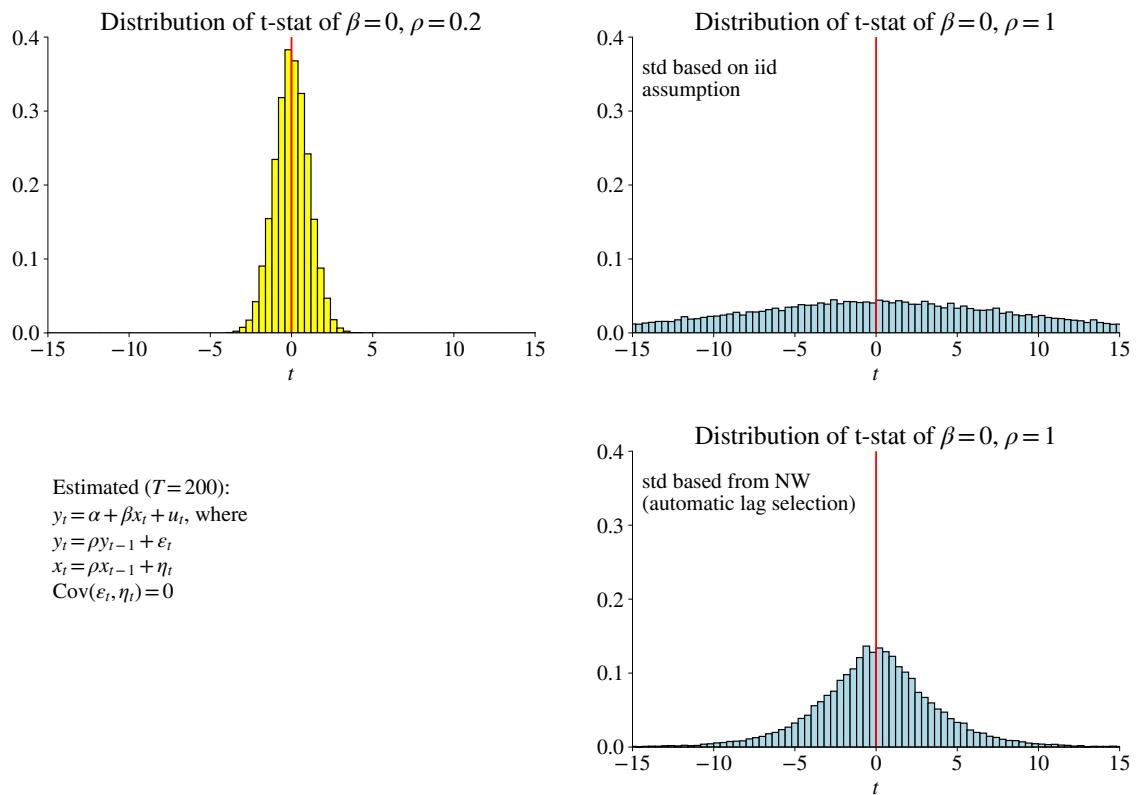


Figure 9.13: Distribution of the t-statistic when y_t and x_t are independent AR(1) processes. See Figure 9.12.

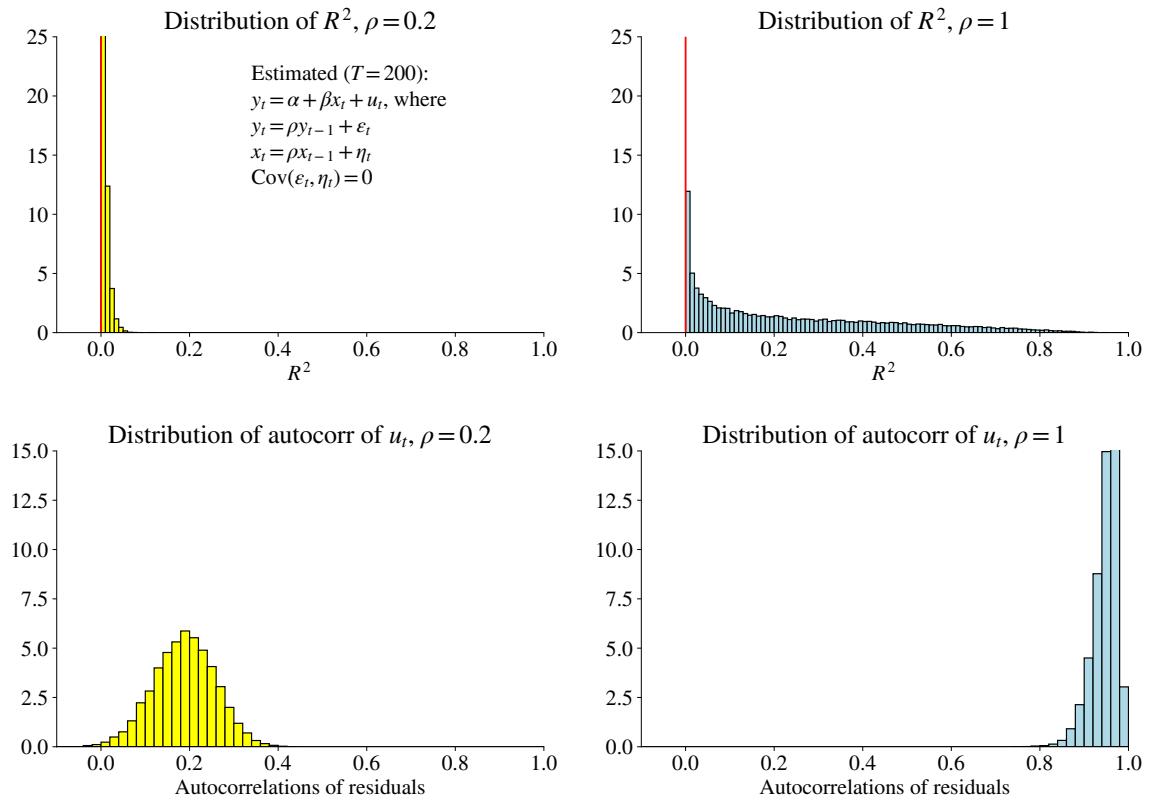


Figure 9.14: Distribution of R^2 and autocorrelation of residuals. See Figure 9.12.

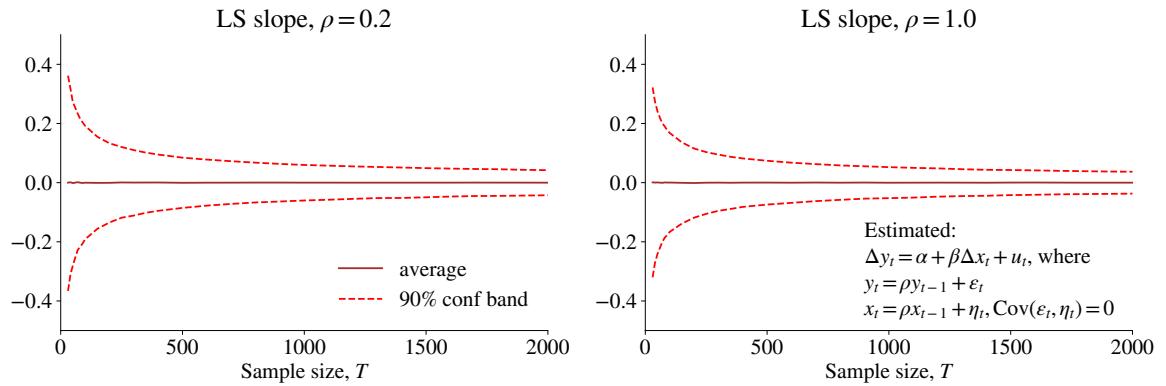


Figure 9.15: Distribution of slope coefficient in regression using first differences.

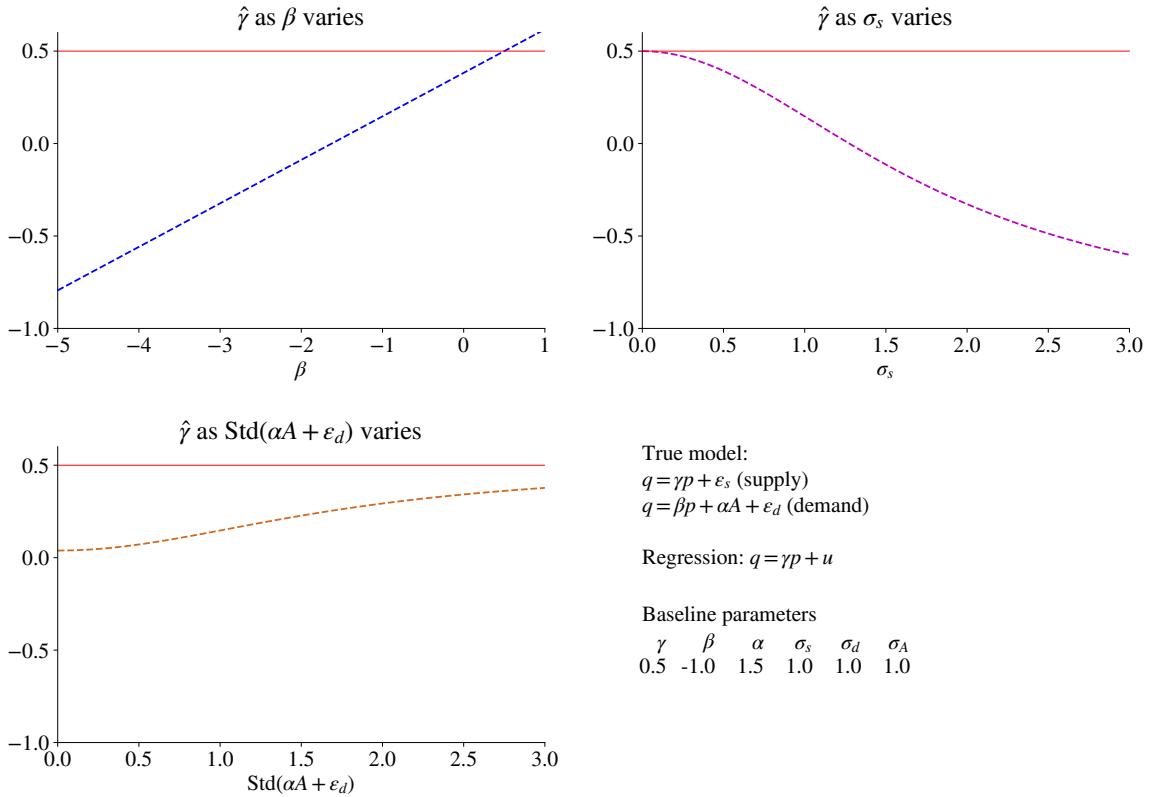


Figure 9.16: OLS estimate of γ in supply equation

9.6 Appendix – Details on Demand and Supply*

This appendix includes some details that are left out from the main text.

Example 9.12 (*Supply and Demand**) The system (the “structural form”) is therefore

$$\begin{bmatrix} 1 & -\gamma \\ 1 & -\beta \end{bmatrix} \begin{bmatrix} q \\ p \end{bmatrix} + \begin{bmatrix} 0 \\ -\alpha \end{bmatrix} A = \begin{bmatrix} \varepsilon_s \\ \varepsilon_d \end{bmatrix}.$$

This can be solved in terms of the exogenous variables (the “reduced form”) as

$$\begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} -\frac{\gamma}{\beta-\gamma}\alpha \\ -\frac{1}{\beta-\gamma}\alpha \end{bmatrix} A + \begin{bmatrix} \frac{\beta}{\beta-\gamma} & -\frac{\gamma}{\beta-\gamma} \\ \frac{1}{\beta-\gamma} & -\frac{1}{\beta-\gamma} \end{bmatrix} \begin{bmatrix} \varepsilon_s \\ \varepsilon_d \end{bmatrix}.$$

Example 9.13 (*Supply equation with LS**) Using the reduced form from Example 9.12, it is straightforward to show that the probability limit of the OLS estimate of γ is (assuming

(that the supply and demand shocks are uncorrelated)

$$\begin{aligned}\text{plim } \hat{\gamma}_{OLS} &= \frac{\text{Cov}(q, p)}{\text{Var}(p)} \\ &= \frac{\gamma\alpha^2 \text{Var}(A) + \gamma \text{Var}(\varepsilon_d) + \beta \text{Var}(\varepsilon_s)}{\alpha^2 \text{Var}(A) + \text{Var}(\varepsilon_d) + \text{Var}(\varepsilon_s)}.\end{aligned}$$

First, suppose the supply shocks are zero, $\text{Var}(\varepsilon_s) = 0$, then $\text{plim } \hat{\gamma} = \gamma$, so we indeed estimate the supply elasticity, as we wanted. Think of a fixed supply curve, and a demand curve which moves around. These point of p and q should trace out the supply curve. It is clearly ε_s that causes a simultaneous equations problem in estimating the supply curve: ε_s affects both q and p and the latter is the regressor in the supply equation. With no movements in ε_s there is no correlation between the shock and the regressor. Second, now suppose instead that the both demand shocks are zero (both $A = 0$ and $\text{Var}(\varepsilon_d) = 0$). Then $\text{plim } \hat{\gamma} = \beta$, so the estimated value is not the supply, but the demand elasticity. Not good. This time, think of a fixed demand curve, and a supply curve which moves around.

Chapter 10

Simulating the Finite Sample Properties

10.1 Introduction

The small sample properties of regression coefficients in linear models with fixed regressors and iid normal error terms are well understood. In other cases, we may either rely on asymptotic (large sample) results or use Monte Carlo/bootstrap simulations to gauge the small sample properties.

Often, it is unclear whether simulations are needed, as methods based on iid assumptions or White's/Newey-West's approaches may be sufficiently effective. The only way to find out is to do the simulations, and then decide.

The implementation of the simulations depends crucially on the properties of the model and data—whether the residuals are autocorrelated, heteroskedastic, or perhaps correlated across regression equations. This chapter summarizes a few typical cases.

10.2 Monte Carlo Simulations

10.2.1 Monte Carlo Simulations in the Simplest Case

A Monte Carlo simulation generates many artificial samples from a parameterized model and then estimates the statistics (for instance, the OLS coefficients) for each of those samples. The distribution of the statistics is then used as the small sample distribution of the estimator.

The following is an example of how Monte Carlo simulations could be done in the case of a linear model

$$y_t = x_t' \beta + u_t, \quad (10.1)$$

where u_t is iid and x_t is stochastic but independent of u_{t+s} for all s . This means that x_t

cannot include lags of y_t . For the simulations, we need information on (i) the coefficients β ; (ii) the properties of u_t ; (iii) and a process for x_t .

The values of β are often derived from a combination of estimation results and theoretical considerations. Details are found in Section 10.2.2. The properties of the residual are mostly calibrated to data, both the variance and the shape of the distribution (normal or t-distribution with few degrees of freedom are commonly used). The process for x_t is typically estimated using data on x_t (for instance, a VAR system $x_t = A_1x_{t-1} + A_2x_{t-2} + \varepsilon_t$). Alternatively, we could simply use the actual sample of x_t and repeat it.

Example 10.1 (*Simulating VAR models*) For instance, a VAR(2,) $x_t = A_1x_{t-1} + A_2x_{t-2} + \varepsilon_t$, where x_t could be a vector of variables. The simulation procedure is straightforward. First, estimate the model on data and record the estimates $(A_1, A_2, \text{Var}(\varepsilon_t))$. Second, draw a new time series of residuals, $\tilde{\varepsilon}_t$ for $t = 1, \dots, T$ and construct an artificial sample recursively (first $t = 1$, then $t = 2$ and so forth) as

$$\tilde{x}_t = A_1\tilde{x}_{t-1} + A_2\tilde{x}_{t-2} + \tilde{\varepsilon}_t.$$

This requires some starting values for \tilde{x}_{-1} and \tilde{x}_0 . One approach is to pick average values, but start the simulation at $t = -100$ (or similarly) and finally discard all values for $t \leq 0$.

To illustrate, suppose these simulations are used to obtain a 5% critical value for testing a null hypothesis. The Monte Carlo experiment takes the following steps:

1. Construct an artificial sample of the regressors (see above), \tilde{x}_t for $t = 1, \dots, T$.
2. Draw random numbers \tilde{u}_t for $t = 1, \dots, T$ from a prespecified distribution (for instance, a normal or t_v) and use those together with the artificial sample of \tilde{x}_t to calculate an artificial sample \tilde{y}_t for $t = 1, \dots, T$ from

$$\tilde{y}_t = \tilde{x}'_t \beta + \tilde{u}_t, \quad (10.2)$$

by using the prespecified values of the coefficients β , adjusted to obey the null hypothesis.

3. Calculate an estimate $\tilde{\beta}$ and record it along with the test statistic of the null hypothesis.

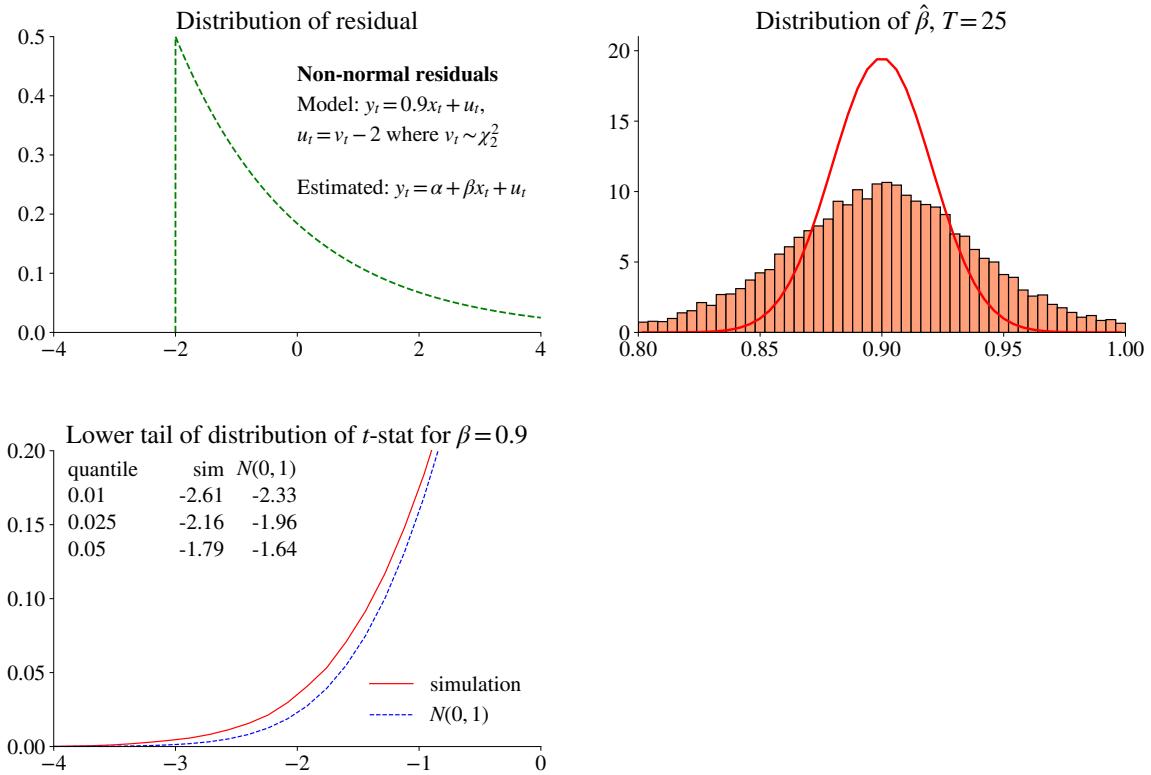


Figure 10.1: Results from a Monte Carlo experiment with thick-tailed errors.

4. Repeat the previous steps N (25,000, say) times. Increasing the number of repetitions improves the approximation of the small sample distribution.
5. Sort your simulated $\tilde{\beta}$ and the test statistic in ascending order. For a one-sided test (for instance, a chi-square test), take the $(0.95N)$ th observation in this sorted vector as your 5% critical values. For a two-sided test (for instance, a t-test), take the $(0.025N)$ th and $(0.975N)$ th observations as the 5% critical values.
6. You may also want to plot a histogram of $\tilde{\beta}$ and the test statistic to study if there is a small sample bias, and more generally, how the distribution looks like. See *Figure 10.1* for an example.

The same basic procedure is used when y_t is a vector, except that correlations across the elements of the residuals vector, u_t , must be considered. For instance, we might want to generate the vector \tilde{u}_t from a $N(\mathbf{0}_{n \times 1}, \Sigma)$ distribution—where Σ is the variance-covariance matrix of u_t .

10.2.2 The Choice of β Coefficients

Whenever the simulations create \tilde{y}_t similar to (10.2), although with possibly different ways of generating the residuals, the choice of β is important. In many cases, the point estimates are used. In other cases, the point estimates are adjusted to *simulate the model under the null hypothesis* in order to study the size of tests and to find valid critical values for small samples. Alternatively, to *simulate the model under an alternative hypothesis* in order to study the power of the test.

An further complication is that the average estimate $\tilde{\beta}$ across simulations may not always equal β due to small sample bias. If we still want to use the simulations to find critical values, then we could center the test statistic on the the average estimate in the simulations, $\tilde{\beta}$. For instance, for a t -test we calculate

$$t = \frac{\tilde{\beta} - \text{average } \tilde{\beta}}{\text{Std}(\tilde{\beta})} \quad (10.3)$$

for each simulation and then take the $(0.025N)$ th and $(0.975N)$ th observations as the 5% critical values (instead of the ± 1.96 from a standard normal distribution). In this expression, $\text{Std}(\hat{\beta})$ should be a consistent estimate of the standard error, which could be either from (a) the original sample ($\text{Std}(\hat{\beta})$ from White, Newey-West, etc); (b) the simulation itself ($\text{Std}(\tilde{\beta})$, again from a consistent estimate for the simulated sample); (c) the standard deviation of $\tilde{\beta}$ across the N simulations. In the latter case, the $\tilde{\beta}$ values are often winsorized (at quantiles 0.01 and 0.99, say) to mitigate the risk that a few very strange simulated samples dominate. These critical values can then be used for t -tests.

A similar reasoning applies to joint tests of coefficients. Consider a linear combination of the coefficients, $R\tilde{\beta}$. If V^* is the variance-covariance matrix (as before, a consistent estimate based on the original sample or each simulation), then for each sample we could calculate the quadratic form

$$\Xi = [R(\tilde{\beta} - \text{average } \tilde{\beta})]'(RV^*R')^{-1}[R(\tilde{\beta} - \text{average } \tilde{\beta})]. \quad (10.4)$$

In this case, we could use the $(0.95N)$ th simulated value as the 5% critical value. This critical value can be used to hypotheses like $R\beta - q$ based on the original sample.

The simulations of test statistics like the t and Ξ are often more precise than the simulations of the regression coefficients themselves—provided that we use consistent estimates of the standard error/covariance matrix. In the limit, these statistics do not depend on model parameters—they asymptotically “pivotal”—which often improves the

convergence rate.

10.2.3 *Monte Carlo Simulations when x_t Includes Lags of y_t

If x_t contains lags of y_t , then we must set up the simulations to preserve the temporal link in every artificial sample. For instance, if x_t includes y_{t-1} and another vector z_t of variables which are independent of u_{t+s} for all s

$$\begin{aligned} y_t &= x_t' \beta + u_t \\ &= \gamma y_{t-1} + z_t' \phi + u_t \end{aligned} \quad (10.5)$$

We can then generate an artificial sample as follows. First, create a sample \tilde{z}_t for $t = 1, \dots, T$ by some time series model (for instance, a VAR) or by taking the observed sample itself. Second, generate \tilde{y}_t recursively

$$\tilde{y}_t = \gamma \tilde{y}_{t-1} + z_t' \phi + \tilde{u}_t \text{ for } t = 1, \dots, T \quad (10.6)$$

See Figures 10.2–10.3 for examples.

We clearly need the initial value \tilde{y}_0 to start up the artificial sample. A startup sample (starting at $t = -100$, say) can be employed to generate the \tilde{y}_t series, although we later use only observations starting at $t = 1$ for the subsequent analysis. This reduces the effect of the choice of the initial point.

10.2.4 Monte Carlo Simulations with non-iid Residuals

With non-iid residuals (for instance, autocorrelation and heteroskedasticity) we need to model the process of residuals.

If the residuals are *autocorrelated*, then we could estimate a time series process from the fitted residuals (or use some theory-based values) and use that to generate artificial samples of residuals. For instance, with an MA(1) we get

$$\tilde{u}_t = v_t + \theta_1 v_{t-1}, \text{ where } v_t \text{ is iid.} \quad (10.7)$$

See Figure 10.3 for an illustration.

Alternatively, *heteroskedastic residuals* can be generated by $N(0, \sigma_t^2)$ process, but where $\ln \sigma_t^2$ is approximated by the fitted values from

$$\ln \hat{\sigma}_t^2 = c' w_t + \eta_t, \quad (10.8)$$

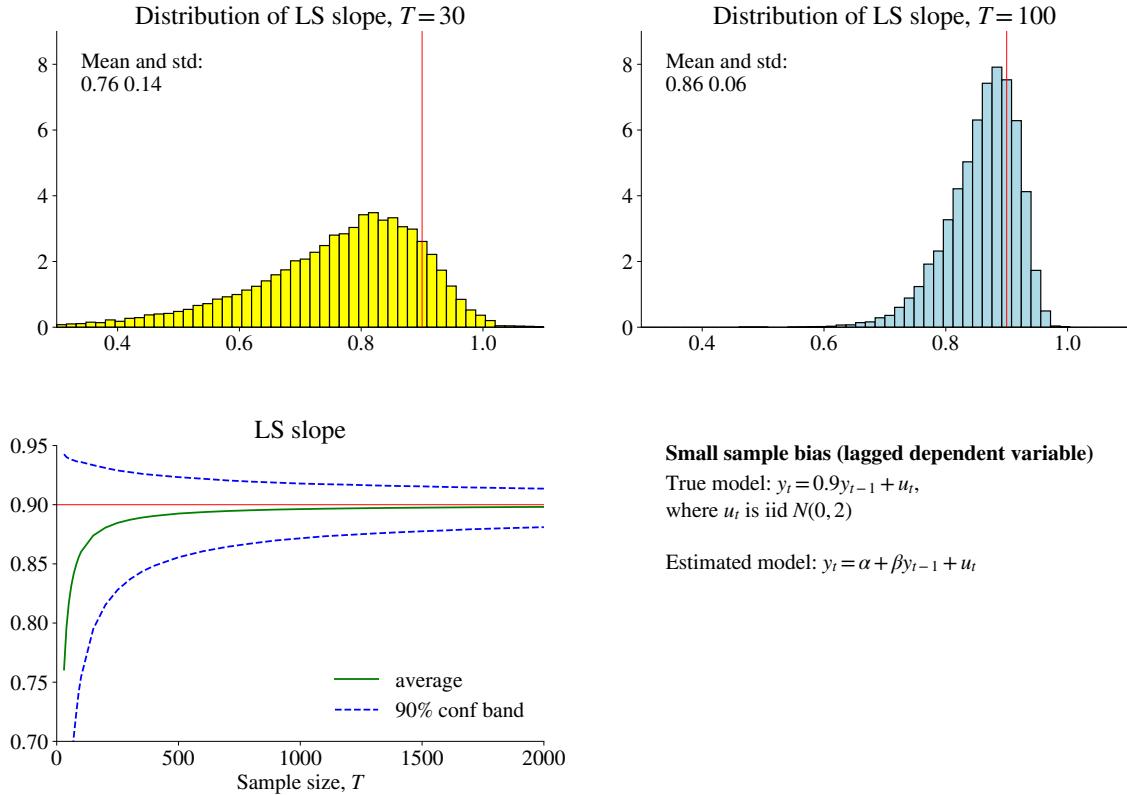


Figure 10.2: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

where w_t includes the squares and cross product of all the regressors, and where we use $\sigma_t^2 = \exp(\hat{c}'w_t)$. Alternatively, use some other transformation that guarantees that $\sigma_t^2 > 0$.

10.3 Bootstrapping

10.3.1 Bootstrapping in the Simplest Case

Bootstrapping is another simulation approach, where we construct artificial samples by drawing from the actual data, rather than a theoretical distribution. The advantage of this is that we do not have to specify/estimate the process of the residuals and regressors, as we do in a Monte Carlo experiment.

The bootstrap approach works particularly well when the residuals are iid and independent of x_{t-s} for all s . (This means that x_t cannot include lags of y_t .) We initially consider bootstrapping the linear model (10.1). The procedure is similar to the Monte

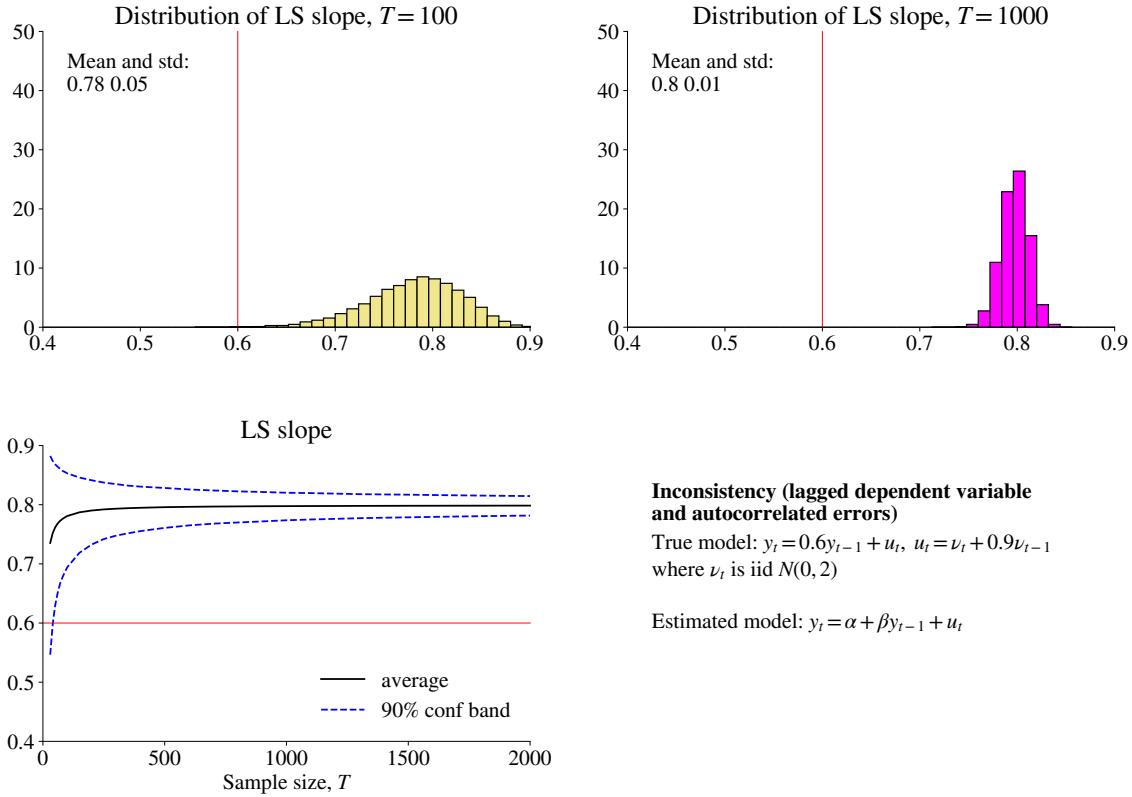


Figure 10.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

Carlo approach, except that the artificial sample is generated differently. In particular, Step 1 in the Monte Carlo simulation is replaced by drawing (with replacement) \tilde{u}_t from the *fitted residual* (“residual resampling”). This approach works also when y_t is a vector of dependent variables—and will then help retain the cross-sectional correlation of the residuals. The same issues with the choice β applies, and is discussed in section 10.2.2.

Remark 10.2 (*Bootstrapped confidence bands**) *Using the simulated 0.025th and 0.975th quantiles of the bootstrapped $\tilde{\beta}$ values is a way of creating a 95% confidence band, sometimes called Efron’s “bootstrap percentile method”. The “bootstrap percentile t-method” (also suggested by Efron) is often considered to be an improvement. To implement it, first define $\tilde{t} = (\tilde{\beta} - \hat{\beta}) / \text{Std}(\tilde{\beta})$, where $\text{Std}(\tilde{\beta})$ is the standard deviation across the bootstrap estimates. (Sometimes the centering is done by subtracting the average of $\tilde{\beta}$ values instead of the point estimate $\hat{\beta}$). Let $Q(\tilde{t}, p)$ be the p th quantile of \tilde{t} . Then, we could define a 95% confidence band as $[\hat{\beta} - Q(\tilde{t}, 0.975) \text{Std}(\tilde{\beta}), \hat{\beta} - Q(\tilde{t}, 0.025) \text{Std}(\tilde{\beta})]$,*

	$\gamma = 0$		$\gamma = 1$	
	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$
Simulated	7.1	19.0	13.5	24.8
OLS formula	7.1	13.3	13.4	19.2
White's	7.0	18.5	13.3	24.3
Bootstrap	7.1	18.5	13.4	24.3
Bootstrap 2	7.0	18.5	13.3	24.3
Jackknife	7.1	18.9	13.5	24.8
FGLS	7.5	17.3	14.0	24.1

Table 10.1: Standard error of OLS slope (%) under heteroskedasticity (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1 + \gamma|z_t| + \alpha|x_t|)^2$, where z_t is iid $N(0, 1)$ and independent of x_t . Sample length: 200. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

where $\text{Std}(\hat{\beta})$ is a consistent estimate of the standard deviation of $\hat{\beta}$ (for instance, from White's or Newey-West's methods). Notice that the quantiles are reversed, compared to what would perhaps be expected and that both are subtracted.

10.3.2 *Bootstrapping when x_t Includes Lags of y_t

When x_t contains lagged values of y_t , then we have to modify the approach since \tilde{u}_t can become correlated with x_t . The easiest way to handle this is as in the Monte Carlo simulations in (10.6), but where \tilde{u}_t are drawn (with replacement) from the sample of fitted residuals.

10.3.3 Bootstrapping when Errors Are Heteroskedastic

Suppose that the residuals are heteroskedastic, but serially uncorrelated. If the heteroskedasticity is unrelated to the regressors, then we can still use the approach discussed above. In contrast, if the heteroskedasticity is related to the regressors, then it would then be wrong to pair x_t with just any $\tilde{u}_t = \hat{u}_s$ since that destroys the relation between x_t and the variance of the residual.

Instead, we may apply an alternative way of bootstrapping: generate the artificial sample by drawing (with replacement) pairs (y_s, x_s) , that is, let the artificial pair in t be $(\tilde{y}_t, \tilde{x}_t) = (y_s, x_s) = (x'_s \beta + \hat{u}_s, x_s)$ for some random draw (with replacement) of s so we are always pairing the residual, \hat{u}_s , with the contemporaneous regressors, x_s .

This approach is also called “case resampling.” Note that we are always sampling with replacement—otherwise the approach of drawing pairs would be to just re-create the original data set. This approach works also when y_t is a vector of dependent variables. See Table 10.1 for results from a simulation.

Example 10.3 With $T = 3$, the artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (y_2, x_2) \\ (y_3, x_3) \\ (y_3, x_3) \end{bmatrix} = \begin{bmatrix} (x'_2\beta + \hat{u}_2, x_2) \\ (x'_3\beta + \hat{u}_3, x_3) \\ (x'_3\beta + \hat{u}_3, x_3) \end{bmatrix}$$

It could be argued (see, for instance, Davidson and MacKinnon (1993)) that bootstrapping the pairs (y_s, x_s) makes little sense when x_s contains lags of y_s , since the random sampling of the pair (y_s, x_s) destroys the autocorrelation pattern on the regressors.

Remark 10.4 (*The wild Bootstrap*) The wild bootstrap is also aimed at solving the heteroskedasticity problem. In this case, the artificial sample is generated as in the basic bootstrap, but we use $\tilde{u}_t = \hat{u}_t \tilde{\varepsilon}_t$ where \hat{u}_t is the fitted (OLS) residual for observation t and $\tilde{\varepsilon}_t$ is drawn from an iid random variable with mean 0 and variance 1. For instance, $\tilde{\varepsilon}_t$ could have a two-point distribution where it is either -1 or 1 with equal probabilities.

10.3.4 Bootstrapping when Errors Are Autocorrelated

It is quite hard to handle the case when the residuals are serially dependent, since we must then sample in such a way that we do not destroy the autocorrelation structure. A common approach is to fit a time series model for the residuals, for instance, an AR(1), and then bootstrap the (hopefully iid) innovations to that process.

Another approach amounts to *resampling blocks* of residuals. For instance, suppose the sample has 10 observations, and we decide to create blocks of 3 observations. The first block is $(\hat{u}_1, \hat{u}_2, \hat{u}_3)$, the second block is $(\hat{u}_2, \hat{u}_3, \hat{u}_4)$, and so forth until the last block, $(\hat{u}_8, \hat{u}_9, \hat{u}_{10})$. If we need a sample of length 3τ , say, then we simply draw τ of those block randomly (with replacement) and stack them to form a longer series. To handle end point effects (so that all data points have the same probability to be drawn), we also create blocks by “wrapping” the data around a circle. In practice, this means that we use the following blocks: $(\hat{u}_{10}, \hat{u}_1, \hat{u}_2)$ and $(\hat{u}_9, \hat{u}_{10}, \hat{u}_1)$. The length of the blocks should clearly depend on the degree of autocorrelation, but $T^{1/3}$ is sometimes recommended as a rough guide. An alternative approach is to have non-overlapping blocks. See Berkowitz and Kilian (2000) for some other approaches. See Table 10.2 for results from a simulation.

Example 10.5 With $T = 9$ and a block size of 3, the artificial sample could be

$$\underbrace{u_2, u_3, u_4}_{block\ 2}, \underbrace{u_7, u_8, u_9}_{block\ 7}, \underbrace{u_4, u_5, u_6}_{block\ 4}.$$

	$\rho = 0.0$	$\rho = 0.75$
Simulated	5.9	23.1
OLS formula	5.8	8.6
Newey-West	5.7	18.5
VARHAC	5.7	22.2
Bootstrapped	5.5	19.6
FGLS	5.9	23.2

Table 10.2: Standard error of OLS intercept (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t$, ξ_t is iid $N()$. NW uses 10 lags. VARHAC uses 10 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300.

Further Reading

Horowitz (2001), Greene (2018) 15, and Hansen (2022a) 10 provide more details.

Also, see Cochrane (2005) 15.2, Davidson and MacKinnon (1993) 21, Davison and Hinkley (1997), Efron and Tibshirani (1993) (bootstrapping, chap 9 in particular), and Berkowitz and Kilian (2000) (bootstrapping in time series models).

Chapter 11

Portfolio Sorts

11.1 Overview

This chapter discusses how portfolio sorts are used to create return series for assets with similar characteristics. Once equipped with several such returns series, we can apply the techniques in, for instance, the chapter on system estimation to test the difference of alphas.

Portfolio sorts are used to construct portfolios (groups) based on some characteristic, for instance, firm size. Once the portfolios have been defined, we often compute the (possibly weighted) average within each portfolio

$$R_{gt} = \sum_{i \in \text{group } g} w_{it} R_{it}, \quad (11.1)$$

where w_{it} is the relative portfolio weight of asset i in the portfolio (with weights summing to one) in period t . An unweighted average is common, but sometimes rank-based weights are applied. The sorting and portfolio construction is typically repeated at regular intervals. For instance, the Fama-French size portfolios are based on the market capitalization and are rebalanced every June. For a daily momentum strategy, we would rather redo the sort every day based on recent performance.

Portfolio sorts are very similar to *dynamic trading strategies*, where the basic idea is to create a portfolio based on some kind of sorting of a trading signal.

The performance of a portfolio is often measured by the mean return, the Sharpe ratio or Jensen's alpha (the intercept from a regression on the market excess return and possibly also other benchmark returns).

11.2 Univariate Sorts

A simple and commonly applied method for studying how an asset characteristic (z_i) is related to returns (or some other performance measure) is to do a *univariate sort*. For instance, we could sort the assets $i = 1, \dots, n$ according to z_i and then construct three portfolios: (1) for those i whose z_i belong to the lowest 1/3; (2) those in the mid 1/3 and (3) those in the highest 1/3. Then, we measure the returns portfolios, and perhaps also analyse the return of portfolio 3 minus the return on portfolio 1.

Empirical Example 11.1 (*Sorting on recent returns*) See Table 11.1 for an empirical example where the 25 FF portfolios are sorted into low/low recent 22-day returns, with 5 portfolios in each. The results indicate strong momentum.

	Portfolio returns
Low 22-day return	4.15 (0.20) [-5.36]
High 22-day return	13.71 (0.74) [5.16]
Difference (H-L)	9.55 (0.85) [10.52]

Table 11.1: Average excess returns, (Sharpe ratios) and $[\alpha]$ for 3 portfolios from a univariate sort on recent (22-day) returns (5/5 assets). Annualized figures. Daily data on 25 FF portfolios 1979:01-2024:12

Example 11.2 (*Simplified version of “Betting against beta” by Frazzini and Pedersen**) First, find those assets with $\beta_{i,t-1} < \text{median}(\beta_{i,t-1})$ where $\beta_{i,t-1}$ is the CAPM beta estimated on data up to and including $t - 1$. (The median is across the assets.) This is the low beta group. Second, calculate the equally weighted portfolio return in t . Third, repeat for all periods. Fourth, do points 1–3 also for the high beta assets, $\beta_{i,t-1} \geq \text{median}(\beta_{i,t-1})$. Fifth, form the excess return as the difference between the two portfolios.

11.3 Bivariate Sorts

Bivariate sorts (also called double sorts) are used when there are two important characteristics (here called x and z) and you want to study how z affects returns—controlling for x (that is, holding x “constant”). This may well be important if x and z are correlated.

Bivariate sorts can be done in several ways. An *independent bivariate sort* first does a univariate sort of x_i (say, forming 3 categories: growth, neutral or value), then it makes another univariate sort according to another sorting variable z_i (say, forming two categories: small or big). Then we find the intersections of the two sorts as in the following matrix

	Low x_i	Medium x_i	High x_i	
Low z_i	(x_L, z_L)	(x_M, z_L)	(x_H, z_L)	(11.2)
High z_i	(x_L, z_H)	(x_M, z_H)	(x_H, z_H)	

where, for instance, (x_M, z_H) denote the set of assets that belong to the medium x category and high z category. (Notice that this matrix has a different structure than a traditional scatter plot with x on the horizontal axis and z on the vertical axis: in this we put low z_i on the first line and high z_i on the second.)

In an independent bivariate sort we cannot directly control how many assets there will be in each group, and some groups might be empty, at least, for some periods (see Example 11.3 and Figure 11.1).

Once the portfolio sort is done, we typically calculate the average return (or some other performance measure of interest) of each portfolio. In the independent sort, you can either compare across rows or across columns.

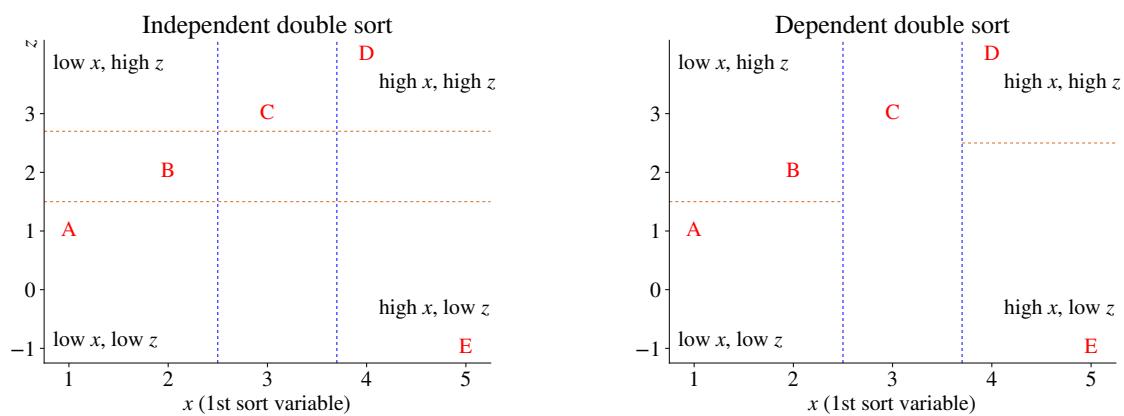


Figure 11.1: Example of bivariate sorts. The data is indicated by letters.

Example 11.3 (Independent double sort) Suppose there are 5 assets (labelled A, B,...) and that the values of x and z are

	x_i	z_i
Asset A	1	1
Asset B	2	2
Asset C	3	3
Asset D	4	4
Asset E	5	-1

We form low/high groups with 2 elements in each

	Assets
Low x	A, B
High x	D, E
Low z	E, A
High z	C, D

The independent double sort then gives

	$Low x$	$High x$
Low z	A	E
High z		D

Notice that there are no assets in the (low x , high z) group. See also Figure 11.1 for an illustration, but notice that the scatter plot has a different structure: low z values are plotted below high z values.

Empirical Example 11.4 (Independent sorting on recent volatility and returns) We first sort the 25 FF portfolios according to recent volatility, putting 10 into the low group and 10 into the high group. Then we sort on recent returns, also putting 10 into a low group and 10 into a high group. Finally, we form intersections. Figure 11.2 illustrates (for a short subsample) how the number of portfolios in the “low vol, low return” group varies over time. Typically, there are 4–5 portfolios in the group, but it varies considerably. Sometimes the group is empty.

In a *dependent bivariate sort* we first sort according to x_i as before. Then, *within* an x category we sort according to z_i . This allows us to control the number of assets in

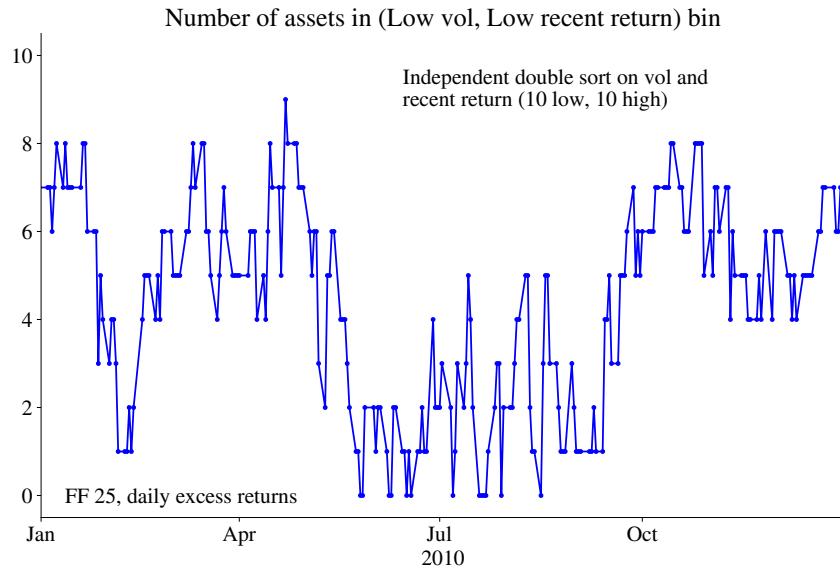


Figure 11.2: Independent bivariate portfolio sort, 25 FF portfolios

each group. Notice that the ordering matters in the dependent sort: letting x represent growth/neutral/value and z small/big will not give the same results as switching the labels. In the dependent sort, we compare across the z categories, that is, *across rows* in (11.2), for instance, the return of (x_L, z_H) minus the return of (x_L, z_L) and so forth. In (11.2) this gives three numbers—which are sometimes averaged: the interpretation is that you are studying the effect of z (here: small/large), but controlling for x (here: one of growth/neutral/value). See Figure 11.1 for an example.

Example 11.5 (Dependent double sort) Continuing the previous example, the dependent double sort (with one asset in each portfolio) gives

	<u>Low x</u>	<u>High x</u>
<u>Low z</u>	A	E
<u>High z</u>	B	D

For instance, among the “high x ” assets D and E , asset E has a lower z value so it is allocated to the (high x , low z) portfolio, while asset D has a higher z value so it is allocated to the (high x , high z) portfolio. Notice that all portfolios are populated. See also Figure 11.1 for an illustration.

Empirical Example 11.6 (Dependent sorting on recent volatility and returns) We first sort the 25 FF portfolios according to recent volatility, putting 10 into the low group and

10 into the high group. Within the “low vol” group, we then sort according to recent returns, putting 5 into the “low vol, low return” group and 5 into the “low vol, high return” group. We do the same within the “high volatility” group. This means that there are always 5 portfolios in each of the 4 groups.

	Low 22-day vol	High 22-day vol
Low 22-day return	8.47 (0.53) [0.99]	4.76 (0.21) [-5.70]
High 22-day return	12.63 (0.81) [5.40]	11.96 (0.56) [2.09]
High - low	4.16 (0.82) [4.41]	7.20 (0.90) [7.79]
High - low, average	5.68 (1.06) [6.1]	

Table 11.2: Average excess returns, (Sharpe ratios) and $[\alpha]$ for 4 portfolios from a dependent bivariate sort. The first sort is on volatility (10/10 assets), the second sort (within each volatility bin) is on recent returns (5/5 assets). Annualized figures. Daily data on 25 FF portfolios 1979:01-2024:12

The bivariate sort is designed to handle some *correlation* between x and z . If there is no correlation, then a single sort is enough. However, the bivariate sort will break down if the correlation is too strong. In the independent sort, it can lead to few (or even zero) assets in the off-diagonal portfolios if the correlation is positive (and vice versa if the correlation is negative). In the dependent sort, it may simply lead to results that cannot be trusted, see Figure 11.3 for an example. The figure illustrates how the “high z ” portfolios have clearly higher x values than the “low z ” portfolios have, so the approach is only moderately successful in controlling for x . This could be solved by having smaller x bins (which may require many assets), so that the variation in x within each bin is small compared to the variation in z . For instance, the low x bin could contain 20% of the assets, the high x also 20%, leaving out the 60% in the middle. As an alternative, we could consider an orthogonalisation (see below).

Remark 11.7 (*When x and z are perfectly correlated**) If we change Example 11.3 so z

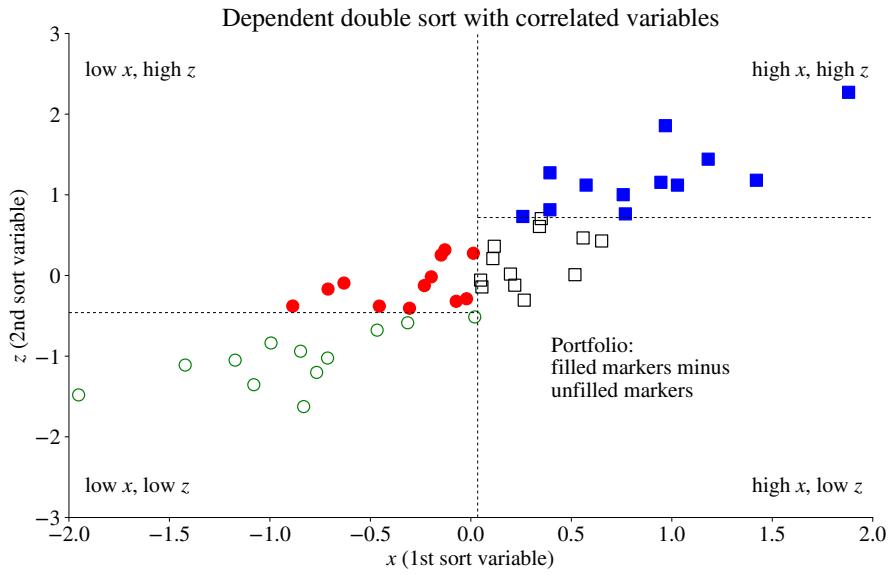


Figure 11.3: Example of dependent bivariate sort with correlated variables

equals x , then the independent double sort gives

	<i>Low x</i>	<i>High x</i>
<i>Low z</i>		<i>D, E</i>
<i>High z</i>	<i>A, B</i>	

This has the problem that the off-diagonal portfolios are empty. In contrast, the dependent sort gives

	<i>Low x</i>	<i>High x</i>
<i>Low z</i>	<i>A</i>	<i>D</i>
<i>High z</i>	<i>B</i>	<i>E</i>

The latter has the problem that comparing across rows does not control for x . For instance, asset *B* has a higher x (and z) value than asset *A*.

Remark 11.8 (*The Fama-French factors**) The SMB and HML are created by an independent bivariate sort. First, classify firms according to the book/market value: low (growth stocks, using 30th percentile as cutoff), neutral or high (value stocks, using 70th percentile as cutoff). Second, classify firms according to size: small or big, using the median as a

cutoff. Create six value weighted portfolios from the intersection of those categories

	<i>Low book/market</i>	<i>Medium book/market</i>	<i>High book/market</i>
<i>Small</i>	<i>Small Growth (SG)</i>	<i>Small Neutral (SN)</i>	<i>Small Value (SV)</i>
<i>Big</i>	<i>Big Growth (BG)</i>	<i>Big Neutral (BN)</i>	<i>Big Value (BV)</i>

The *SMB* is the average of the small portfolios minus the average of the big portfolios: $SMB = 1/3(SG + SN + SV) - 1/3(BG + BN + BV)$. Rearranging gives $SMB = 1/3(SG - BG) + 1/3(SN - BN) + SV + 1/3(SV - BV)$, which shows that it represents the return on small stocks (for a given book/market) minus the return on big stocks (for same book/market). The *HML* is the average of the value stocks minus the growth stocks, $HML = 1/2(SV + BV) - 1/2(SG + BG)$, which can be rearranged as $HML = 1/2(SV - SG) + 1/2(BV - BG)$, which shows that it represents the return on value stocks (for a given size) minus the return on growth stocks (for the same size).

11.4 Orthogonalisation

Single sort on orthogonalised data is an alternative to a double sort and may be better at handling strong (linear) correlation of x and z . It involves two steps. First, run a regression of (z on x and a constant) to get coefficients (a, b). Second, do a single sort on the residual

$$\varepsilon_i = z_i - (a + bx_i). \quad (11.3)$$

The regression can be done in several different ways: (1) a cross-sectional regression as in Figure 11.4 ; (2) time series regressions; (3) a panel regression. There is also a choice between using the full sample or just data up to $t - 1$ (and then redo for each period).

Further Reading

Bali, Engle, and Murray (2016) discusses many aspects of portfolio sorts.

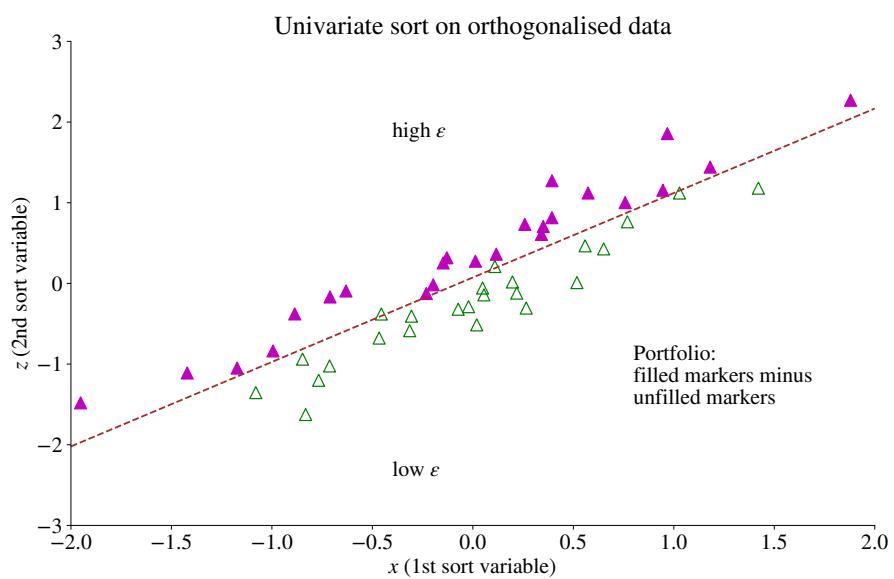


Figure 11.4: Example of univariate sort on orthogonalised data

Chapter 12

Financial Panel Data

12.1 Introduction to Panel Data

A panel data set (also called a longitudinal data set) has data on a cross-section ($i = 1, 2, \dots, N$, individuals or firms) over many time periods ($t = 1, 2, \dots, T$). Our aim is to estimate a linear relation between the dependent variable and the regressors

$$y_{it} = \alpha_i + x'_{it}\beta_i + u_{it}, \quad (12.1)$$

where the coefficients (α_i, β_i) may or may not be different for different individuals (this is discussed in detail below). As examples of such applications, we may want to evaluate if alphas or betas of different mutual funds are related to fund characteristics, for instance, age or trading activity. Alternatively, we want to investigate whether firms with different types of board compositions perform differently.

Data on the dependent variable has this structure

$$\begin{array}{cccc} & i = 1 & i = 2 & \cdots & i = N \\ \hline t = 1 : & y_{11} & y_{21} & & y_{N1} \\ t = 2 : & y_{12} & y_{22} & & y_{N2} \\ \vdots & & & & \\ t = T : & y_{1T} & y_{2T} & & y_{NT} \end{array} \quad (12.2)$$

The structure for each of the regressors is similar, although it can also be the case that (some of) the regressors are the same for all N cross-sectional units (for instance, when the regressors are pricing factors like the market excess return). When needed for clarity we will use the $y_{i,t}$ notation instead of y_{it} .

The structure in (12.2) implicitly assumes that we have a *balanced panel*, that is, have

data for all the cells. However, it is often the case that the panel is *unbalanced* in the sense that some data is missing. For instance, we may not have data on regressor 2 for $i = 7$ and $t = 3$. If data is *missing in a random way*, then we can simply exclude (y_{it}, x_{it}) for the missing (i, t) . (Practical tips on how to do that is discussed later.) In our example that means just excluding $(y_{7,3}, x_{7,3})$ but keeping all other data. In contrast, if data is missing in a non-random way (for instance, depending on the value of y_{it}), then we have to apply more sophisticated sample-selection models (not discussed in this chapter).

12.2 Calendar Time Regressions

Reference: Bali, Engle, and Murray (2016)

This chapter uses a calendar time (CalTime) regressions as a benchmark for some of the panel regressions. The reason is that the CalTime approach and panel regressions are sometimes very similar—and we know how to estimate and test calendar time regressions (with trustworthy standard errors). In such cases, the two methods should give similar results. If not, this may indicate issues.

In the CalTime approach, we first define M discrete groups by a portfolio sort (for instance, construct ten portfolios for different firm size deciles) and calculate their respective excess returns.

Then, we run M regressions

$$y_{jt} = z_t' \theta_j + u_{jt}, \text{ for } j = 1, 2, \dots, M \quad (12.3)$$

where y_{jt} represents the excess return of group j and z_t typically includes a constant and various return factors (for instance, the three Fama-French factors). By estimating these M equations as a system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the “alpha”) is the same for all portfolios.

Example 12.1 (*CalTime with three groups*) *With three groups, estimate the following system*

$$\begin{aligned} y_{1t} &= z_t' \theta_1 + u_{1t} \\ y_{2t} &= z_t' \theta_2 + u_{2t} \\ y_{3t} &= z_t' \theta_3 + u_{3t}. \end{aligned}$$

The CalTime approach is straightforward and the cross-sectional correlations are handled. However, it forces us to define discrete asset groups—which makes it hard to handle several different types of characteristics at the same time. In contrast, panel data models are more flexible but also come with the challenge of getting the standard errors right.

Empirical Example 12.2 (*Investor activity vs performance*) See Table 12.1 for results on a ten-year panel of some 60,000 Swedish pension savers from Dahlquist, Martinez, and Söderlind (2017). In this case, the dependent variables are the daily return of one of three pension investment portfolios (defined in terms of trading activity). The regressors include a constant, 7 risk factors (global and Swedish market, SMB, HML as a well as a bond factor) on ± 2 days ($1 + 7 \times 5$ regressors).

	Inactive	Active	Highly Active
coef	-0.76	3.08	8.65
t-tstat NW	-0.69	1.77	2.73

Table 12.1: Calendar time regressions, estimated as a system. Annualised coefficients and t-stats from Table 10, in Dahlquist et al (RFS 2017). For Inactive the coefficient is the annualised alpha, but for the other two categories it is the difference in alpha to the Inactive. Three EW portfolios based on 62640 individuals, 2116 days. The dependent variables are the returns of the EW portfolio based on the activity indicators. The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

12.3 An Overview of Panel Data Models

A *pooled model* assumes that all individuals have the same coefficients (no subscript on α and β), so (12.1) becomes

$$y_{it} = \alpha + x'_{it}\beta + u_{it}. \quad (12.4)$$

This model can be estimated by pooled OLS (see below).

A *fixed effects model* assumes that all individuals have the same slope coefficients, but that their intercepts might differ

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}. \quad (12.5)$$

An extension of the fixed effects model is to also allow for *time fixed effects*

$$y_{it} = \lambda_t + \alpha_i + x'_{it}\beta + u_{it}. \quad (12.6)$$

Estimation of these models is discussed below. One way of handling the fixed effects are a first-difference model

$$\Delta y_{it} = \Delta \lambda_t + \Delta x'_{it} \beta + u_{it}^*, \quad (12.7)$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ denotes a first difference (in the time dimension).

A *random effects model* is similar to a fixed effects model, except that the individual α_i now contains a common component (α) and a random individual component (μ_i). We can then write the model as

$$y_{it} = \alpha + x'_{it} \beta + u_{it} \text{ where } u_{it} = \mu_i + \varepsilon_{it}. \quad (12.8)$$

The ε_{it} is typically assumed to be uncorrelated across time and individuals, but the μ_i terms make the u_{it} residuals correlated over time (for the same individual). The estimation of this model is discussed later.

The *unrestricted model* allows all individuals to have different coefficients (hence a subscript i on β_i). These regressions could be estimated by OLS for each individual separately.

12.4 Pooled OLS

Consider the regression model

$$y_{it} = z'_{it} \theta + u_{it}, \quad (12.9)$$

where z_{it} is an $k \times 1$ vector. For notational convenience, this section assumes that any constant is included in the z_{it} vector along with the other regressors (x_{it}). Notice that the coefficients are the same across individuals (and time), but that the regressors may vary along both the time series and cross-sectional dimensions. As usual, we assume $E u_{it} = 0$ and $Cov(z_{it}, u_{jt}) = 0$ (across all i and j).

Define the matrices

$$S_{zz} = \sum_{t=1}^T \sum_{i=1}^N z_{it} z'_{it} \text{ (a } k \times k \text{ matrix)} \quad (12.10)$$

$$S_{zy} = \sum_{t=1}^T \sum_{i=1}^N z_{it} y_{it} \text{ (a } k \times 1 \text{ vector).} \quad (12.11)$$

The LS estimator (stacking all TN observations) is then

$$\hat{\theta} = S_{zz}^{-1} S_{zy}. \quad (12.12)$$

Recall that we can (conceptually) decompose the point estimate $\hat{\theta}$ by using (12.9) to

substitute for y_{it} in S_{zy} (12.11) and then in (12.12). The result is

$$\hat{\theta} = \theta + S_{zz}^{-1} \sum_{t=1}^T \sum_{i=1}^N z_{it} u_{it}. \quad (12.13)$$

The variance-covariance matrix can then be written

$$\text{Var}(\hat{\theta}) = S_{zz}^{-1} S S_{zz}^{-1}, \text{ where} \quad (12.14)$$

$$S = \text{Var}(\sum_{t=1}^T \sum_{i=1}^N g_{it}), \quad (12.15)$$

where $g_{it} = z_{it} u_{it}$ is used as short hand notation.

Remark 12.3 (*Background to (12.14)*) As discussed in earlier chapters, the variance-covariance matrix (12.14) can be motivated by either (a) assuming that x_t are fixed regressors or (b) by asymptotic results. A more precise statement is thus that $S_{zz}^{-1} \hat{S} S_{zz}^{-1}$ is an estimate of the variance-covariance matrix. In practice, this gives the same estimate of the variance-covariance matrix as under assumption (a).

Remark 12.4 (*Unbalanced panels**) With missing values in (y_{it}, z_{it}) we want to exclude that observation. This can be done in several ways. For instance, by changing the summations in (12.10), (12.11) and (12.15) to skip over such data points. Alternatively, we can set $(y_{it}, z_{it}) = (0, \mathbf{0}_k)$ so all variables related to (t, i) are set to zero (also the constant).

When we assume that the residuals in (12.15) are iid and also independent of the regressors, then $\text{Var}(\hat{\theta})$ simplifies to the usual OLS expression. Instead, with heteroskedasticity, we can apply White's approach, and with autocorrelation the Newey-West approach (although this requires some care in order to not mix data for different cross-sectional units). Instead, with cross-sectional correlations we need to make further adjustments (see below for a discussion).

Remark 12.5 (**Panel regression vs average coefficient in the case of common regressors*) Consider the regression for cross-sectional unit i

$$y_{it} = z_t' \theta_i + u_{it}, i = 1 \dots N,$$

where the regressors are the same in all regressions—but where the coefficients might be different across i . It is straightforward to show that the cross-sectional average of the regression coefficients equals the results from a pooled panel regression.

12.4.1 Panel Regression with Clustering of Residuals

Different *cluster methods* account for a non-zero covariance within the same period (for instance, between u_{it} and u_{jt}). If there is no autocorrelation, then we can write (12.15) as

$$S = \sum_{t=1}^T \text{Var}(\Sigma_{i=1}^N g_{it}). \quad (12.16)$$

Positive correlations in the cross-section (g_{it} is correlated with g_{jt}) increases the right hand side of (12.16). (Autocorrelations are discussed in a later section.)

Example 12.6 ($N = 4$) To save space, let $v()$ denote a variance (possibly a variance-covariance matrix) and $\gamma(,)$ a covariance (possibly a matrix). For $N = 4$, the $\text{Var}()$ term in (12.16) can then be written (dropping the time subscripts to save further space)

$$\begin{aligned} \text{Var}(\Sigma_{i=1}^4 g_i) &= v(g_1) + v(g_2) + v(g_3) + v(g_4) + \\ &\quad 2\gamma(g_1, g_2) + 2\gamma(g_1, g_3) + 2\gamma(g_1, g_4) + \\ &\quad 2\gamma(g_2, g_3) + 2\gamma(g_2, g_4) + 2\gamma(g_3, g_4). \end{aligned}$$

Cross-sectional correlations mean that some $\gamma(,)$ terms are non-zero.

A cluster method makes assumptions about which cross-sectional units (i and j) can be correlated, that is, we first define C clusters ($c = 1, \dots, C$) and rewrite (12.16) as

$$S = \sum_{c=1}^C \sum_{t=1}^T \text{Var}(G_{ct}), \text{ where } G_{ct} = \sum_{i \in \text{cluster } c} g_{it}. \quad (12.17)$$

In this expression, G_{ct} is a k -vector with the sum of each of the k moment conditions in period t (g_{it}) across the members in cluster c . This will handle the correlations within that cluster, as illustrated in the next example.

Example 12.7 (Cluster method on $N = 4$) Assume that individuals 1 and 2 form cluster A and that individuals 3 and 4 form cluster B—and disregard correlations across clusters. This means setting the covariances across clusters to zero,

$$\begin{aligned} \text{Var}(\Sigma_{i=1}^4 g_i) &= v(g_1) + v(g_2) + v(g_3) + v(g_4) + \\ &\quad 2\gamma(g_1, g_2) + 2\underbrace{\gamma(g_1, g_3)}_0 + 2\underbrace{\gamma(g_1, g_4)}_0 + \\ &\quad 2\underbrace{\gamma(g_2, g_3)}_0 + 2\underbrace{\gamma(g_2, g_4)}_0 + 2\gamma(g_3, g_4). \end{aligned}$$

Notice that this can be written

$$\text{Var}(\sum_{i=1}^N g_i) = \text{Var}(g_1 + g_2) + \text{Var}(g_3 + g_4),$$

which is non the same form as (12.17).

To estimate the S matrix in (12.17), we use

$$\hat{S} = \sum_{c=1}^C \sum_{t=1}^T G_{ct} G'_{ct}. \quad (12.18)$$

In practice, this means that we use $\sum_{t=1}^T G_{ct} G'_{ct}$ to estimate $\sum_{t=1}^T \text{Var}(G_{ct})$, which is the contribution of cluster c to (12.17). The iid case is when each i is her/his own cluster. The *Driscoll-Kraay* approach puts everyone in the same single cluster.

Remark 12.8 ($T = 1^*$) When we have a pure cross-sectional data set, so $T = 1$, then we can still apply (12.18). However, it is common to scale \hat{S} by $C/(C - 1)$ to improve the small sample properties. Also, notice that the Driscoll-Kraay approach (letting everyone be in one big cluster) does not work when $T = 1$ since $\sum_{i=1}^N g_i = 0$.

	coef	t-tstat W	tstat DK
Inactive	-0.76	-56.89	-0.69
Active	3.08	37.48	1.77
Highly Active	8.65	28.73	2.73

Table 12.2: Panel regression, annualised coefficients and different t-stats from Table 10, regressions I and II in Dahlquist et al (RFS 2017). For Inactive the coefficient is the annualised alpha, but for the other two categories it is the difference in alpha to the Inactive. Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t , individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

Empirical Example 12.9 (*Investor activity vs performance*) For an empirical illustration, see Table 12.2 where White's t-stats look massively inflated compared to the earlier findings from the calendar time approach in Table 12.1. In contrast, the Driscoll-Kraay (DK) t-stats are actually the same as in Table 12.1. Table 12.3 extends the panel estimation by including more regressors (age, gender and pension rights). This would be difficult to handle in a calendar time approach, and thus illustrates that a panel regression can handle more general cases. Notice that the investor characteristics are here allowed to

change across time. For instance, an investor can be active during the early years and then become inactive.

	coef	t-tstat W	tstat DK
Inactive	-1.10	-1.63	-0.69
Active	3.10	34.61	1.79
Highly Active	8.69	28.44	2.74
Age	0.00	0.19	0.11
Male	0.62	2.94	2.22
Pension rights	-0.03	-0.39	-0.33

Table 12.3: Panel regression, annualised coefficients and different t-stats from Table 10, regressions I and II in Dahlquist et al (RFS 2017). Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t , individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

12.4.2 Reintroducing Autocorrelations*

The previous discussion of clustering disregarded autocorrelations. We now reintroduce this possibility. Clearly, this should only be applied if there are good reasons to suspect autocorrelation (for instance, as indicated by tests on the fitted residuals). Otherwise, this effectively introduces many new parameters which comes with the usual drawbacks.

The cluster method is to first define the autocovariance matrix for lag s

$$\Gamma_s = \sum_{c=1}^C \sum_{t=s+1}^T G_{c,t} G'_{c,t-s}. \quad (12.19)$$

If we allow for m lags, then the estimate of S is

$$\hat{S} = \hat{S}_0 + \sum_{s=1}^m w_s (\Gamma_s + \Gamma'_s), \quad (12.20)$$

where $w_s = 1 - s/(m + 1)$ in case we use the Bartlett weights as in the Newey-West approach, but also $w_s = 1$ can be motivated (see Petersen (2009) for a discussion). When $m = 0$, then this is clearly the same as in (12.18).

Notice that with flat weights ($w_s = 1$) and $m = T$, (12.20) gives the same as the (Arellano (2003)) estimator, which is used by many software packages. This has the advantage of few assumptions, but the disadvantage of (effectively) estimating many more parameters.

Remark 12.10 (*Arellano 2003 estimator of S) First, define $G_c = \sum_{t=1}^T G_{ct}$ as the sum (over time) for cluster c , a k -vector. Second, let $\hat{S} = \sum_{c=1}^C G_c G_c'$. This is the same as (12.20) with $w_p = 1$ and $m = T$. (For a comparison with the literature, notice that if X^c is the matrix of regressors for all observations in cluster c , $TN_c \times k$, and \hat{u}^c is the corresponding TN_c -vector of fitted residuals, then $X^{c'} u^c u^{c'} X^c$ is the same as $G_c G_c'$.)

12.5 The Within Estimator

In the fixed effects model, we allow for different individual intercepts

$$y_{it} = \alpha_i + x'_{it} \beta + u_{it}, \quad (12.21)$$

where u_{it} is iid with zero mean and variance σ_u^2 .

Figure 12.1 illustrates how a correlation between individual intercepts and a regressor (x) would affect the estimate of the slope coefficient, and how the FE estimator solves that problem. This highlights that omitted variables (here individual intercepts) can bias the estimates for included variables, if there is a correlation.

There are several ways to estimate this model. The conceptually most straightforward is to include individual dummies (N) where dummy i takes the value of one if the data refers to cross-sectional unit i and zero otherwise and estimate the model with pooled OLS. (Clearly, the regression can then not include any intercept. Alternatively, include an intercept but only $N - 1$ dummies, for $i = 2 - N$.) However, this approach can be difficult to implement since it may involve a very large number of regressors.

For an alternative which gives the same results, consider the following approach. *First*, take average across time (for a given i) of y_{it} and x_{it} in (12.21). That is, think of forming the cross-sectional regression (but do not run any estimation yet...)

$$\bar{y}_i = \alpha_i + \bar{x}'_i \beta + \bar{u}_i, \text{ where} \quad (12.22)$$

$$\bar{y}_i = \sum_{t=1}^T y_{it}/T \text{ and } \bar{x}_i = \sum_{t=1}^T x_{it}/T. \quad (12.23)$$

Second, transform the data as

$$y_{it}^* = y_{it} - \bar{y}_i \quad (12.24)$$

$$x_{it}^* = x_{it} - \bar{x}_i. \quad (12.25)$$

Figure 12.1 illustrates the transformed data in the right-hand side subfigures.

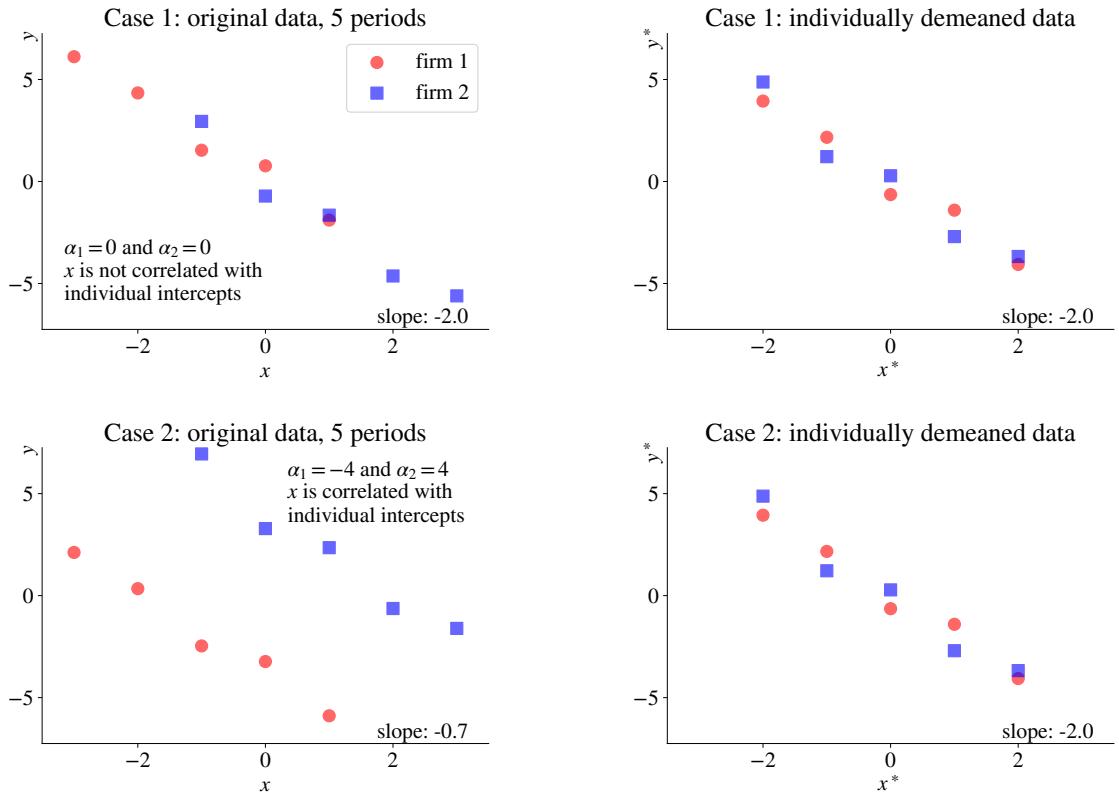


Figure 12.1: Interpretation of the FE estimator

The difference between (12.21) and (12.22) and notice that this is the same as using (12.24)–(12.25) in a regression model for transformed data

$$y_{it}^* = x_{it}^* \beta + u_{it}^*. \quad (12.26)$$

At this stage, estimate β by running pooled OLS on all observations of (12.26). There is no intercept in this regression, but adding an intercept should not affect the slope coefficients (since y_{it}^* and x_{it}^* have zero means). All the results on the variance-covariance matrix (clustering, etc.) discussed in Section 12.4.1 apply to this regression.

We denote the estimate from (12.26) by $\hat{\beta}_{FE}$ (FE stands for fixed effects) and it is also often called the *within estimator*. The interpretation of this approach is that we estimate the slope coefficients by using only the movements around individual means (not how the individual means differ). Notice that it gives the same results as OLS with dummies. If

needed, get estimates of individual intercepts as

$$\alpha_i = \bar{y}_i - \bar{x}'_i \hat{\beta}_{FE}. \quad (12.27)$$

Clearly, the within estimator wipes out all regressors that are constant across time for a given individual (e.g., gender and schooling), as they are effectively merged with the individual means. In practice, such variables must be excluded from the x_{it} vector since otherwise there will be some transformed variables, $x_{it} - \bar{x}_i$, that are always zero—causing numerical problems. See Table 12.4 for an example (see "south") where this almost happens and thus leads to problems with empirically identifying some coefficients.

Remark 12.11 (*The within estimator and the Frisch-Waugh-Lovell theorem**) Regressing y_t and x_t on a set of dummies gives \bar{y}_i and \bar{x}_i . The FWL theorem says that regressing y_{it}^* on x_{it}^* gives the same slope coefficients as regressing y_{it} on x_{it} and the dummies.

Measures of fit (R^2) are often calculated as the squared correlation between the dependent variable and the fitted value. The R^2 from (12.26), $\text{Corr}(y_{it}^*, \hat{y}_{it}^*)^2$, is called the within- R^2 . Instead, if we add back the individuals means and calculate $\text{Corr}(y_{it}, \hat{y}_{it}^* + \bar{y}_i)^2$, then we have the “overall R^2 .”

Remark 12.12 (*Fixed effects in unbalanced panels*) When (y_{it}, x_{it}) include one or more missing values, then we typically exclude that observation from the estimation. For that reason, they should also be excluded from calculating (\bar{y}_i, \bar{x}_i) . In this way, (y_{it}^*, x_{it}^*) will have zero means in the sample that is used in the estimation. An alternative way of implementing this is (a) exclude those observation from estimating (\bar{y}_i, \bar{x}_i) ; (b) set the corresponding observations in (y_{it}^*, x_{it}^*) to zero.

Remark 12.13 (*Lagged dependent variable as regressor**) If $y_{i,t-1}$ is among the regressors x_{it} , then the within estimator (12.26) is biased in short samples (that is, when T is small)—and increasing the cross-section (that is, N) does not help. This is essentially the same issue as with estimating an AR(1) from a short sample.

12.5.1 The Within Estimator with Time Fixed Effects

When we allow for time fixed effects (but no individual fixed effects)

$$y_{it} = \lambda_t + x'_{it} \beta + u_{it}, \quad (12.28)$$

	LS	Fixed eff	Between	GLS	1st diff
exper/100	7.84 (8.25)	4.11 (6.21)	10.64 (4.05)	4.57 (7.12)	3.55 (2.33)
exper ² /100	-0.20 (-5.04)	-0.04 (-1.50)	-0.32 (-2.83)	-0.06 (-2.37)	-0.05 (-0.93)
tenure/100	1.21 (2.47)	1.39 (4.25)	1.25 (0.90)	1.38 (4.32)	1.29 (2.98)
tenure ² /100	-0.02 (-0.85)	-0.09 (-4.36)	-0.02 (-0.20)	-0.07 (-3.77)	-0.08 (-2.45)
south	-0.20 (-13.51)	-0.02 (-0.45)	-0.20 (-6.67)	-0.13 (-5.70)	-0.02 (-0.56)
union	0.11 (6.72)	0.06 (4.47)	0.12 (3.09)	0.07 (5.57)	0.04 (3.31)

Table 12.4: Panel estimation of log wages for women, $T = 5$ and $N = 716$ from NLS (1982,1983,1985,1987,1988). Example of fixed and random effects models, Hill et al (2008), Table 15.9. Numbers in parentheses are t-stats.

then we could once again introduce dummies (now for each time period) and apply pooled OLS. (There is no need for an intercept, since that would correspond to the average value of λ_t . Alternatively, introduce a constant and drop one time dummy.)

Figure 12.2 illustrates how omitted time varying (common) variables can affect the slope coefficient of included regressors, if there is a correlation.

As before, it is often easier to transform the data before estimating with pooled OLS.

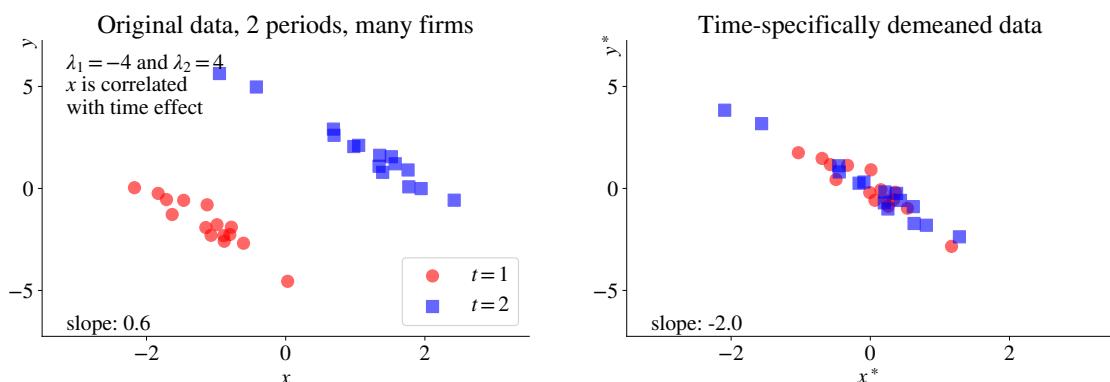


Figure 12.2: Interpretation of the FE estimator

In this case, we run the regression on transformed variables

$$y_{it}^* = x_{it}'\beta + u_{it}. \quad (12.29)$$

The transformations are

$$y_{it}^* = y_{it} - \bar{y}_t \quad (12.30)$$

$$x_{it}^* = x_{it} - \bar{x}_t, \quad (12.31)$$

where

$$\bar{x}_t = \sum_{i=1}^N x_{it}/N \text{ for each } t. \quad (12.32)$$

(Similarly for the transformation of y_{it} .)

As before, the estimation and testing of (12.29) is a pooled OLS, so the results on the variance-covariance matrix (clustering) apply.

Remark 12.14 (**Fixed effects in unbalanced panels*) As with individual fixed effects, the averages should only be calculated from those data points that will be used in the estimation.

12.5.2 The Within Estimator with Individual and Time Fixed Effects

When we allow for both time fixed effects and individual fixed effects

$$y_{it} = \lambda_t + \alpha_i + x_{it}'\beta + u_{it}, \quad (12.33)$$

then we could once again introduce dummies (now for both time periods and individuals) and apply pooled OLS. A double transformation of the data (similar to calculating y_{it}^* and x_{it}^* discussed before) works when the panel is balanced (see, for instance, Greene (2018)) 11). As this is often not the case, this section focuses on another approach.

We define T time dummies and make N copies of them all (yes, there is a reason for making those copies), and put them in a vector τ_{it} for individual i in period t . Then, consider the regression equation

$$y_{it} = \alpha_i + [x_{it}', \tau_{it}]\theta + u_{it}. \quad (12.34)$$

The time dummies have, at this stage, the same values for all individuals, but this will change in the next step. First, apply the “within transformation” (12.24)–(12.25) on this model. Notice that the transformation must be applied to each of $(y_{it}, x_{it}, \tau_{it})$ to create

$(y_{it}^*, x_{it}^*, \tau_{it}^*)$, that is, also to the time dummies (which now differ across individuals, if the panel is unbalanced). *Second*, apply the pooled regression of y_{it}^* on (x_{it}^*, τ_{it}^*) .

Note, however, that the number of regressors is $K + T$, so this approach described above might only work with a moderate number of time periods. An alternative approach, based on a repeated application of the Frisch-Waugh theorem is discussed in the remark below. It works with many time periods and also for three-way (or more) FE model.

Remark 12.15 (*L-way FE by repeated Frisch-Waugh**) *The following approach avoids a large regression (on many dummies) by applying a sequence of many smaller regressions. Let Z_{it}^1 be a set of dummies (say, for individuals), Z_{it}^2 another set of dummies (say, time periods), and so forth until Z_{it}^L . Actually, the ordering and grouping of the Z dummies is arbitrary, and it might be numerically preferable to just create groups of, say, 10 dummies. First, regress each variable in (y, x, Z^2, \dots, Z^L) on Z^1 and retrieve the residuals $(y^*, x^*, Z^{2*}, \dots, Z^{L*})$. (This step can be replaced by a within-transformation, which might be quicker. For instance, if $N > T$, then let Z^1 be individual dummies.) Second, regress each of the variables in $(y^*, x^*, Z^{3*}, \dots, Z^{L*})$ on Z^{2*} and retrieve the residuals $(y^{**}, x^{**}, Z^{3**}, \dots, Z^{L**})$. Continue this until only $(y^{****\dots}, x^{****\dots})$ are left. Finally, do a pooled regression on those. A special case is to let Z^1 be individual dummies and Z^2 time dummies ($L = 2$), to get the individual and time fixed effects model.*

12.6 The First-Difference Estimator

An another way of estimating the fixed effects model is to difference away the α_i by taking *first-differences* (in time)

$$\Delta y_{it} = \Delta \lambda_t + \Delta x'_{it} \beta + u_{it}^*, \quad (12.35)$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and similarly for the regressors. (Be careful not to take differences that involve different cross-sectional units.) Quite often, we interpret $\Delta \lambda_t$ as just a constant. Notice that

$$u_{it}^* = u_{it} - u_{i,t-1}, \quad (12.36)$$

so there are reasons to suspect that u_{it}^* is (negatively) autocorrelated.

Notice that the first-difference approach focuses on how changes in the regressors (over time, for the same individual) affect changes in the dependent variable. Also this method wipes out all regressors that are constant across time (for a given individual).

Regression (12.35) is once again estimated by pooled OLS. However, unadjusted standard errors are likely to overstate the uncertainty, because of the (possibly) negative autocorrelation of u_{it}^* . (Testing for autocorrelation is recommended.) This suggests that using the unadjusted standard errors constitutes a conservative approach (that is, making it harder to reject the null hypothesis). Again, the results on clustering discussed in Section 12.4.1 still apply.

Example 12.16 (*Negative autocorrelation of u_{it}^*) If u_{it} in (12.36) is iid, then $\text{Cov}(u_{it}^*, u_{it-1}^*) = \text{Cov}(u_{it} - u_{i,t-1}, u_{it-1} - u_{i,t-2})$, which equals $-\sigma_u^2$.

Remark 12.17 (*Lagged dependent variable as regressor**) If $y_{i,t-1}$ is among the regressors x_{it} , then the first-difference method (12.35) does not work (OLS is inconsistent and a larger sample does not help). The reason is that the (autocorrelated) residual is then correlated with the lagged dependent variable. This model cannot be estimated by OLS (the instrumental variable method might work).

12.7 Differences-in-Differences Estimator

Consider the model (12.35) when one of the regressors is a dummy variable indicating whether cross-sectional unit i was “treated” (for instance, received investment advise) in period t . We can estimate this as before—and interpret the coefficient as the effect of the “treatment” (conditional on all other variables)

In the classical difference-in-difference estimator there are only two periods ($T = 2$): before and after the treatment. This is

If there are no other regressors, then (12.35) can be written

$$\Delta y_{it} = \Delta \lambda_t + \beta Q_{it} + u_{it}^*, \quad (12.37)$$

where Q_{it} is the dummy variable. (The restriction that all individuals have the same $\Delta \lambda_t$ term is the so called “parallel trend assumption.”) In this case β can be estimated by the difference between the average Δy_{it} among the treated ($\Delta \bar{y}_{B2}$) and the average Δy_{it} among the non-treated ($\Delta \bar{y}_{A2}$)

$$\hat{\beta} = \Delta \bar{y}_{B2} - \Delta \bar{y}_{A2}. \quad (12.38)$$

(Notice that the change of the average is the same as the average of the change.)

More generally, consider the regression specification

$$y_{it} = \alpha_i + \kappa Q_i + \lambda_t + \delta T_t Q_i + x'_{it} \beta + \varepsilon_{it}, \quad (12.39)$$

where Q_i is a cross-sectional dummy variable (0 if i is non-treated and 1 otherwise) and T_t is a time-series dummy variable (0 before the treatment, 1 after) and x_{it} contains other regressors. In this specification, α_i , Q_i and the cross-sectional variation in x_{it} capture the differences across individuals that are not related to the treatment. In contrast, λ_t and the time-series variation in x_{it} capture changes over time that are also unrelated to the treatment. In contrast, $T_t Q_i$ captures the treatment effect, so δ is the key coefficient.

Suppose we only have two time periods (before and after the treatment), then the first difference of (12.39) gives

$$\Delta y_{it} = \underbrace{\Delta \alpha_i + \kappa \Delta Q_i}_{0} + \Delta \lambda_t + \delta \underbrace{\Delta(T_t Q_i)}_{Q_{it}} + \Delta x'_{it} \beta + u^*_{it}, \quad (12.40)$$

where Q_{it} is 0 for non-treated and 1 for treated (like in (12.37)). Notice that the “parallel trend assumption” now amounts to assuming that $\Delta \lambda_t + \Delta x'_{it} \beta$ is the same across the treated and non-treated. If this is questionable, then (12.37) should not be used. Rather, we should estimate (12.40).

12.8 Fama-MacBeth

The Fama and MacBeth (1973) approach (called FMB below) is a different method for handling panel data.

The method has two main steps, described below.

First, estimate λ_t and β_t

$$y_{it} = \lambda_t + x'_{it} \beta_t + u_{it} \quad (12.41)$$

period by period (using the cross section). Notice that this can easily handle unbalanced data sets (the cross-sectional regressions (12.41) are run on the available cross section for each time period).

Second, estimate the time averages

$$\hat{\beta} = \frac{1}{T} \sum_{t=1}^T \hat{\beta}_t. \quad (12.42)$$

Remark 12.18 (*Step 0**) The FMB can also be used to test CAPM (or other linear factor

models). In this case, y_{it} in (12.41) are the excess returns on asset i in period t (R_{it}^e) and x_{it} are the loadings (θ_{it}) of the excess return on the market excess return (or other factors) according to the regression $R_{it}^e = \alpha + f_t' \theta_{it} + \varepsilon_{it}$. (Often the θ coefficients are referred to as “betas”, which should not be confused with the coefficients β_t in (12.41).) In many cases, the θ_{it} values used as x_{it} are estimated during a previous sample, for instance, during the five years up to and including $t - 1$. In other cases, the θ_{it} values are estimated from the full sample, and are thus constant across periods. The latter has the advantage of being more precise estimates, provided the assumption of constant loadings is correct.

Fama and MacBeth (1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\beta}_t$. In particular, they suggest that the variance of $\hat{\beta}_t$ (notice, not $\hat{\beta}$) can be estimated by the (average) squared variation around its mean

$$\widehat{\text{Var}}(\hat{\beta}_t) = \frac{1}{T} \sum_{t=1}^T (\hat{\beta}_t - \hat{\beta})^2. \quad (12.43)$$

Since $\hat{\beta}$ is the sample average of $\hat{\beta}_t$, the variance of the former is the variance of the latter divided by T (the sample size)—provided $\hat{\beta}_t$ is iid. That is,

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{T} \text{Var}(\hat{\beta}_t) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\beta}_t - \hat{\beta})^2. \quad (12.44)$$

When x_{it} are common risk factors ($x_{it} = x_t$), then FMB and pooled OLS give the same point estimates (provided (12.41) is estimated without an intercept, effectively setting $\lambda_t = 0$). However, FMB’s variance-covariance matrix automatically handles the cross sectional correlations between residuals, while the pooled OLS would require applying a cluster method. See Cochrane (2005) 12.3 and Campbell, Lo, and MacKinlay (1997) 5.8 for further details.

Empirical Example 12.19 (*Estimated factor risk premia from different methods*) Table 12.5 shows estimates of the factor risk premia from several methods based on the 25 FF portfolios.

Further Reading

See Verbeek (2017) 10; Baltagi (2008), Greene (2018) 6.3 and 11, Hansen (2022a) 17-18 and Wooldridge (2010).

	Data	CR	FMB1	FMB2
Market	7.30 (2.15)	6.69 (2.22)	6.69 (2.18)	-7.24 (3.89)
SMB	1.18 (1.44)	1.06 (1.51)	1.06 (1.48)	0.66 (1.48)
HML	3.44 (1.45)	4.27 (1.60)	4.27 (1.48)	3.92 (1.48)

Table 12.5: Different estimates of factor risk premia, annualized %. Numbers in (parentheses) are standard deviations. The 25 FF portfolios 1970:01-2024:12. Data are the mean excess returns of the factors; CR are estimates of the factor risk premia from a cross-sectional regression; FMB1 are from Fama-MacBeth without intercept in the cross-sectional regression; FMB2 are from Fama-MacBeth with intercept in the cross-sectional regression. In both FMB regressions, the betas are estimated from the full sample.

12.9 Appendix – Random Effects Model*

The random effects model allows for *random* individual “intercepts” (μ_i)

$$y_{it} = \alpha + x'_{it}\beta + \mu_i + \varepsilon_{it}, \text{ where} \quad (12.45)$$

$$\varepsilon_{it} \text{ is iid } N(0, \sigma_\varepsilon^2) \text{ and } \mu_i \text{ is iid } N(0, \sigma_\mu^2). \quad (12.46)$$

Notice that μ_i is random (across agents) but constant across time, while ε_{it} is just random noise. Hence, μ_i can be interpreted as the permanent “luck” of individual i .

It is sometimes argued that the random effect only makes sense if the data is a sample from a larger population—and then captures the peculiar (relative to the population) features of the individuals that end up in the sample. It is then convenient to merge μ_i with ε_{it} , because it gives fewer parameters to estimate (and thus, saves degrees of freedom). In contrast, if the cross-section effectively contains the population (all mutual funds on a market, say), then a fixed effect is perhaps more reasonable.

Clearly, if we regard μ_i as non-random, then we are back in the fixed-effects model. (The choice between the two models is not always easy, so it may be wise to try both—and compare the results.)

We could write the regression as

$$y_{it} = \alpha + x'_{it}\beta + u_{it}, \text{ where } u_{it} = \mu_i + \varepsilon_{it}, \quad (12.47)$$

and we typically assume that u_{it} is uncorrelated across individuals, but correlated across time (only because of μ_i). In addition, we assume that ε_{jt} and μ_i are not correlated with each other or with x_{it} .

There are several ways to estimate the random effects model. First, the methods for fixed effects (the within and first-difference estimators) all work—so the “fixed effect” can actually be a random effect. Second, the *between estimator* using only individual time averages (from (12.23))

$$\bar{y}_i = \alpha + \bar{x}'_i\beta + \underbrace{\mu_i + \bar{\varepsilon}_i}_{\text{residual}_i}, \quad (12.48)$$

is also consistent, but discards all time-series information. Third, LS on

$$y_{it} = \alpha + x'_{it}\beta + \underbrace{\mu_i + \varepsilon_{it}}_{\text{residual}_{it}} \quad (12.49)$$

is consistent (but not really efficient). However, in this case we may need to adjust $\text{Var}(\hat{\beta})$

since the covariance matrix of the residuals is not diagonal.

In the random effects model, the μ_i variable can be thought of as an *excluded variable*. Excluded variables typically give a bias in the coefficients of all included variables—unless the excluded variable is uncorrelated with all of them. This is the assumption in the random effects model (recall: we assumed that μ_i is uncorrelated with x_{jt}). If this assumption is wrong, then we cannot estimate the RE model by either OLS or GLS, but the within-estimator (compare with the FE model) works, since it effectively eliminates the excluded variable from the system.

Example 12.20 $N = 2, T = 2$. If we stack the data for individual $i = 1$ first and those for individual $i = N$ second

$$\text{Var} \begin{pmatrix} u_{1,T-1} \\ u_{1T} \\ u_{N,T-1} \\ u_{NT} \end{pmatrix} = \begin{bmatrix} \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 & 0 & 0 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 \\ 0 & 0 & \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 \end{bmatrix},$$

which has elements off the main diagonal.

Remark 12.21 (*Generalized least squares*^{*}) GLS is an alternative estimation method that exploits correlation structure of residuals to increase the efficiency. In this case, it can be implemented by running OLS on

$$y_{it} - \vartheta \bar{y}_i = \alpha(1 - \vartheta) + (x_{it} - \vartheta \bar{x}_i)' \beta + v_{it}, \text{ where}$$

$$\vartheta = 1 - \sqrt{\sigma_u^2 / (\sigma_u^2 + T\sigma_\mu^2)}.$$

In this equation, σ_u^2 is the variance of the residuals in the “within regression” as estimated in (12.26) and $\sigma_\mu^2 = \sigma_B^2 - \sigma_u^2 / T$, where σ_B^2 is the variance of the residuals in the “between regression” (12.48). Here, σ_μ^2 can be interpreted as the variance of the random effect μ_i . However, watch out for negative values of σ_μ^2 and notice that when $\vartheta \approx 1$, then GLS is similar to the “within estimator” from (12.26). This happens when $\sigma_\mu^2 \gg \sigma_u^2$ or when T is large. The intuition is that when σ_μ^2 is large, then it is smart to get rid of that source of noise by using the within estimator, which disregards the information in the differences between individual means.

Chapter 13

Instrumental Variables Method (IV)

13.1 Instrumental Variables Method

When OLS is inconsistent, then we typically apply MLE or the instrumental variables (IV/2SLS) method. This section describes the latter.

We want to estimate β in

$$y_t = x_t' \beta + u_t, \quad (13.1)$$

where x_t and β are vectors with k elements. Recall that OLS is defined by making the fitted residuals orthogonal (uncorrelated) with the regressors

$$\mathbf{0}_{kx1} = \sum_{t=1}^T x_t (y_t - x_t' \hat{\beta}). \quad (13.2)$$

Example 13.1 (ARMA(1,1)) Consider the time series process $y_t = \rho y_{t-1} + \varepsilon_t$ where $\varepsilon_t = v_t + \theta v_{t-1}$. Notice that the regressor (y_{t-1}) is correlated with the residual (the v_{t-1} part), so OLS is inconsistent. See Figure 13.1 for an illustration.

The IV method replaces (13.2) by

$$\mathbf{0}_{kx1} = \sum_{t=1}^T z_t (y_t - x_t' \hat{\beta}_{iv}), \quad (13.3)$$

where z_t is a vector of k elements that have two key properties: (1) z_t is uncorrelated with the true residual (u_t) so z_t are *valid instruments*; but (2) correlated with the regressors (x_t) so z_t are *relevant instruments*. The first property cannot be directly checked since we never observe the true residuals. Instead, theoretical arguments must be employed, though the Hausman test may provide some assistance (see below). In particular, a valid instrument *cannot be endogenous* with respect to (that is, caused by) y_t and it *cannot be an erroneously excluded regressor*, because both cases would lead to $\text{Cov}(z_t, u_t) \neq 0$. A good application of the IV method should argue why that is not the case.

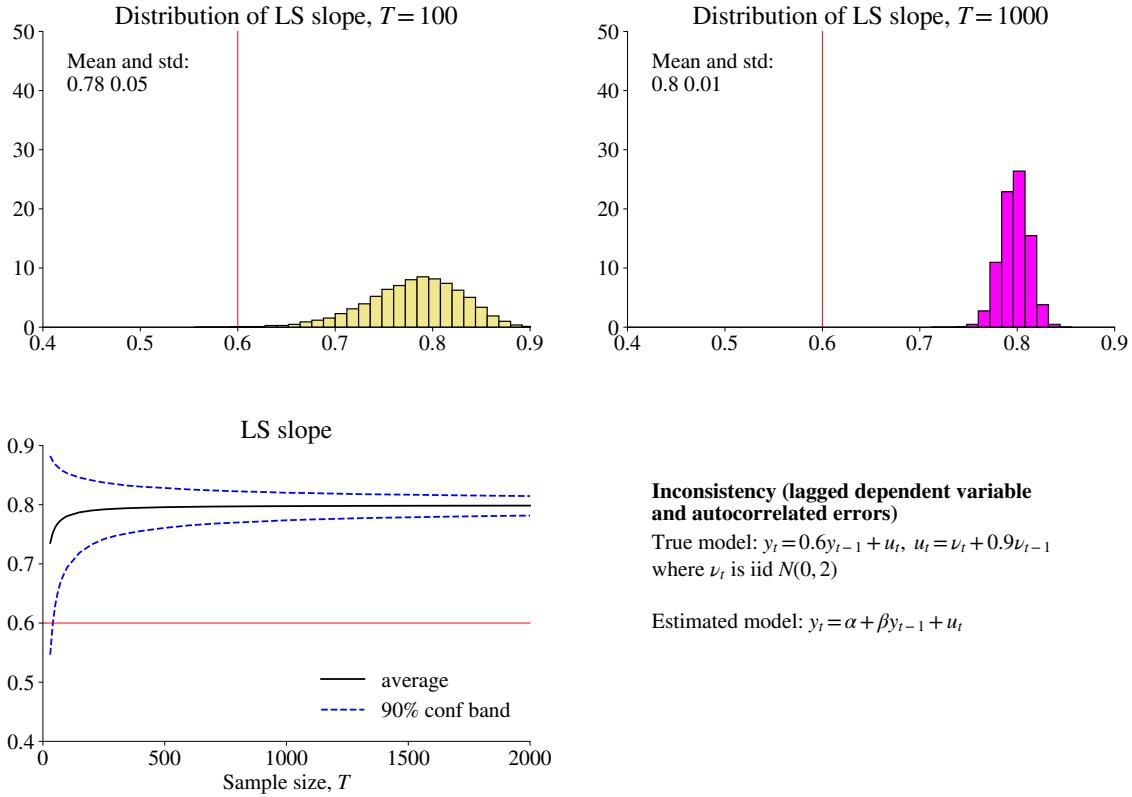


Figure 13.1: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

In contrast, the second property is easily checked by, for instance, regressing x_t on z_t and studying the t-statistics. This is discussed in some detail in the section on 2SLS below.

Example 13.2 (ARMA(1,1) continued) Continuing Example 13.1, notice that y_{t-2} (or earlier lags) are not correlated with the residual so they could be used as instruments.

Some regressors (elements of x_t) may also be used as instruments (z_t). For instance, if just one of the regressors is an endogenous variable then we need (at least one) new instrument for that regressor, while the other regressors can be instruments for themselves.

Solving (13.3) gives the IV estimator

$$\hat{\beta}_{iv} = \left(\sum_{t=1}^T z_t x_t' \right)^{-1} \sum_{t=1}^T z_t y_t. \quad (13.4)$$

Clearly, this is the same as OLS when $z_t = x_t$. That z_t are relevant instruments (z_t and x_t are correlated) means that the probability limit of $\Sigma z_t x_t'$ is invertible. There are few

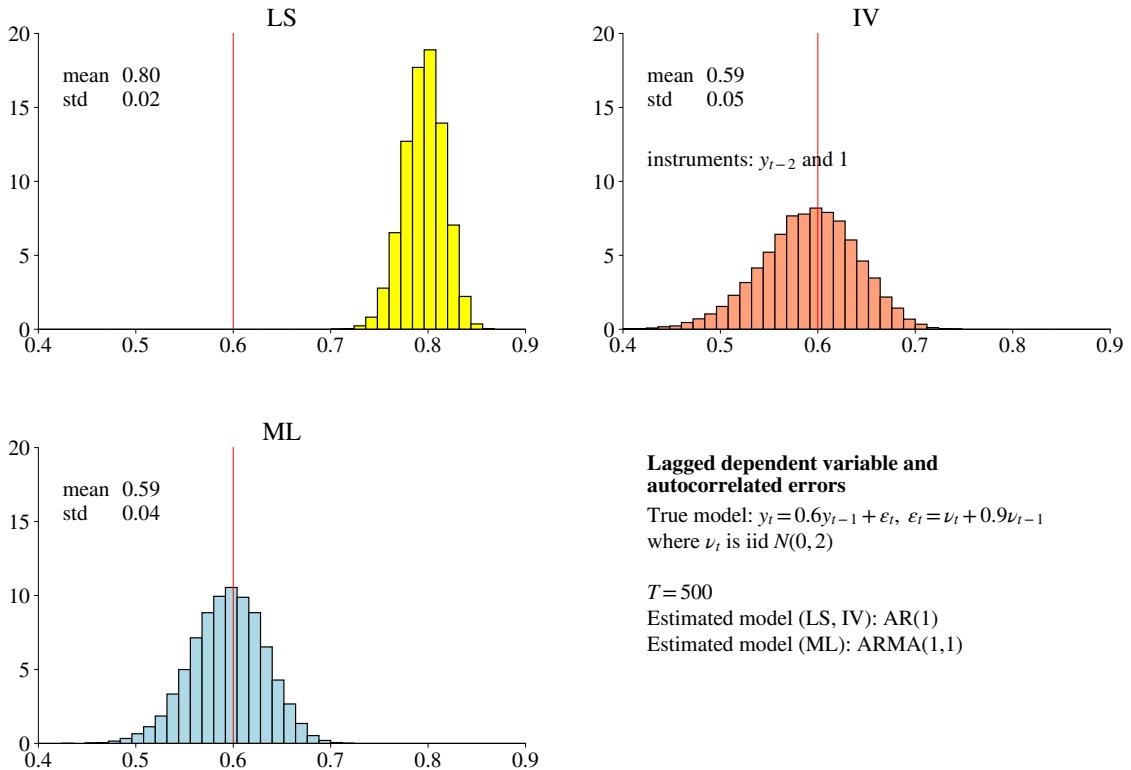


Figure 13.2: Results from a Monte Carlo experiment when data is from an ARMA process.

results on the small sample properties, although it is often found that there is small-sample bias.

Remark 13.3 (Matrix notation) Let z'_t be the t^{th} row of Z and similarly for X . We then have $\hat{\beta}_{iv} = (Z'X)^{-1}(Z'Y)$.

Figure 13.2 shows a simulation of an ARMA(1,1) process. The regressor (y_{t-1}) is correlated with the residual (v_{t-1}), so OLS is inconsistent. The IV method uses $(1, y_{t-2})$ as instruments for $(1, y_{t-1})$. Notice that $(1, y_{t-2})$ are indeed uncorrelated with the residual (which include shocks in t and $t - 1$ but not earlier), but correlated with the regressors (because of the persistence of the y_t series). (The figure also shows that the model can be estimated by maximum likelihood, but disregard that for now.)

$\hat{\beta}_{iv}$ is (asymptotically) normally distributed so

$$\hat{\beta}_{iv} \xrightarrow{a} N(\beta, V), \text{ with} \quad (13.5)$$

$$V = S_{zx}^{-1} S S_{xz}^{-1} \text{ where } S = \text{Var}(\Sigma_{t=1}^T z_t u_t)$$

and $S_{zx} = \sum_{t=1}^T z_t x'_t$. (See Appendix 13.4 for details.) This general expression is valid for both autocorrelated and heteroskedastic residuals—both features are loaded into the S matrix. In practice (with stochastic regressors and instruments), we estimate V by plugging in $S_{zx} = \sum_{t=1}^T z_t x'_t$ and an estimate of S .

We can estimate S by replacing u_t by fitted residuals

$$\hat{u}_t = y_t - x'_t \beta_{iv}. \quad (13.6)$$

If the residuals are iid and independent of z_t (so $S = \sigma^2 S_{zz}$), then

$$V = \sigma^2 S_{zx}^{-1} S_{zz} S_{xz}^{-1}, \text{ if } u_t \text{ are iid,} \quad (13.7)$$

where $S_{zz} = \sum_{t=1}^T z_t z'_t$. Otherwise, with non-iid residuals, (13.5) where S is estimated by the methods of White or Newey-West.

The IV estimator often has large standard errors, especially with “weak instruments” (weak correlation with regressors). This is seen in Figure 13.2 where the histogram for IV is much wider than that for OLS, and also in Figure 13.3 which shows how the choice of instruments affects the standard errors.

Example 13.4 ($\text{Var}(\hat{\beta}_{iv})$ in the simplest case) Assume y_t , x_t and z_t are zero mean variables and that z_t and u_t are independent. Equation (13.7) for a simple regression can then be written

$$\begin{aligned} \text{Var}(\hat{\beta}_{iv}) &= \frac{\sigma^2 \text{Var}(z_t)/T}{\text{Cov}(x, z)^2} \\ &= \frac{\sigma^2/T}{\text{Var}(x_t)} \frac{1}{\text{Corr}(x_t, z_t)^2}. \end{aligned}$$

If $\text{Corr}(x_t, z_t) = 1$ or -1 , then the result is the same as with OLS, although consistency can be questioned in this case. Instead, with a low $\text{Corr}(x_t, z_t)^2$ value (weak instruments), then the uncertainty is higher.

13.2 Two-stages-least squares (2SLS)

2SLS is the same as IV when there are as many instruments (L) as there are regressors (k). When there are more instruments than regressors ($L > k$), then 2SLS can produce more precise (efficient) estimates than IV. It proceeds in two steps.

Example 13.5 ($\text{ARMA}(2,1)$) $y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t$ where $\varepsilon_t = v_t + \theta v_{t-1}$. Notice

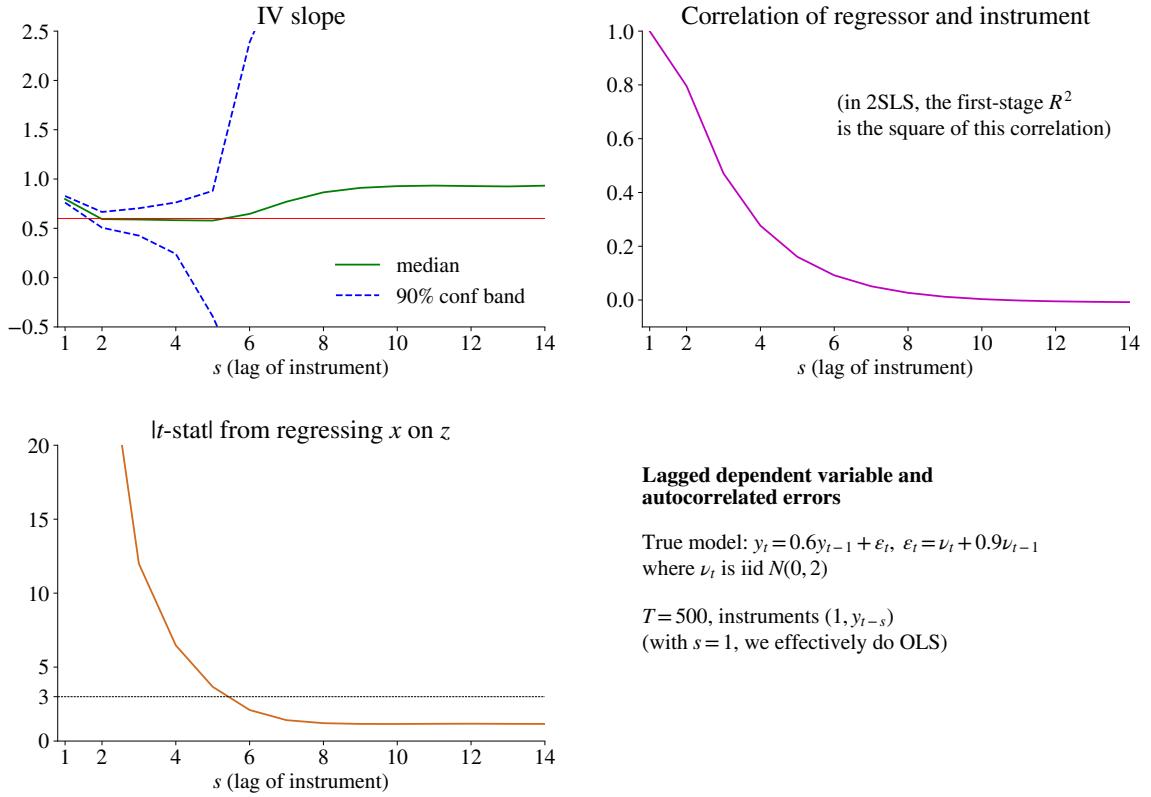


Figure 13.3: Results from a Monte Carlo experiment when data is from an ARMA process.

that y_{t-1} is correlated with ν_{t-1} but y_{t-2} is not. We could therefore use y_{t-2}, y_{t-3}, \dots as instruments for the two regressors. See Figure 13.4.

First, regress each of the elements in x_t on z_t

$$x_{it} = \delta_i' z_t + \varepsilon_t, \text{ for } i = 1 \text{ to } k, \quad (13.8)$$

where δ_i is a vector with L elements and let \hat{x}_{it} be the fitted values

$$\hat{x}_{it} = \delta_i' z_t. \quad (13.9)$$

(Clearly, $\delta_i' z_t = z_t' \delta_i$, but the former is more straightforward when we stack the equations below.) We can stack the equations as

$$\hat{x}_t = \delta' z_t, \quad (13.10)$$

where δ is an $L \times k$ matrix with δ_i in column i (and thus in row i of δ' , which is $k \times L$). The

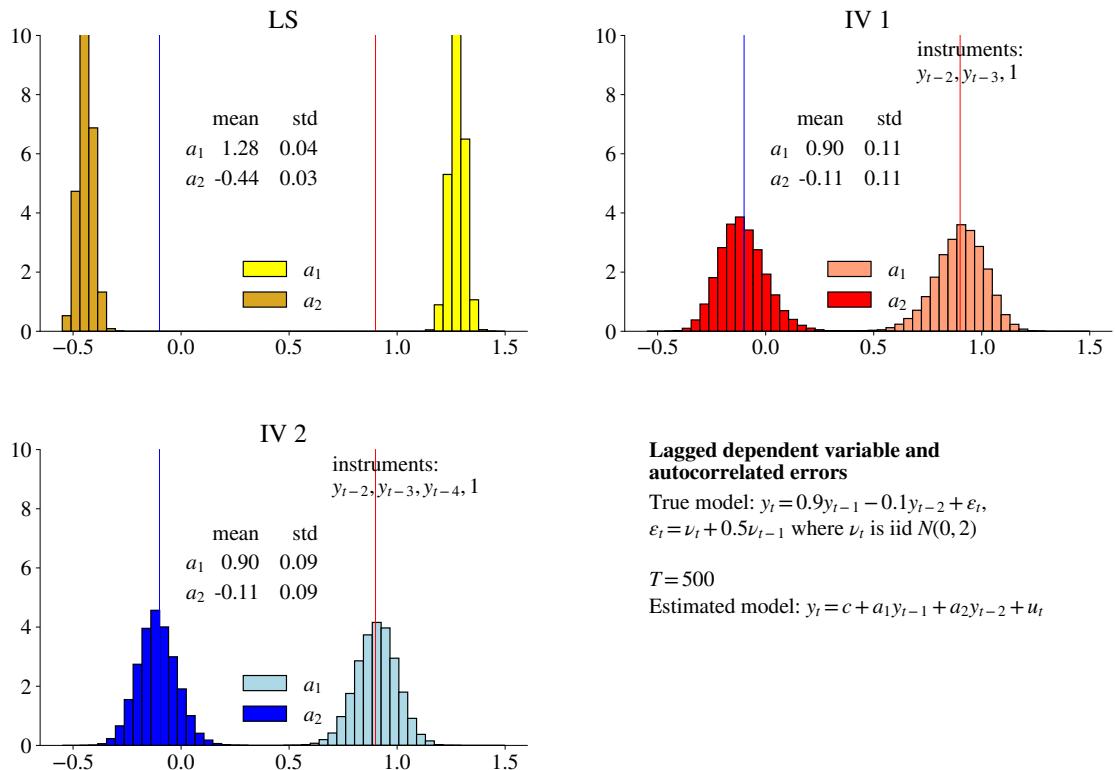


Figure 13.4: Results from a Monte Carlo experiment when data is from an ARMA process.

fit (or t-stats) of these regressions are often used to assess if the instruments are relevant. Often a $|t| > 3$ is considered necessary. See Figure 13.3.

Second, regress y_t on the fitted values \hat{x}_t

$$y_t = \hat{x}'_t \beta + v_t. \quad (13.11)$$

Remark 13.6 (Alternative to (13.11)*) We could equally well use \hat{x}_t instead of z_t in the IV estimator (13.4). This gives the same result as (13.11), provided that the instruments in the first stage estimation (13.8) include all “non-problematic” regressors.

Similarly to IV, the small sample properties are poor if the first-stage regression has a low R^2 (“weak instruments”), see Figure 13.3. However, using more instruments can improve precision; see Figure 13.4.

Example 13.7 (Supply and Demand) Consider the simplest simultaneous equations model

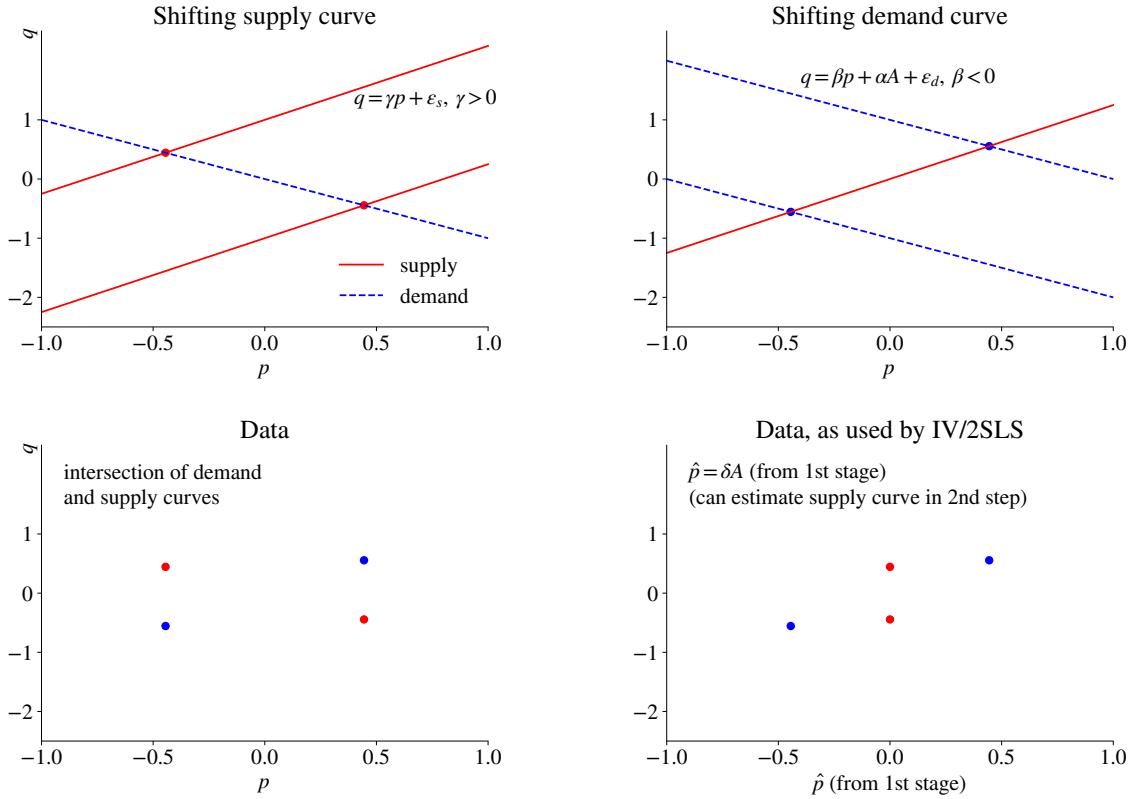


Figure 13.5: Illustration of demand and supply curves

for supply and demand on a market are

$$q_t = \gamma p_t + u_t^s, \gamma > 0 \text{ (supply)}$$

$$q_t = \beta p_t + \alpha A_t + u_t^d, \beta < 0 \text{ (demand)},$$

where A_t is an observable demand shock (perhaps income). To estimate the supply curve, the observable demand shocks A_t can be used as an instrument. See Figure 13.5 for an illustration.

The 2SLS approach highlights the key idea of IV (and 2SLS): in the regression (13.11) we only consider those movements in the regressors that are correlated with z_t (as captured by \hat{x}_t). Since z_t is chosen to be uncorrelated with the residuals, but correlated with x_t , we are only using the “clean” co-movements of x_t and y_t to estimate the coefficients. See Example 13.7 and Figure 13.5 for an illustration.

It can be shown (see Appendix 13.4 for details) that the (asymptotically valid) variance-

covariance matrix for $\hat{\beta}_{SLS}$ is

$$V = B S B', \text{ where} \\ B = (S_{xz} S_{zz}^{-1} S_{zx})^{-1} S_{xz} S_{zz}^{-1} \text{ and } S = \text{Var}(\Sigma_{t=1}^T z_t u_t).$$

This general expression is valid for both autocorrelated and heteroskedastic residuals since those features are loaded into the S matrix. We can estimate S as in the IV case: plug in (the sample values of) S_{xz} and S_{zz} and an estimate of S . For the latter, we replace u_t by the fitted residuals

$$\hat{u}_t = y_t - x_t' \beta_{iv}. \quad (13.13)$$

Notice that these are *not* the same as the fitted residuals from the 2nd stage regression. If the residuals are iid and independent of z_t (so $S = \sigma^2 S_{zz}$), then V simplifies to

$$V = \sigma^2 (S_{xz} S_{zz}^{-1} S_{zx})^{-1} \text{ if } u_t \text{ are iid.} \quad (13.14)$$

Empirical Example 13.8 (*Wage equation*) Tables 13.1 –13.2 shows results from an example in Hill, Griffiths, and Lim (2008) 10.3.3. The purpose is to estimate how log wages depend on education, experience and experience², while treating education as an endogenous variable. The instruments are experience, experience² and the education of the mother: see the first stage regression in 13.1. The result is fairly different from the OLS regression: see Table 13.2.

	1st stage
c	9.775 (23.753)
exper	0.049 (1.152)
exper ²	−0.001 (−0.959)
mothereduc	0.268 (8.481)
R^2	0.153
T	428

Table 13.1: First stage estimation of the 'educ' variable. Example of IV estimation, Hill et al (2008), section 10.3.3. Instruments: c, exper, exper², and mothereduc. Numbers in parentheses are t-stats (from White's method).

	OLS	IV/2SLS
c	-0.522 (-2.641)	0.198 (0.407)
educ	0.107 (7.634)	0.049 (1.301)
exper	0.042 (3.170)	0.045 (2.888)
exper ²	-0.001 (-2.073)	-0.001 (-2.145)
T	428	428

Table 13.2: IV estimation of wage equation. Example of IV estimation, Hill et al (2008), section 10.3.3. Instruments: c, exper, exper², and mothereduc. Numbers in parentheses are t-stats (from White's method).

Remark 13.9 (*Overidentifying restrictions in 2SLS**) When we use 2SLS, then we can test if instruments affect the dependent variable only via their correlation with the regressors. If not, something is wrong with the model since some relevant variables are excluded from the regression. A simple test is to first estimate with 2SLS to get the fitted residuals \hat{u}_t , then regress those on z_t . The TR^2 from this second regression is (under the null hypothesis) χ^2_{df} with df being the number of overidentifying restrictions.

13.3 Hausman's Specification Test

Reference: Greene (2018) 8.6

This test is investigating whether an efficient estimator (like LS) gives (approximately) the same estimate as a consistent estimator (like IV). If not, the efficient estimator is most likely inconsistent. It is therefore a way to test for the presence of endogeneity, measurement errors, etc.

Let $\hat{\beta}_e$ be an estimator that is consistent and asymptotically efficient when the null hypothesis, H_0 , is true, but inconsistent when H_0 is false (eg. LS). Let $\hat{\beta}_c$ be an estimator that is consistent under both H_0 and the alternative hypothesis (eg. IV). When H_0 is true, the asymptotic distribution is such that

$$\text{Cov}(\hat{\beta}_e, \hat{\beta}_c) = \text{Var}(\hat{\beta}_e). \quad (13.15)$$

Proof (of 13.15, univariate version*) Consider the estimator $\lambda\hat{\beta}_c + (1 - \lambda)\hat{\beta}_e$, which

is clearly consistent under H_0 since both $\hat{\beta}_c$ and $\hat{\beta}_e$ are. The asymptotic variance of this estimator is $\lambda^2 \text{Var}(\hat{\beta}_c) + (1 - \lambda)^2 \text{Var}(\hat{\beta}_e) + 2\lambda(1 - \lambda) \text{Cov}(\hat{\beta}_c, \hat{\beta}_e)$, which is minimized at $\lambda = 0$ (since $\hat{\beta}_e$ is asymptotically efficient). The first order condition with respect to λ , $2\lambda \text{Var}(\hat{\beta}_c) - 2(1 - \lambda) \text{Var}(\hat{\beta}_e) + 2(1 - 2\lambda) \text{Cov}(\hat{\beta}_c, \hat{\beta}_e) = 0$ should therefore be zero at $\lambda = 0$ so $\text{Var}(\hat{\beta}_e) = \text{Cov}(\hat{\beta}_c, \hat{\beta}_e)$. (See [Davidson \(2000\) 8.1](#)) \square

This means that we can write

$$\begin{aligned} \text{Var}(\hat{\beta}_e - \hat{\beta}_c) &= \text{Var}(\hat{\beta}_e) + \text{Var}(\hat{\beta}_c) - 2 \text{Cov}(\hat{\beta}_e, \hat{\beta}_c) \\ &= \text{Var}(\hat{\beta}_c) - \text{Var}(\hat{\beta}_e). \end{aligned} \quad (13.16)$$

We can use this to test, for instance, if the estimates from least squares ($\hat{\beta}_e$), and instrumental variable method ($\hat{\beta}_c$) are the same. In this case, H_0 is that the true residuals are uncorrelated with the regressors.

All we need for this test are the point estimates and consistent estimates of the variance-covariance matrices. Testing one of the coefficient can be done by a t test, and testing all the parameters by a χ^2 test

$$(\hat{\beta}_e - \hat{\beta}_c)' \text{Var}(\hat{\beta}_e - \hat{\beta}_c)^{-1} (\hat{\beta}_e - \hat{\beta}_c) \sim \chi_j^2, \quad (13.17)$$

where the covariance matrix is from (13.16) and where j equals the number of regressors that are potentially endogenous or measured with error. Note that the covariance matrix is likely to have a reduced rank, so the inverse needs to be calculated as a generalized (pseudo) inverse.

Further Reading

See [Verbeek \(2017\) 5](#), [Greene \(2018\) 8.3](#), [Hamilton \(1994\) 9.2](#), [Pindyck and Rubinfeld \(1998\) 7](#), and [Hansen \(2022a\) 12](#).

13.4 Appendix – Asymptotics of the IV and 2SLS Estimators*

This section gives some details of the asymptotic properties of IV and 2SLS.

13.4.1 Asymptotic Results on the IV Estimator*

There are few results on small sample properties, but IV is often imprecise and even biased. In large samples, we typically get consistency and a normal distribution.

Use (13.1) to substitute for y_t in (13.4), multiply both sides by \sqrt{T} and rearrange as

$$\sqrt{T}(\hat{\beta}_{iv} - \beta) = \hat{\Sigma}_{zx}^{-1}\sqrt{T}\sum_{t=1}^T z_t u_t / T, \quad (13.18)$$

where $\hat{\Sigma}_{zx} = \sum_{t=1}^T z_t x'_t / T$. Notice that $\hat{\Sigma}_{zx}$ is a symbol for a matrix, not for a summation. Divide by \sqrt{T} to show that $\hat{\beta}_{iv}$ is consistent if $\text{Cov}(z_t, u_t) = 0$.

Since $\sqrt{T}\sum_{t=1}^T z_t u_t / T$ in (13.18) is $\sqrt{T} \times$ a sample average, it is plausible that a CLT applies so the asymptotic distribution, so $\sqrt{T}(\hat{\beta}_{iv} - \beta)$ might be normal with a zero mean and a variance-covariance matrix

$$\text{Var}(\sqrt{T}\hat{\beta}_{iv}) = \Sigma_{zx}^{-1} \Sigma \Sigma_{xz}^{-1}, \text{ where } \Sigma = \text{Var}\left(\sum_{t=1}^T z_t u_t / \sqrt{T}\right). \quad (13.19)$$

and where Σ_{zx} is the probability limit of $\hat{\Sigma}_{zx}$. The last matrix in the covariance matrix follows from $(\Sigma_{zx}^{-1})' = (\Sigma_{zx}')^{-1} = \Sigma_{xz}^{-1}$. Dividing both sides by T and rewriting gives (13.5). (Details: Σ_{zx} is the probability limit of S_{zx}/T and $\Sigma = S/T$. Use this in (13.19) and simplify to get the probability limit of $TS_{zx}^{-1}SS_{xz}^{-1}$. Divide both sides by T to get $\text{Var}(\hat{\beta}_{iv})$ which then equals (13.5).)

13.4.2 Asymptotic Results on 2SLS*

From (13.8)–(13.9) we have

$$\hat{\delta} = \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}, \text{ and } \hat{\beta} = \hat{\Sigma}_{\hat{x}\hat{x}}^{-1} \hat{\Sigma}_{\hat{x}y}, \quad (13.20)$$

where $\hat{\Sigma}_{zz} = \sum_{t=1}^T z_t z'_t / T$, and so forth. Notice that $\hat{\Sigma}_{zz}^{-1}$ is an $L \times L$ matrix and $\hat{\Sigma}_{zx}$ is an $L \times k$ matrix, so $\hat{\delta}$ is $L \times k$. The fitted values in (13.10) can then be written

$$\hat{x}_t = \hat{\delta}' z_t = \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} z_t, \quad (13.21)$$

so

$$\begin{aligned}\hat{\Sigma}_{\hat{x}\hat{x}} &= \sum_{t=1}^T \hat{x}_t \hat{x}'_t / T = \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zx} \text{ and} \\ \hat{\Sigma}_{\hat{x}y} &= \sum_{t=1}^T \hat{x}_t y_t / T = \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zy}.\end{aligned}\quad (13.22)$$

(Substitute for \hat{x} from (13.21) and simplify to derive this.)

Using these results in (13.20) gives

$$\hat{\beta} = (\hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zx})^{-1} \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zy}. \quad (13.23)$$

Substituting for y_t by using (13.1) and expanding gives

$$\begin{aligned}\hat{\beta} &= (\hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zx})^{-1} \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \sum_{t=1}^T z_t (x'_t \beta + u_t) / T \\ &= \beta + \underbrace{(\hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zx})^{-1} \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1}}_A \sum_{t=1}^T z_t u_t.\end{aligned}\quad (13.24)$$

Consistency follows from $\text{plim } \Sigma_{t=1}^T z_t u_t / T = 0$ and asymptotic normality from a CLT applied to $\sqrt{T} \Sigma_{t=1}^T z_t u_t / T$ and the asymptotic variance-covariance matrix is

$$\text{Var}(\sqrt{T} \hat{\beta}_{iv}) = A \Sigma A', \text{ where } \Sigma = \text{Var}(\Sigma_{t=1}^T z_t u_t / \sqrt{T}) \quad (13.25)$$

This can be rewritten as (13.12). (Details: Σ_{zx} is the probability limit of S_{zx} / T etc and $\Sigma = S / T$. Divide both sides by T .)

Remark 13.10 (*Alternative expression for A) By using (13.20), A in (13.24) can also be written $A = (\hat{\delta}' \Sigma_{zz} \hat{\delta})^{-1} \hat{\delta}'$.

Chapter 14

GMM

14.1 The Basic GMM

The Generalized Method of Moments (GMM) can be used to construct a new estimator when there is none suitable, but also to embed classical methods. In fact, OLS, IV and MLE are special cases of GMM.

The key steps in GMM are to (1) define a number of “moment conditions” which relates data in period t to model parameters; (2) calculate the sample averages of those moment conditions; (3) choose parameter estimates to minimize a loss function in terms of those sample averages. The intuition for the approach is to make the properties (moments) of the data as similar as possible to those implied by the model.

GMM is typically consistent, provided the moment conditions are correctly defined. The standard errors for GMM estimates are based on asymptotic results related to the Delta method.

In general, the $q \times 1$ vector of sample moment conditions in GMM are written

$$\bar{g}(\beta) = \sum_{t=1}^T g_t(\beta)/T, \quad (14.1)$$

where $\bar{g}(\beta)$ is short hand notation for the sample average. The notation $g_t(\beta)$ is meant to show that the moment conditions depend on the parameter vector β and on data for period t . We let β denote the true value of the $k \times 1$ parameter vector.

Example 14.1 (*Moment condition for estimating the mean*) With $g_t(\mu) = x_t - \mu$, we have $\bar{g}(\mu) = \sum_{t=1}^T (x_t - \mu)/T$.

The GMM estimate $\hat{\beta}_{k \times 1}$ minimizes the loss function

$$\bar{g}(b)' W \bar{g}(b), \quad (14.2)$$

where W is some symmetric positive definite $q \times q$ weighting matrix.

Remark 14.2 (*Positive definite matrix*) *The $n \times n$ matrix A is positive definite if for any non-zero $n \times 1$ vector x , $x'Ax > 0$. Such a matrix has a positive determinant, all diagonal elements are positive and also A^{-1} is positive definite. For instance, a covariance matrix of random variables is positive definite, unless there perfect collinearity.*

When the model is *exactly identified* (there are as many parameters as there are moment conditions, $q = k$), then we do not have to perform an explicit minimization, since all sample moment conditions can be set equal to zero

$$\bar{g}(\hat{\beta}) = \mathbf{0}_{q \times 1} \quad (14.3)$$

This will clearly minimize (14.2) since W is positive definite. However, we may still have to apply a numerical algorithm to find the $\hat{\beta}$ values that make (14.3) hold, in particular, if $g_t()$ are non-linear functions.

Example 14.3 (*Estimating the mean*) *With the moment condition in Example 14.1, $\Sigma_{t=1}^T (x_t - \hat{\mu})/T = 0$ gives the traditional sample average, $\hat{\mu} = \Sigma_{t=1}^T x_t / T$.*

In contrast, when the model is *overidentified* (there are more moment conditions than parameters, $q > k$), then we need to explicitly minimize (14.2). This often involve a numerical optimization routine or solving the first order conditions, $(\partial \bar{g}(b)/\partial b')' W \bar{g}(b) = \mathbf{0}_{k \times 1}$.

It can be shown that choosing $W = \Sigma^{-1}$, where Σ is the covariance matrix of $\sqrt{T}\bar{g}(\beta)$ evaluated at the true parameter values, gives the most efficient estimates (for a given set of moment conditions). To approximate this, an iterative procedure is often used: start with $W = I_q$ (or another reasonable weighting matrix), estimate the parameters and use them to create a $T \times q$ matrix of moment conditions, estimate Σ (see below for a discussion), then (in a second step) use $W = \hat{S}_0^{-1}$ and reestimate. In most cases this iteration is stopped at this stage, but you could also continue iterating until the point estimates converge.

Example 14.4 (*Moment conditions for estimating a normal distribution*) *Suppose you specify four moments for estimating the mean and variance ($\beta = (\mu, \sigma^2)$) of a normal distribution*

$$g_t = \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix}$$

This case is overidentified ($q = 4$ and $k = 2$), so a weighting matrix is needed. Instead, if we only use the first two moment conditions, then the model is exactly identified.

Example 14.5 (Moments conditions for OLS) Consider the linear model $y_t = x_t' \beta + u_t$, where x_t and β are $k \times 1$ vectors. The k moments are $g_t = x_t(y_t - x_t' \beta)$. There are as many parameters as moment conditions: exactly identified.

Empirical Example 14.6 (Estimating the mean and variance with GMM) Table 14.1 reports estimates of the mean and variance of the FF equity market return. The first approach (column) uses only the first two moment conditions of Example 14.4 (an exactly identified case). Other approaches apply different (suboptimal) W matrices in solving (14.2). (The columns marked $A_1\bar{g}$ and $A_2\bar{g}$ are discussed below.) Table 14.2 reports results from iterating on the W matrix when solving (14.2). The W matrix used in the final iteration is shown in Table 14.3.

	trad.	ex. ident.	$\bar{g}' W_1 \bar{g}$	$\bar{g}' W_2 \bar{g}$	$A_1 \bar{g}$	$A_2 \bar{g}$
μ	0.61	0.61	0.61	0.38	0.61	0.56
σ^2	21.12	21.12	21.12	21.17	21.12	21.12

Table 14.1: Estimates of mean and variance of the FF equity market factor, 1970:01-2024:12. The W_1 and A_1 matrices put equal weights on moment conditions 1–2 and zero weight moment conditions 3–4, while W_2 and A_2 put also a very small weight on moment condition 3.

	iteration				
	0	1	2	3	4
μ	0.61	0.74	0.75	0.75	0.75
σ^2	21.12	19.00	18.94	18.94	18.94

Table 14.2: Estimates of mean and variance of the FF equity market factor, 1970:01-2024:12. The estimates minimize $\bar{g}' W_i \bar{g}$, where W_i is the inverse of the variance-covariance matrix of the moment conditions from the previous iteration.

14.1.1 Distribution of the Basic GMM

GMM estimators are typically asymptotically normally distributed, with a covariance matrix that depends on the covariance matrix of the moment conditions (Σ) and the

	g_1	g_2	g_3	g_4
g_1	1215.63	58.14	-9.80	-0.41
g_2	58.14	18.98	-0.63	-0.07
g_3	-9.80	-0.63	0.14	0.01
g_4	-0.41	-0.07	0.01	0.00

Table 14.3: $W_i \times 10000$ used in the last iteration when, minimizing $\bar{g}' W_i \bar{g}$ to estimate the mean and variance of the FF equity market factor, 1970:01-2024:12.

mapping from the parameters to the moment conditions (D). This section first discusses Σ and D , and then combines the results to get the variance-covariance matrix of $\hat{\beta}$.

The Covariance Matrix

Let Σ be the $(q \times q)$ covariance matrix of $\sqrt{T}\bar{g}(\beta)$, evaluated at the true parameter values

$$\Sigma = \text{Var}[\sqrt{T}\bar{g}(\beta)], \quad (14.4)$$

where $\text{Var}()$ is a variance-covariance matrix. For instance, when there is no autocorrelation in the moments, then $\Sigma = \text{Var}[g_t(\beta)]$, but when there is autocorrelation, then we may use the Newey-West approach to estimate Σ . In practice, Σ is estimated by using the estimated coefficients in the moments to get the data series $g_t(\hat{\beta})$, a $T \times q$ matrix, from which we estimate the covariances needed for (14.4).

Remark 14.7 (*Variance-covariance of $\sqrt{T}\bar{g}(\beta)$ or of $\Sigma_{t=1}^T g_t$?) Previous chapters have presented the Newey-West approach to estimate the variance-covariance of $\Sigma_{t=1}^T g_t$, which we denoted $\hat{\Sigma}$. Since $\sqrt{T}\bar{g}(\beta) = \Sigma_{t=1}^T g_t / \sqrt{T}$, we clearly have $\hat{\Sigma} = \hat{S} / T$.

Example 14.8 (OLS, covariance) For the moments in Example 14.5 we have $\Sigma = \text{Var}(\sqrt{T}\Sigma_{t=1}^T x_t u_t / T)$, since $u_t = y_t - x_t' \beta$.

The Derivatives

Let D be the $(q \times k)$ probability limit of the Jacobian (transpose of the gradients of each element in $\bar{g}(\beta)$) of the sample moment conditions with respect to the parameters (also evaluated at the true parameters)

$$D = \text{plim} \frac{\partial \bar{g}(\beta)}{\partial \beta'}. \quad (14.5)$$

In practice, the Jacobian D is approximated by using the point estimates and the available sample of data. Modern software has fast and accurate routines for numerical calculations of derivatives of the moment conditions, reducing the need for explicitly coding the derivatives.

Remark 14.9 (*Jacobian*) The Jacobian is of the following format

$$\frac{\partial \bar{g}(\beta)}{\partial \beta'} = \begin{bmatrix} \frac{\partial \bar{g}_1(\beta)}{\partial \beta_1} & \dots & \frac{\partial \bar{g}_1(\beta)}{\partial \beta_k} \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_q(\beta)}{\partial \beta_1} & \dots & \frac{\partial \bar{g}_q(\beta)}{\partial \beta_k} \end{bmatrix}$$

Example 14.10 (*Estimating/testing a normal distribution, covariance*) For the moments in Example 14.4 we have

$$D = -\text{plim} \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & 0 \\ 2(x_t - \mu) & 1 \\ 3(x_t - \mu)^2 & 0 \\ 4(x_t - \mu)^3 & 6\sigma^2 \end{bmatrix} = -\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 3\sigma^2 & 0 \\ 4\mu_3 & 6\sigma^2 \end{bmatrix},$$

where μ_3 is the third centered moment (which is zero for a symmetric distribution). If we only use the first two moment conditions, then $D = -I_2$.

Example 14.11 (*OLS, Jacobian*) For the moments in Example 14.5 $D = \text{plim}(-\sum_{t=1}^T x_t x_t' / T) = -\Sigma_{xx}$. This does not contain any parameters, but data. In practice, we replace Σ_{xx} by a sample estimate.

The Distribution of $\hat{\beta}$

The distribution of the GMM estimates is

$$\begin{aligned} \sqrt{T}(\hat{\beta} - \beta) &\xrightarrow{d} N(0, V) \text{ if } W = \Sigma^{-1}, \text{ where} \\ V &= (D' \Sigma^{-1} D)^{-1}, \end{aligned} \tag{14.6}$$

provided we have used Σ^{-1} as the weighting matrix ($W = \Sigma^{-1}$) in (14.2). (See Appendix 14.4 for a proof.)

The choice of the weighting matrix is irrelevant if the model is exactly identified, so (14.6) can be applied to this case (even if we did not specify any weighting matrix at all). It can also be noticed that when the model is exactly identified, then we can typically

rewrite the covariance matrix as $V = D^{-1}\Sigma(D^{-1})'$, which might be easier to calculate. (The case of using another W matrix is discussed below.)

Most GMM analysis use expressions for the distribution of $\sqrt{T}(\hat{\beta} - \beta)$, so we follow that convention here. However, we can also write the asymptotic distribution as

$$\hat{\beta} \xrightarrow{a} N(\beta, V/T), \quad (14.7)$$

where \xrightarrow{a} means “is asymptotically distributed as.”

Example 14.12 (OLS, distribution) For the moment conditions in Example 14.5 $V = (\Sigma_{xx}\Sigma^{-1}\Sigma_{xx})^{-1}$. Under the assumption of iid residuals, $\Sigma = \sigma^2\Sigma_{xx}$ so $V = \sigma^2\Sigma_{xx}^{-1}$.

Empirical Example 14.13 (Standard errors) The standard errors for the estimations in Example 14.6 are given in Table 14.4.

	trad.	ex. ident.	$\bar{g}'W_1\bar{g}$	$\bar{g}'W_2\bar{g}$	$A_1\bar{g}$	$A_2\bar{g}$
Std(μ)	0.18	0.18	0.18	0.23	0.18	0.19
Std(σ^2)	1.16	1.70	1.70	1.72	1.70	1.70

Table 14.4: Standard errors of the mean and variance of the FF equity market factor. See Table 14.1 for details.

Remark 14.14 (Test of overidentifying restrictions*) A CLT applies to $\sqrt{T}\bar{g}(\beta)$, but there are effectively only $q - k$ nondegenerate random variables since (a linear combination of) k moment conditions are always set to zero by the first order conditions (see Davidson and MacKinnon (1993) 17.6 for a detailed discussion). When $W = \Sigma^{-1}$, we can thus test the hypothesis that $\bar{g}(\beta) = 0$ by the J test, $T\bar{g}(\hat{\beta})'\Sigma^{-1}\bar{g}(\hat{\beta}) \xrightarrow{d} \chi_{q-k}^2$. Notice that this is a χ^2 distribution with $q - k$ degrees of freedom.

Example 14.15 (Non-linear least squares) Consider the non-linear regression $y_t = F(x_t; \beta) + u_t$, where $F(x_t; \beta)$ is a potentially non-linear equation of the regressors x_t , with a k -vector of parameters β . The non-linear least squares (NLS) approach is minimize the sum of squared residuals, that is, to choose b to minimize $\sum_{t=1}^T [y_t - F(x_t; b)]^2$. Use the first order conditions as moment conditions

$$\bar{g}(\beta) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial F(x_t; \beta)}{\partial \beta} [y_t - F(x_t; \beta)].$$

The model is exactly identified so the point estimates are found by setting all moment conditions to zero, $\bar{g}(\hat{\beta}) = \mathbf{0}_{k \times 1}$. As usual, $S = \text{Cov}[\sqrt{T}\bar{g}(\beta)]$, while the Jacobian is

$$D = \text{plim} \frac{1}{T} \sum_{t=1}^T \frac{\partial F(x_t; \beta)}{\partial \beta} \frac{\partial F(x_t; \beta)}{\partial \beta'} - \text{plim} \frac{1}{T} \sum_{t=1}^T [y_t - F(x_t; \beta)] \frac{\partial^2 F(x_t; \beta)}{\partial \beta \partial \beta'}.$$

For instance, when $F(x_t; \beta) = x_t' \beta$, then $\partial F(x_t; \beta)/\partial \beta = x_t$, so the moment conditions are $\bar{g}(\beta) = -\sum_{t=1}^T x_t(y_t - x_t' \beta)/T$. Since the second derivatives are zero, we get $D = \text{plim} \sum_{t=1}^T x_t x_t'/T$, which is the same in the LS case.

14.2 GMM with a Suboptimal Weighting Matrix

The distribution of the GMM estimates when we use a sub-optimal weighting matrix is similar to (14.6), but the variance-covariance matrix is different (basically, reflecting the fact that the approach does not produce the lowest possible variances anymore).

Example 14.16 (*Estimating/testing a normal distribution*) Example 14.4 is overidentified since there are four moment conditions but only two parameters. Instead of using the optimal weighting matrix, we could use any other symmetric positive definite 4×4 matrix. For instance, $W = I_4$ or a matrix that puts almost all weight on the first two moment conditions.

It can be shown (see Appendix 14.4) that if we use another weighting matrix than $W = \Sigma^{-1}$, then the V matrix in (14.6) and (14.7) should be changed to

$$V_2 = V_{A2} D' W \Sigma W D V_{A2}', \text{ where } V_{A2} = (D' W D)^{-1}. \quad (14.8)$$

Remark 14.17 (*Test of overidentifying restrictions**) The test is now $T \bar{g}(\hat{\beta})' \Psi_2^+ \bar{g}(\hat{\beta}) \xrightarrow{d} \chi_{q-k}^2$, where Ψ_2^+ is a generalized (pseudo) inverse of $\Psi_2 = \Psi_{A2} \Sigma \Psi_{A2}'$, where $\Psi_{A2} = I_q - D (D' W D)^{-1} D' W$. The variance-covariance matrix Ψ_2 has a reduced rank, so we must use a generalized inverse.

14.3 GMM without a Loss Function

Suppose we bypass the optimization process and directly specify k linear combinations of the q moment conditions

$$\mathbf{0}_{k \times 1} = \underbrace{A}_{k \times q} \underbrace{\bar{g}(\hat{\beta})}_{q \times 1}, \quad (14.9)$$

where the matrix A is chosen by the researcher. We can solve (possibly with a numerical algorithm) for the $\hat{\beta}$ values that make these equations hold.

Example 14.18 (*Overidentified example: estimating/testing a normal distribution*) Example 14.4 is overidentified since there are four moment conditions but only two parameters. One possible A matrix would put all weight on the first two moment conditions

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Empirical Example 14.19 (*Estimating a mean and a variance with GMM*) Table 14.1 reports estimates of the mean and variance of the FF equity market return. The columns marked $A_1\bar{g}$ and $A_2\bar{g}$ use different A matrices.

It is straightforward to show (see Appendix 14.4) that the V matrix in (14.6) and (14.7) should be changed to

$$V_3 = V_{A3} A \Sigma A' V'_{A3}, \text{ where } V_{A3} = (AD)^{-1}. \quad (14.10)$$

and where A should be understood as a probability limit.

Example 14.20 (*Estimating/testing a normal distribution*) Continuing Example 14.18, we have that AD in (14.10) is

$$V_{A3} = \left(\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_A \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_D \right)^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Remark 14.21 (*Test of overidentifying restrictions**) The test is the same as in Remark 14.17, but with Ψ_2 replaced by $\Psi_3 = \Psi_{A3} \Sigma \Psi'_{A3}$, where $\Psi_{A3} = I_q - D (AD)^{-1} A$.

Further Reading

See Greene (2018) 13, Hamilton (1994) 14 and Hansen (2022a) for general treatment of GMM. Also, see Cochrane (2005) 11 and 14, Campbell (2018) 4, Singleton (2006) 2–4 for further details and financial applications.

14.4 Appendix – Proofs

Proof (The asymptotic distribution (14.6), (14.8) or (14.10)) By the mean-value theorem the sample moment condition evaluated at the GMM estimate, $\hat{\beta}$, is

$$\bar{g}(\hat{\beta}) = \bar{g}(\beta) + \frac{\partial \bar{g}(\tilde{\beta})}{\partial \beta'} (\hat{\beta} - \beta) \quad (14.11)$$

for some values $\tilde{\beta}$ between $\hat{\beta}$ and β . Premultiply with $[\partial \bar{g}(\hat{\beta}) / \partial \beta']' W$. By the first order conditions, $(\partial \bar{g}(b) / \partial b')' W \bar{g}(b) = \mathbf{0}_{k \times 1}$, the left hand side is then zero. Multiply with \sqrt{T} and solve $\sqrt{T}(\hat{\beta} - \beta)$ as

$$\sqrt{T}(\hat{\beta} - \beta) = - \underbrace{\left[\left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'} \right)' W \frac{\partial \bar{g}(\tilde{\beta})}{\partial \beta'} \right]^{-1} \left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'} \right)' W \sqrt{T} \bar{g}(\beta)}_{\Gamma}. \quad (14.12)$$

If $\text{plim } \partial \bar{g}(\hat{\beta}) / \partial \beta' = \partial \bar{g}(\beta) / \partial \beta' = D$, then $\text{plim } \partial \bar{g}(\tilde{\beta}) / \partial \beta' = D$, since $\tilde{\beta}$ is between β and $\hat{\beta}$. Then $\text{plim } \Gamma = -(D' W D)^{-1} D' W$. The last term in (14.12), $\sqrt{T} \bar{g}(\beta)$, is \sqrt{T} times a vector of sample averages, so by a CLT it converges in distribution to $N(0, \Sigma)$, where Σ is defined as in (14.4). By the continuous mapping theorem we then have that

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N\left[\mathbf{0}_{k \times 1}, (\text{plim } \Gamma) \Sigma (\text{plim } \Gamma')\right].$$

The variance-covariance matrix is

$$\text{Var}[\sqrt{T}(\hat{\beta} - \beta)] = (D' W D)^{-1} D' W \Sigma W D (D' W D)^{-1},$$

where we use the facts that W and $D' W D$ are symmetric matrices. If $W = \Sigma^{-1}$, this expression simplifies to (14.6). Otherwise, it is the same as (14.8). To get (14.10), substitute A for $D' W$. See, for instance, Hamilton (1994) 14 (appendix) for more details.

□

Proof (Proof of Remark 14.17 and 14.21) Multiply (14.11) by \sqrt{T} and (14.12) to substitute for $\sqrt{T}(\hat{\beta} - \beta)$

$$\sqrt{T} \bar{g}(\hat{\beta}) = \sqrt{T} \bar{g}(\beta) + \sqrt{T} \frac{\partial \bar{g}(\tilde{\beta})}{\partial \beta'} \Gamma \bar{g}(\beta) = \left[I + \frac{\partial \bar{g}(\tilde{\beta})}{\partial \beta'} \Gamma \right] \sqrt{T} \bar{g}(\beta_0).$$

The term in brackets has a probability limit of $I - D (D' W D)^{-1} D' W$ (see below (14.12)). Since $\sqrt{T} \bar{g}(\beta_0) \xrightarrow{d} N(\mathbf{0}_{q \times 1}, \Sigma)$ we get Remark 14.17. Substitute A for $D' W$ to get Remark 14.21. □

Chapter 15

Time Series Analysis

Time series analysis is used for both predictions and theoretical models. This chapter will summarise some of the key concepts and tools. The perhaps most useful insight is that many models can be rewritten as a 1st-order vector autoregression, VAR(1), for which there are fairly simple expressions for forecasts and various diagnostic tools like impulse response functions and implied autocovariances.

15.1 Sample Autocorrelations

The starting point of time series modelling is often to describe the autocorrelation pattern of data. This section discusses autocorrelation, while a later section 15.5.4 will focus on partial autocorrelations.

The s th *autocovariance* of y_t is estimated as

$$\widehat{\text{Cov}}(y_t, y_{t-s}) = \sum_{t=1}^T (y_t - \bar{y})(y_{t-s} - \bar{y}) / T, \text{ where } \bar{y} = \sum_{t=1}^T y_t / T. \quad (15.1)$$

This chapter sometimes uses γ_s to denote an autocovariance (and $\hat{\gamma}_s$ its estimate). The convention in time series analysis is that we use the same sample average (\bar{y}) in both places and to divide by T . (In a reasonably long sample, this has small importance.)

The s th *autocorrelation* is estimated as

$$\hat{\rho}_s = \widehat{\text{Cov}}(y_t, y_{t-s}) / \widehat{\text{Std}}(y_t)^2 \quad (15.2)$$

Compared with a traditional estimate of a correlation we here impose that the standard deviation of y_t and y_{t-s} are the same (which typically does not make much of a difference).

The sampling properties of $\hat{\rho}_s$ are complicated, but there are several useful large sample results for Gaussian processes (these results often carry over to processes which

are similar to a Gaussian process—a homoskedastic process with finite 6th moment is typically enough, see Priestley (1981) 5.3 or Brockwell and Davis (1991) 7.2–7.3). When the true autocorrelations are all zero, then for any i and j different from zero

$$\sqrt{T} \begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \quad (15.3)$$

This result can be used to construct tests for both single autocorrelations (t-test or χ^2 test) and several autocorrelations at once (χ^2 test). In particular,

$$\sqrt{T} \hat{\rho}_s \xrightarrow{d} N(0, 1), \quad (15.4)$$

so $\sqrt{T} \hat{\rho}_s$ can be used as a t-stat. We can then define a 90% confidence band for $\hat{\rho}$ as $\pm 1.64/\sqrt{T}$ around the point estimate $\hat{\rho}$ or around the null hypothesis (0).

Empirical Example 15.1 (*Autocorrelations of returns and absolute values of returns, different lags*) See Figure 15.1 for results on S&P 500.

Example 15.2 (*t-test*) We want to test the hypothesis that $\rho_1 = 0$. Since the $N(0, 1)$ distribution has 5% of the probability mass below -1.64 and another 5% above 1.64, we can reject the null hypothesis at the 10% level if $\sqrt{T}|\hat{\rho}_1| > 1.64$. With $T = 100$, we therefore need $|\hat{\rho}_1| > 1.64/\sqrt{100} = 0.165$ for rejection, and with $T = 1000$ we need $|\hat{\rho}_1| > 1.64/\sqrt{1000} \approx 0.052$.

Equation (15.3) shows that $\sqrt{T} \hat{\rho}_i$ and $\sqrt{T} \hat{\rho}_j$ are independent $N(0, 1)$ variables, so the sum of the square of them is distributed as a χ^2_2 variable. More generally, the *Box-Pierce test* is that

$$Q_L = T \sum_{s=1}^L \hat{\rho}_s^2 \xrightarrow{d} \chi^2_L. \quad (15.5)$$

However, you could also test $T(\rho_1^2 + \rho_4^2)$, and it would have a χ^2_2 distribution.

Example 15.3 (*Box-Pierce*) Let $\hat{\rho}_1 = 0.165$, and $T = 100$, so $Q_1 = 100 \times 0.165^2 = 2.72$. The 10% critical value of the χ^2_1 distribution is 2.71, so the null hypothesis of no autocorrelation is rejected.

The choice of lag order in (15.5), L , should be guided by theoretical considerations, but it may also be wise to try different values. There is clearly a trade off: too few lags may miss a significant high-order autocorrelation, but too many lags can destroy the power of the test (as the test statistic is not affected much by increasing L , but the critical values increase).

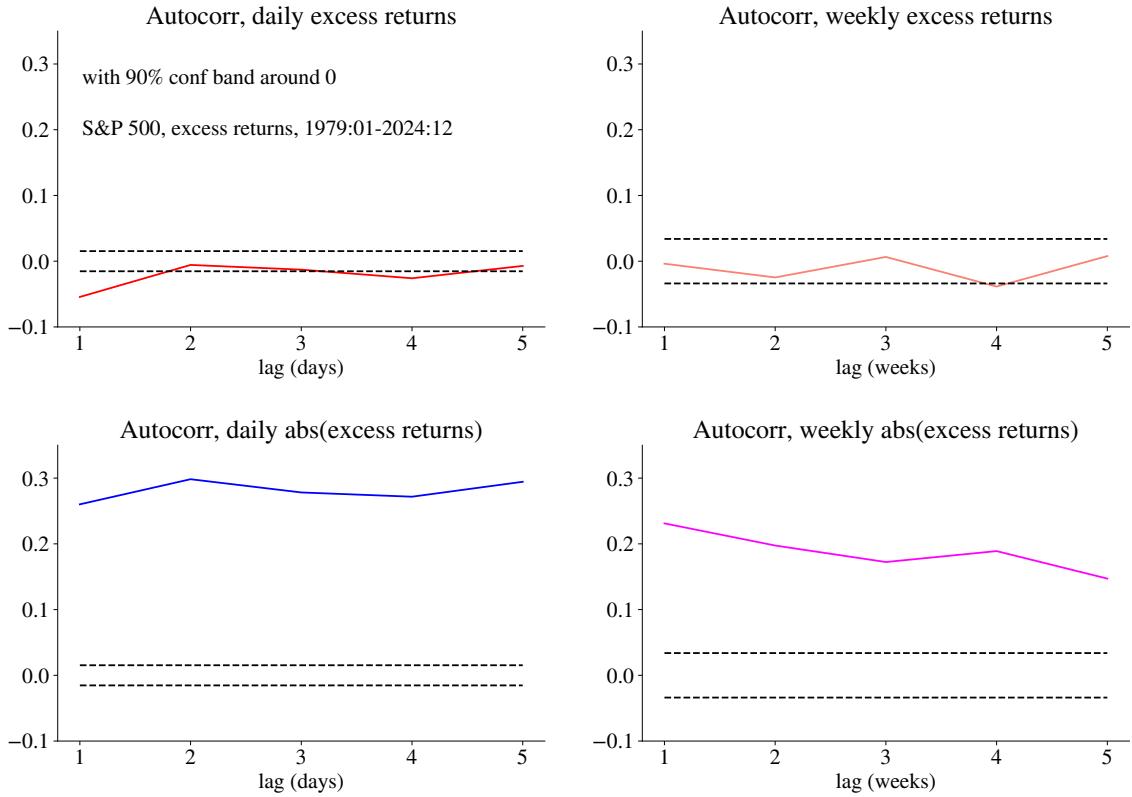


Figure 15.1: Predictability of US stock returns

15.2 Stationarity

The process y_t is (weakly) stationary if the mean, variance, and covariances are finite and constant across time. The *autocorrelation function* is just the autocorrelation coefficient ρ_s as a function of s . Notice that for any stationary series, ρ_s goes to zero at long lags. See Figure 15.2 for an example.

The autocorrelation function is strongly related to the *impulse response function* (IRF) which shows the dynamic response of y_{t+s} ($s = 0, 1, 2, \dots$) to a shock in t , that is $\partial y_{t+s} / \partial u_t$. For any stationary series, the IRF converges to zero as the horizon (s) is increased. See Figure 15.3 for an example. A later section will discuss a more formal way of assessing the stationarity.

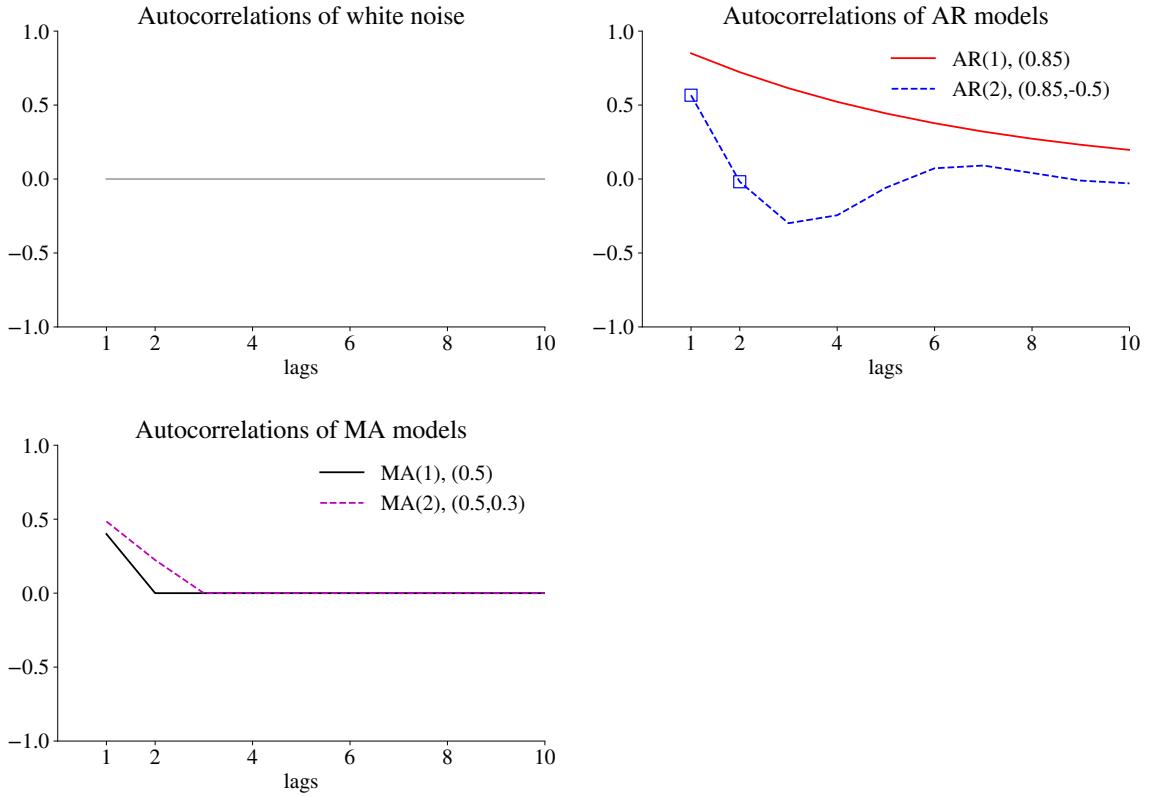


Figure 15.2: Example of autocorrelation functions. Some results discussed in the examples are especially marked.

15.3 White Noise

The *white noise* process is the basic building block used in most other time series models. It is characterized by a zero mean, a constant variance, and no autocorrelation

$$\begin{aligned} E \varepsilon_t &= 0 \\ \text{Var} (\varepsilon_t) &= \sigma^2, \text{ and} \\ \text{Cov} (\varepsilon_{t-s}, \varepsilon_t) &= 0 \text{ if } s \neq 0. \end{aligned} \tag{15.6}$$

This process can clearly not be forecasted. If, in addition, ε_t is normally distributed, then it is said to be Gaussian white noise.

To construct a white noise with a non-zero mean, we can add a constant μ ,

$$y_t = \mu + \varepsilon_t. \tag{15.7}$$

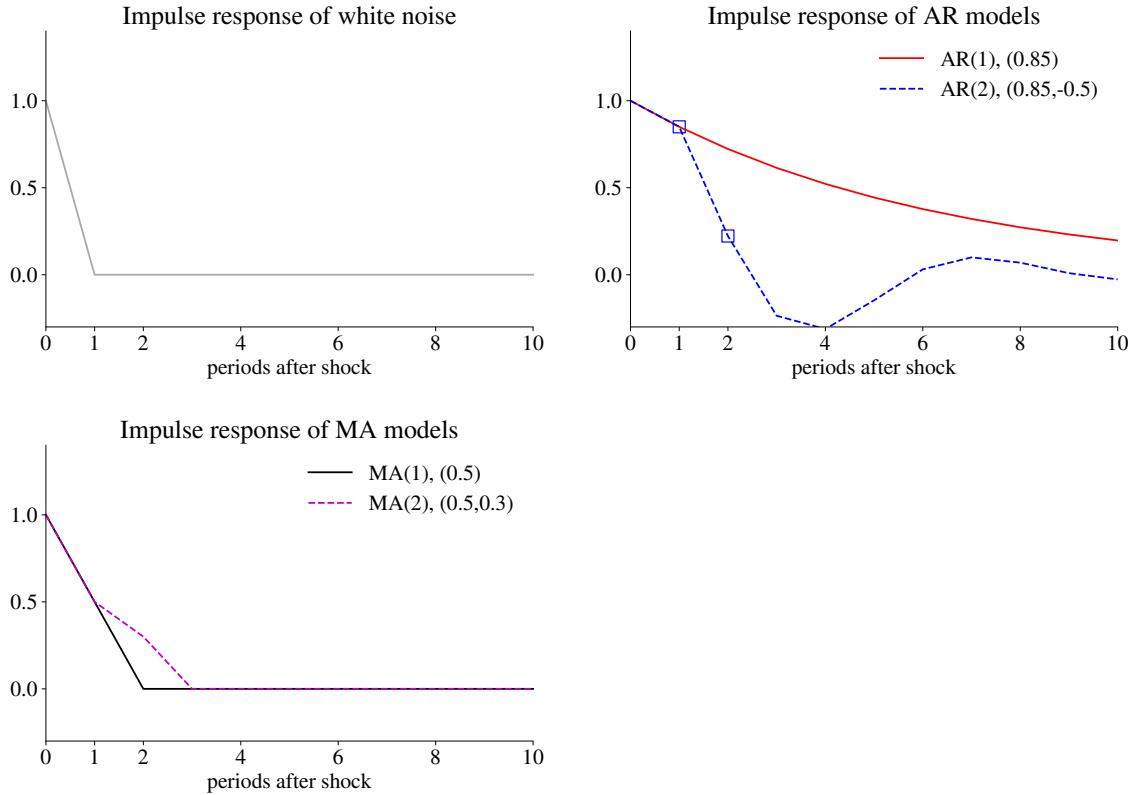


Figure 15.3: Impulse responses Some results discussed in the examples are especially marked.

The *impulse response function* of a white noise process is just a blip (no persistent effect), as illustrated in Figure 15.3.

The model parameters (μ, σ^2) are most easily estimated by the sample mean and variance, or by OLS with a constant term as the sole regressor.

15.4 Moving Average (MA)

A q^{th} -order moving average process $\text{MA}(q)$ is

$$y_t = \varepsilon_t + \sum_{s=1}^q \theta_s \varepsilon_{t-s}, \quad (15.8)$$

where the innovation ε_t is white noise (usually Gaussian). It is straightforward to add a constant to capture a non-zero mean. If the order of the MA is finite ($q < \infty$), then the MA model is stationary.

The *impulse response function* of an MA(q) is simply the MA coefficients $(1, \theta_1, \dots, \theta_q)$. This can feature many different patterns, but they are all 0 at horizons beyond q . See Figure 15.3 for an illustration.

Remark 15.4 (*Impulse response function**) For instance, write out (15.8) for y_t, y_{t+1}, \dots

$$\begin{aligned}y_t &= \underline{\varepsilon_t} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots \\y_{t+1} &= \varepsilon_{t+1} + \theta_1 \underline{\varepsilon_t} + \theta_2 \varepsilon_{t-1} + \dots \\y_{t+2} &= \varepsilon_{t+2} + \theta_1 \varepsilon_{t+1} + \theta_2 \underline{\varepsilon_t} + \dots,\end{aligned}$$

which shows that the dynamic effect of $\varepsilon_t = 1$ is $(1, \theta_1, \theta_2, \dots)$.

Estimation of an MA processes is typically done by *maximum likelihood* or GMM. LS does not work since the right hand side variables are unobservable. This is one reason why MA models play a limited role in applied work. Moreover, most MA models can be well approximated by an AR model of low order.

Remark 15.5 (*MLE of an MA(q)**) Assume that $(\varepsilon_{-q+1}, \dots, \varepsilon_0) = (0, \dots, 0)$. For $t \geq 1$, calculate $\varepsilon_t = y_t - \sum_{s=1}^q \theta_s \varepsilon_{t-s}$. If $\varepsilon_t \sim N(0, \sigma^2)$, then the log-likelihood contribution of observation t is $\ln L_t = \ln \phi(\varepsilon_t / \sigma_t) - \ln \sigma_t$, where $\phi()$ is the pdf of an $N(0, 1)$ variable. Maximise $\sum_{t=1}^T L_t$ with respect to $(\theta_1, \dots, \theta_q)$ and σ^2 (or σ). Clearly, the ε_t series must be recalculated in each iteration.

Remark 15.6 (*GMM of an MA(1)**) To estimate θ_1 and σ^2 for a zero mean variable y_t , use the two moment conditions

$$\begin{bmatrix} y_t^2 - \sigma^2(1 + \theta_1^2) \\ y_t y_{t-1} - \sigma^2 \theta_1 \end{bmatrix}$$

where the second numbers are the autocovariances implied by an MA(1), see Example 15.7.

15.4.1 MA(q): Implied Autocovariances*

The autocovariances implied by an MA(q) model are

$$\text{Cov}(y_t, y_{t-s}) = \sigma^2 \sum_{i=0}^{q-s} \theta_{i+s} \theta_i \text{ for } 0 \leq s \leq q, \quad (15.9)$$

with the convention that $\theta_0 = 1$. For $s > q$, the covariance is zero.

Example 15.7 (*Autocovariances of MA(1) and MA(2)*) For an MA(1), we get $\text{Cov}(y_t, y_{t-1}) = \sigma^2\theta_1$ and zero for further lags. See Figure 15.2 for an example. Instead, for an MA(2), $\text{Cov}(y_t, y_{t-1}) = \sigma^2(\theta_1 + \theta_2\theta_1)$, $\text{Cov}(y_t, y_{t-2}) = \sigma^2\theta_2$, and zero for further lags.

Proof (of (15.9)*) Write (15.8) as $y_t = \sum_{i=0}^{s-1} \theta_i \varepsilon_{t-i} + \sum_{i=s}^q \theta_i \varepsilon_{t-i}$. For instance, with $q = 3$ and $s = 2$, the two terms are $(\theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1}) + (\theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3})$. Similarly, $y_{t-s} = \sum_{i=0}^{q-s} \theta_i \varepsilon_{t-s-i} + \sum_{i=q-s+1}^q \theta_i \varepsilon_{t-s-i}$. For instance, $(\theta_0 \varepsilon_{t-2} + \theta_1 \varepsilon_{t-3}) + (\theta_2 \varepsilon_{t-4} + \theta_3 \varepsilon_{t-5})$. The only terms that correlate are the 2nd term in y_t and the 1st in y_{t-2} . Since ε_t are iid, the covariance is as in (15.9), which in our example is $\text{Cov}(\theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3}, \theta_0 \varepsilon_{t-2} + \theta_1 \varepsilon_{t-3}) = \sigma^2(\theta_2\theta_0 + \theta_3\theta_1)$. \square

Remark 15.8 (*VMA**) In case of a Vector Moving Average (VMA), where y_t and ε_{t-s} in (15.8) are $n \times 1$ vectors and θ_s are $n \times n$ matrices, the autocovariance-covariance matrix in (15.9) becomes $\text{Cov}(y_t, y_{t-s}) = \sum_{i=0}^{q-s} \theta_{i+s} \Omega \theta_i'$, where Ω is the variance-covariance matrix of ε_t .

15.5 Autoregressions

15.5.1 AR(1)

In this section we study the *first-order autoregressive* process, AR(1), in some detail in order to understand the basic concepts of AR processes. Later sections will extend to AR(p) models.

An AR(1) is

$$y_t = ay_{t-1} + \varepsilon_t, \text{ with } \text{Var}(\varepsilon_t) = \sigma^2, \quad (15.10)$$

where ε_t is the white noise process in (15.6), assumed to be uncorrelated with y_{t-s} for $s \geq 1$. To include a mean, μ , substitute $x_t - \mu$ for y_t and rearrange to get

$$x_t = (1 - a)\mu + ax_{t-1} + \varepsilon_t. \quad (15.11)$$

The AR(1) model parameters (a, μ, σ^2) can be *estimated with OLS* (since ε_t and y_{t-1} are uncorrelated).

The basic properties of an AR(1) process are (provided $|a| < 1$)

$$\text{Var}(y_t) = \sigma^2 / (1 - a^2) \quad (15.12)$$

$$\text{Corr}(y_t, y_{t-s}) = a^s. \quad (15.13)$$

The impulse response function equals the autocorrelation function for an AR(1). With $0 < a < 1$, they show an exponentially decaying pattern, and with $-1 < a < 0$ a zigzag pattern that decreases in amplitude. See Figures 15.2–15.3 for illustrations. The case of $|a| = 1$ creates a random walk process, which is non-stationary. It will be discussed in section 15.10.

Example 15.9 With $a = 0.85$ and $\sigma^2 = 0.5^2$, we have $\text{Var}(y_t) = 0.25/(1 - 0.85^2) \approx 0.9$, which is much larger than the variance of the residual since the uncertainty is accumulating when $a > 0$. See Figure 15.2 for the autocorrelations.

Empirical Example 15.10 (AR(1) for different investment horizons) See Figure 15.4 for AR(1) results for different investment horizons.

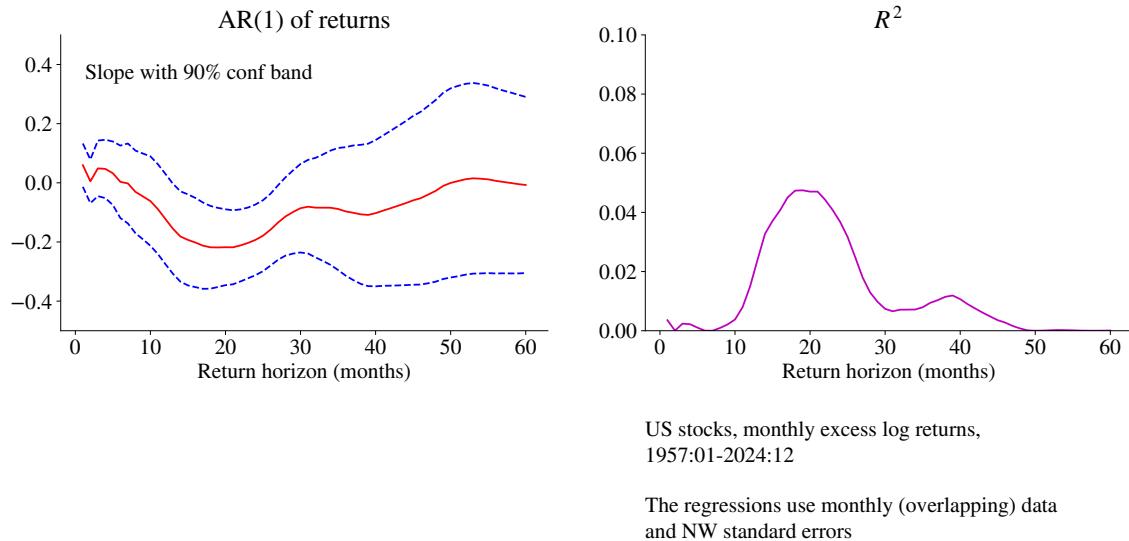


Figure 15.4: Predicting long run US stock returns

Proof (of (15.12)–(15.13)) Substitute for y_{t-1} in (15.10) using the same equation, but lagged once. Keep substituting (for y_{t-2}, \dots) to get $y_t = a^s y_{t-s} + \sum_{i=0}^{s-1} a^i \varepsilon_{t-i}$. (The last term is $\varepsilon_t + a\varepsilon_{t-1} + \dots + a^{s-1}\varepsilon_{t-s+1}$.) This shows directly that $\text{Cov}(y_t, y_{t-s}) = a^s \text{Var}(y_t)$, since y_{t-s} is uncorrelated with $(\varepsilon_t, \dots, \varepsilon_{t-s+1})$. This gives (15.13). In addition, assuming $|a| < 1$ and taking the limit as $s \rightarrow \infty$, gives $y_t = \sum_{i=0}^{\infty} a^i \varepsilon_{t-i}$. Since ε_t is iid, the variance is $\sigma^2 \sum_{i=0}^{\infty} a^{2i}$, which equals (15.12). \square

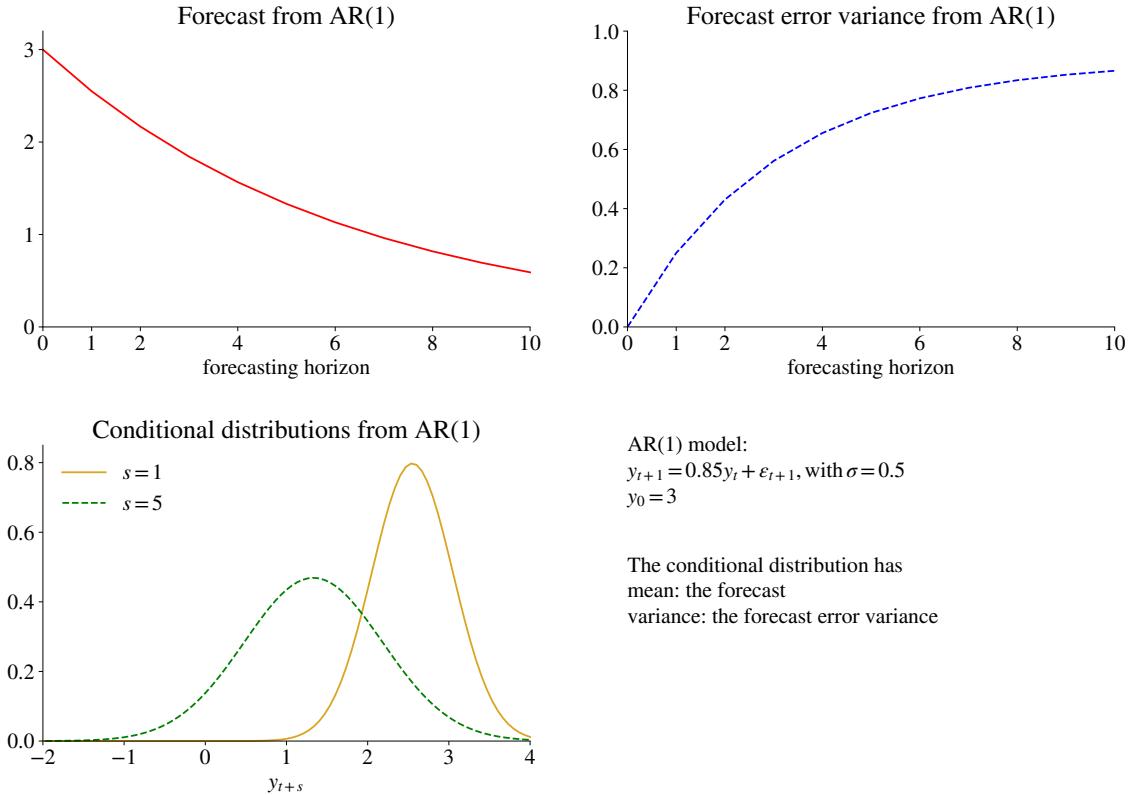


Figure 15.5: Properties of forecasts from AR(1) process

15.5.2 Forecasting with an AR(1)

Suppose we have estimated an AR(1). To simplify the exposition, we assume that we actually know a and $\text{Var}(\varepsilon_t)$, which might be a reasonable approximation if they were estimated on a long sample.

We want to *forecast* y_{t+1} using information available in t . From (15.10) we get

$$y_{t+1} = ay_t + \varepsilon_{t+1}. \quad (15.14)$$

Since the best guess of ε_{t+1} is that it is zero, the best forecast and the associated forecast error are

$$\mathbb{E}_t y_{t+1} = ay_t, \text{ and} \quad (15.15)$$

$$y_{t+1} - \mathbb{E}_t y_{t+1} = \varepsilon_{t+1} \text{ with variance } \sigma^2. \quad (15.16)$$

More generally, we have

$$\mathbb{E}_t y_{t+s} = a^s y_t, \quad (15.17)$$

$$\text{Var}(y_{t+s} - \mathbb{E}_t y_{t+s}) = \sigma^2(1 - a^{2s})/(1 - a^2). \quad (15.18)$$

Notice that the point forecast converges towards zero and that the variance of the forecast error converges to the unconditional variance in (15.12). Figure 15.5 illustrates this property.

Example 15.11 If $y_t = 3$, $a = 0.85$ and $\sigma = 0.5$, then the forecasts and the forecast error variances become

Horizon s	$\mathbb{E}_t y_{t+s}$	$\text{Var}(y_{t+s} - \mathbb{E}_t y_{t+s})$
1	$0.85 \times 3 = 2.55$	0.25
2	$0.85^2 \times 3 = 2.17$	$(0.85^2 + 1) \times 0.5^2 = 0.43$
25	$0.85^{25} \times 3 = 0.05$	$\frac{0.85^{50}-1}{0.85^2-1} \times 0.5^2 = 0.90$

See Figure 15.5.

Proof (of (15.17)–(15.18)) From the proof of (15.12)–(15.13), recall that $y_t = a^s y_{t-s} + \sum_{i=0}^{s-1} a^i \varepsilon_{t-i}$. Since $\mathbb{E}_{t-s} \varepsilon_{t-s+i} = 0$ for $i \geq 1$, we have $\mathbb{E}_{t-s} y_t = a^s y_{t-s}$. Shift time subscripts to get (15.17). Also, the variance of the second term in the expression of y_t is $\sigma^2 \sum_{i=0}^{s-1} a^{2i}$, which can be simplified as (15.18). \square

15.5.3 AR(p)

The p th-order autoregressive process, AR(p), is a straightforward extension of the AR(1)

$$y_t = \sum_{s=1}^p a_s y_{t-s} + \varepsilon_t. \quad (15.19)$$

Models with $p \geq 2$ can capture rich dynamics: see Figure 15.3 for a comparison of impulse response functions of an AR(1) and an AR(2). Adding a constant to the theoretical expressions is straightforward: substitute $x_t - \mu$ for y_t everywhere. In some cases, it may make sense to skip some lags to get, for instance, $y_t = a_1 y_{t-1} + a_4 y_{t-4} + \varepsilon_t$. The AR(1) model parameters $(a_1, \dots, a_p, \mu, \sigma^2)$ can be estimated with OLS (since ε_t uncorrelated with y_{t-s}).

Section 15.9 shows how an AR(p) can be rewritten as a 1st order Vector Autoregression, VAR(1), which facilitates computations (of forecasts, impulse response functions, etc).

15.5.4 Descriptive Statistics: Sample Partial Autocorrelations

The p th partial autocorrelation measures the direct relation between y_t and y_{t-p} , where the indirect effects of $y_{t-1}, \dots, y_{t-p+1}$ are eliminated. That is,

$$\text{partial autocorrelation}(s) = \text{Corr}(y_t - \hat{y}_t, y_{t-s} - \hat{y}_{t-s}), \quad (15.20)$$

where \hat{y}_t and \hat{y}_{t-s} are the best linear estimates of y_t and y_{t-s} using $(y_{t-1}, \dots, y_{t-s+1})$ as regressors (same in both regressions).

Example 15.12 (*The 3rd partial autocorrelation coefficient*) Regress

$$\begin{aligned} y_t &= a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t \\ y_{t-3} &= b_0 + b_1 y_{t-1} + b_2 y_{t-2} + u_{t-3}. \end{aligned}$$

The third partial autocorrelation coefficient is then $\text{Corr}(\hat{\varepsilon}_t, \hat{u}_{t-3})$.

To see the distinction between autocorrelations and partial autocorrelations, consider an AR(1) model, $y_t = a y_{t-1} + \varepsilon_t$: the 2nd autocorrelation is a^2 , whereas the 2nd partial autocorrelation is zero (since y_{t-2} does not have any direct effect on y_t once you have controlled for y_{t-1}). The partial autocorrelation is therefore a way to gauge how many lags that are needed in an AR(p) model.

In practice, the first partial autocorrelation is often *estimated* by a in an AR(1)

$$y_t = \omega + \underline{a} y_{t-1} + \varepsilon_t, \quad (15.21)$$

while the second partial autocorrelation is estimated by the second slope coefficient (a_2) in an AR(2)

$$y_t = \omega + a_1 y_{t-1} + \underline{a_2} y_{t-2} + \varepsilon_t, \quad (15.22)$$

and so forth. The general pattern is that the p th partial autocorrelation is estimated by the slope coefficient of the p th lag in an AR(p), where we let p first be 1, then 2, and then 3, and so on. See Figure 15.6 for an illustration.

15.5.5 Choice of Lag Order (p)

To choose the model, (1) study the autocorrelations and partial autocorrelations in data; and (2) and check that residual are close to white noise. To avoid overfitting, punish models with many parameters by applying Akaike's Information Criterion (AIC) or the

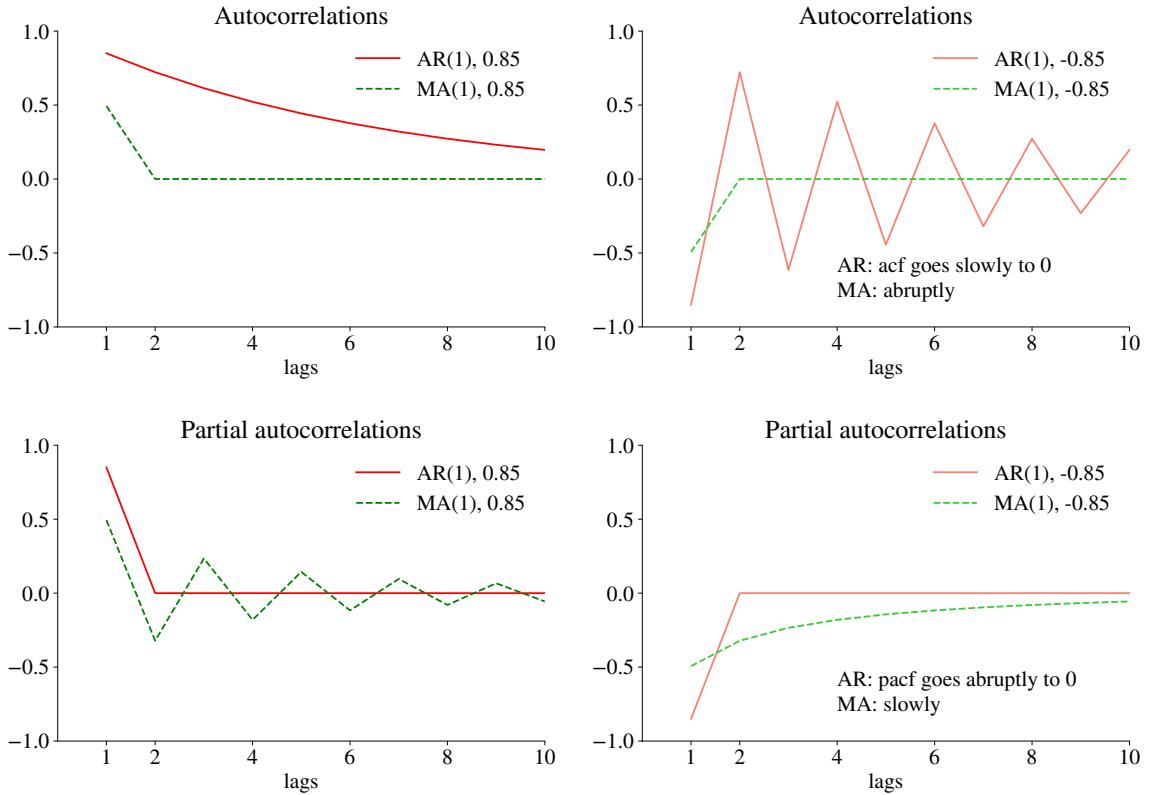


Figure 15.6: Autocorrelations and partial autocorrelations

Bayesian information criterion (BIC)

$$AIC = \ln \hat{\sigma}^2 + 2(p + 1)/T \quad (15.23)$$

$$BIC = \ln \hat{\sigma}^2 + (p + 1)/T \times \ln T, \quad (15.24)$$

where $\hat{\sigma}^2$ is the variance of the fitted residuals. Choose the model with the lowest AIC or BIC. (Note, however, that AIC often exaggerates the lag length.) This provides a trade-off between fit (low $\hat{\sigma}^2$) and number of parameters ($p + 1$).

Empirical Example 15.13 Figure 15.7 illustrates the lag order choice for modelling monthly S&P 500 volatility. The results indicate that an AR(2) might be a reasonable choice (or possibly an AR(1)).

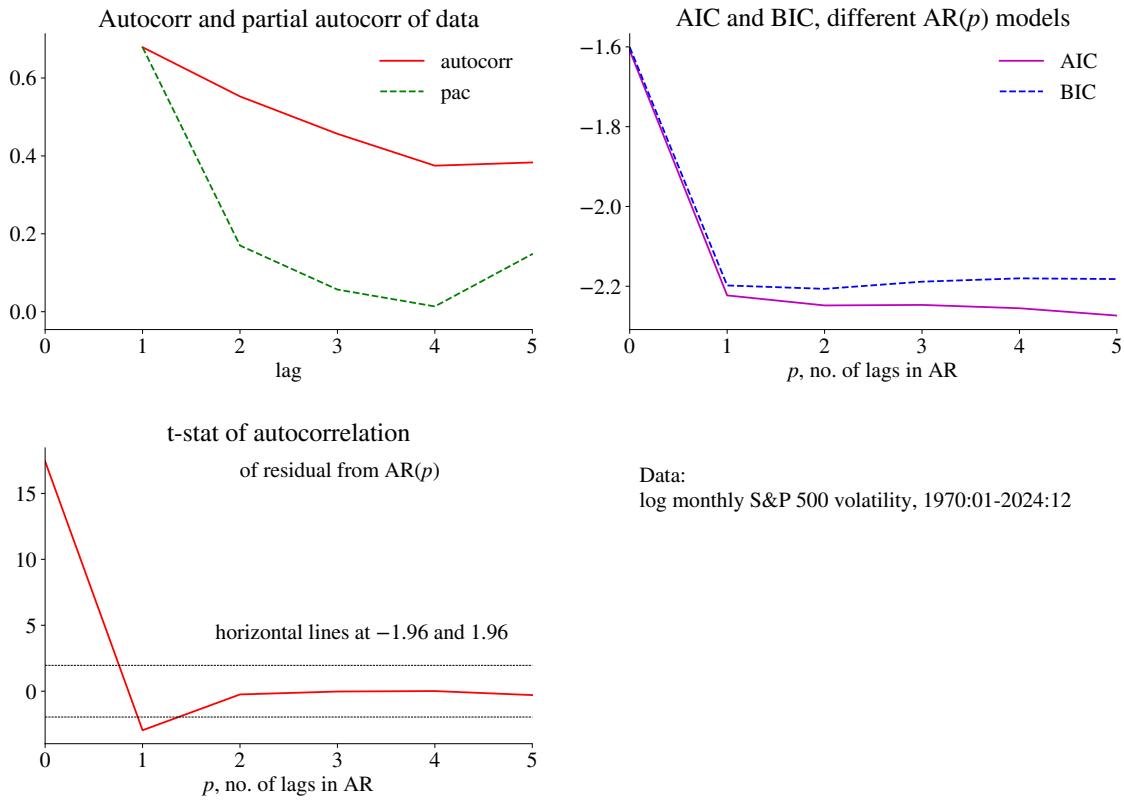


Figure 15.7: Example of choosing lag length in an AR model

15.6 ARMA(p,q)

When both the autocorrelations and partial autocorrelations show mixed patterns, then a combination of AR and MA models might be appropriate. Alternatively, a higher-order AR model might work as an approximation.

Autoregressive moving average (ARMA) models add a moving average structure to an AR model. For instance, an ARMA(2,1) is

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t + \theta \varepsilon_{t-1}, \quad (15.25)$$

where ε_t is white noise. (It is straightforward to add a constant to capture a non-zero mean.) If the AR part of the ARMA is stationary, then the whole ARMA model is (since MA models are stationary). In an ARMA model, both the autocorrelations and partial autocorrelations decay to slowly zero. Even low-order ARMA models can generate complicated dynamics. ARMA models are harder to estimate than an autoregressive

model, and we typically use MLE or GMM.

Remark 15.14 (MLE of an ARMA(2,1)^{*}) Assume $\varepsilon_0 = 0$ and calculate $\varepsilon_1 = y_1 - a_1 y_0 - a_2 y_{-1}$ and $\varepsilon_t = y_t - a_1 y_{t-1} - a_2 y_{t-2} - \theta \varepsilon_{t-1}$ for $t \geq 2$. Use in the log likelihood function, $\ln L_t = \ln \phi(\varepsilon_t/\sigma) - \ln \sigma$ and maximize with respect to $(a_1, a_2, \theta, \sigma^2)$ or σ . As usual, the ε_t series must be recalculated for every new guess of the parameter vector.

15.7 Approximating MA and ARMA Models with an AR Model

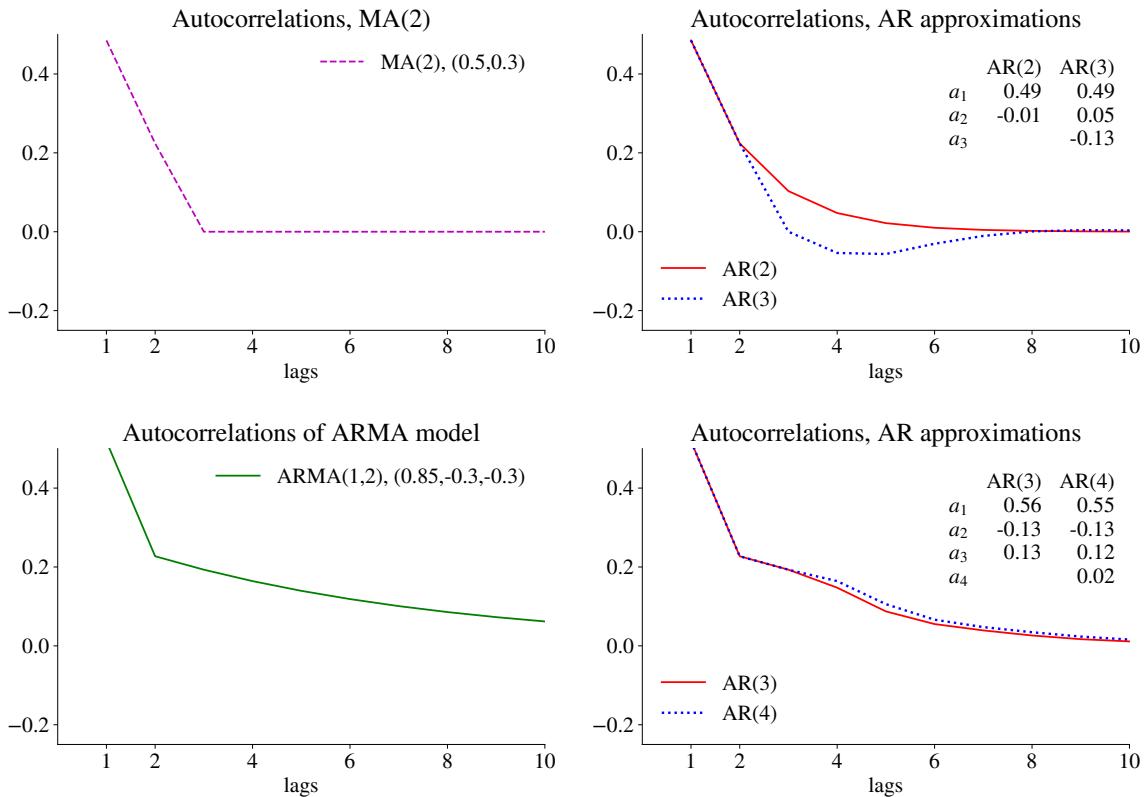


Figure 15.8: Examples of autocorrelation functions for some MA and ARMA models, and AR approximations

15.7.1 Autocovariances versus Autoregression Coefficients

There is a direct relation between autocovariances and the autoregression (AR) coefficients, the *Yule-Walker equations*. This allows us to supply p autocovariances or autocorrelations

from data or another time series model, and calculating the coefficients of an AR(p) needed to replicate those results. This is thus useful for both estimation and approximation.

With the variance γ_0 and p autocovariances (denoted $\gamma_1, \dots, \gamma_p$) we have the following relation to the autoregression coefficients (denoted a_1, \dots, a_p)

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}. \quad (15.26)$$

This can be solved for the vector a_1, \dots, a_p . For instance, with $p = 1$ we have $\gamma_1 = \gamma_0 a_1$, so $a_1 = \gamma_1 / \gamma_0$, which is the first autocorrelation. Notice that dividing all γ_s elements by γ_0 would not change the solution of (a_1, \dots, a_p) , which means that γ_s could effectively be either autocovariances or autocorrelations (in the latter case, $\gamma_0 = 1$). The matrix in (15.26) is a symmetric Toeplitz matrix, which modern software easily constructs from the vector $\gamma_1, \dots, \gamma_p$.

This also provides a mechanism for calculating the partial autocorrelations (15.21)–(15.22) from the autocovariances. First, set $p = 1$ in (15.26) to find a_1 . Second, set $p = 2$ and solve to find a_1, a_2 , where we save the last coefficient, and so forth.

A straightforward way to do the reverse calculations (calculate the autocovariances γ_s from the autoregression coefficients a_s) is described in section 15.9. (It involves rewriting the AR(p) on VAR(1) form and the applying straightforward formulas.)

Example 15.15 With $(\gamma_0, \gamma_1, \gamma_2) = (1, 0.567, -0.018)$ as in Figure 15.2 (notice the markers), the Yule-Walker equations (15.26) show that the AR(2) coefficients must be $(a_1, a_2) \approx (0.85, -0.5)$.

Proof (of (15.26) for $p = 2$) Consider an AR(2) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$. Clearly, the covariance of y_t and y_{t-s} must be the same as the covariance between the right hand side of and y_{t-s} , $\text{Cov}(y_t, y_{t-s}) = \text{Cov}(a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t, y_{t-s})$. Apply this for $s = 1, 2$ and use the short hand notation γ_s to get $\gamma_1 = \gamma_0 a_1 + \gamma_1 a_2$ and $\gamma_2 = \gamma_1 a_1 + \gamma_0 a_2$. Rewrite on matrix form as

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

□

15.7.2 Relation of AR(p) and MA(q) Models

For low-order MA and AR models, there are distinct differences, so the autocorrelations and partial autocorrelations (for different lags) can help us gauge if the time series *looks more like an AR or an MA*. In particular, in a low-order AR(p) model, the autocorrelations decay to zero for long lags, while the $p + 1$ partial autocorrelation (and beyond) goes abruptly to zero. The reverse is true for an MA model. See Figure 15.6 for an illustration. This might help making the choice between a MA and AR model. However, the situation is more blurred when we allow for high-order models.

In fact, it is always possible to rewrite any stationary time series model (also AR(p) ones) on MA(∞) form, as formulated in Wold's decomposition (see Hamilton (1994) 4.8). Also, many MA(q) models can be inverted as an AR(p). In case they cannot be inverted (there are conditions for invertibility, see Hamilton (1994) 3.7), then there is an equivalent MA model with the same autocovariance function that *can* be inverted.

For practical purposes, it might be more interesting to approximate an MA or ARMA model by a low-order AR since they are often easier to work with. The work flow is then (1) calculate the impulse response function of the exact/correct model; (2) find the autocovariances by using (15.9); (3) apply the Yule-Walker equations (15.26) for different choices of p to get the AR(p) coefficients. For instance, Figure 15.8 shows how MA and ARMA models can be approximated. (To show the autocorrelations implied by those AR(p) models, we also use the methods discussed in 15.9.) The figures show that an AR(p) can match the first p autocovariances exactly, but that autocorrelations beyond have errors. With a sufficiently high p , this approach may create a useful approximation.

15.8 VAR(p)

Let y_t be an $n \times 1$ vector of variables. The VAR(p) is

$$y_t = A_1 y_{t-1} + \dots + A_p y_{t-p} + \varepsilon_t, \quad (15.27)$$

where A_1, \dots, A_p are $n \times n$ matrices and ε_t is an $n \times 1$ vector of iid shocks, although possibly correlated in the cross-section. The vector autoregression is a multivariate version of an AR process. As usual, it is straightforward to add a constant. To gauge the dynamics we can calculate the *impulse response function* of (all the n elements of) y_t to a shock to the j th element of the $n \times 1$ vector ε_0 . See Figure 15.9 for an illustration. It suggests that a low-order VAR model (here a VAR(1)) can create more complicated dynamics than a

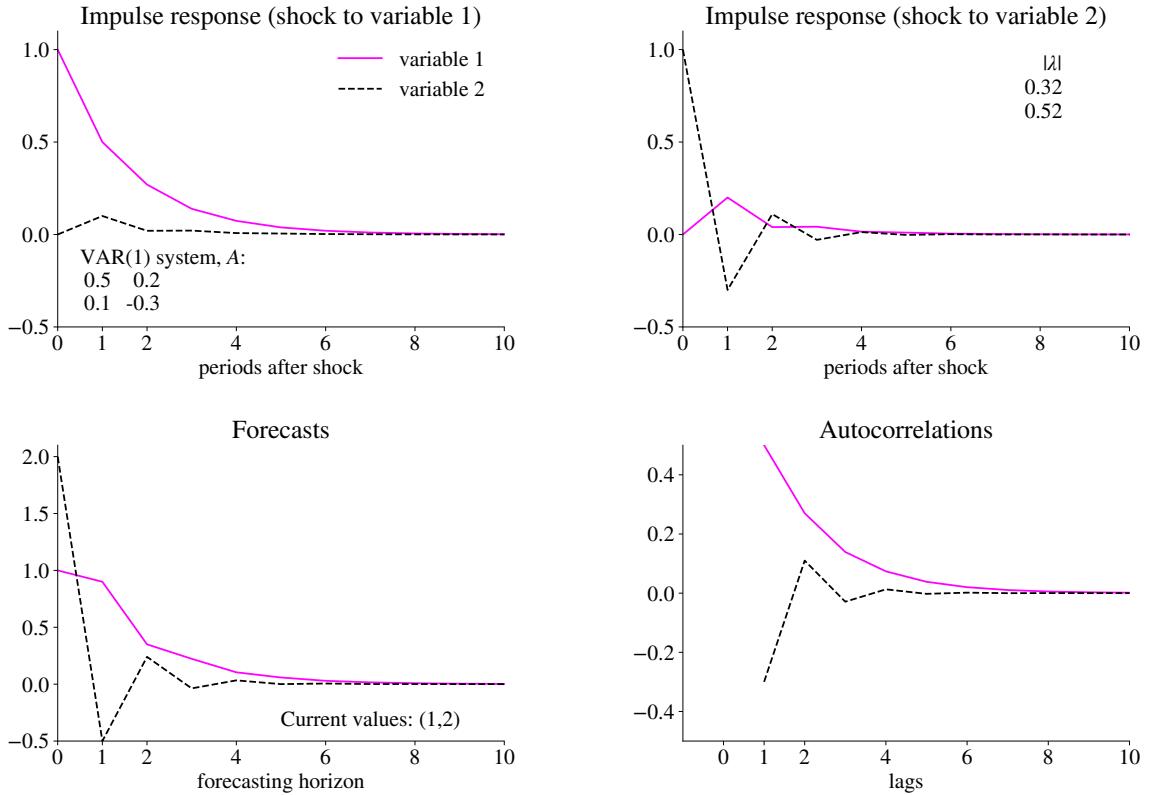


Figure 15.9: Properties of a VAR(1) model

low-order AR model. (A later section discusses how rewriting the model on an extended VAR(1) form can help with these calculations.)

For instance, a VAR(2) with two variables is

$$\begin{bmatrix} w_t \\ z_t \end{bmatrix} = \begin{bmatrix} A_{1,11} & A_{1,12} \\ A_{1,21} & A_{1,22} \end{bmatrix} \begin{bmatrix} w_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} A_{2,11} & A_{2,12} \\ A_{2,21} & A_{2,22} \end{bmatrix} \begin{bmatrix} w_{t-2} \\ z_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{wt} \\ \varepsilon_{zt} \end{bmatrix}. \quad (15.28)$$

Each line defines a regression equation, which can be *estimated with OLS* (since ε_{wt} and ε_{zt} are uncorrelated with lags of w and z). Also, the covariance matrix of the residuals can be estimated from the fitted residuals.

If z_t can help predict future w , over and above what lags of w itself can, then z is said to *Granger Cause* w . This is a statistical notion of causality, and may not necessarily have much to do with true causality. In (15.28) z does Granger cause w if $A_{1,12} \neq 0$ and/or $A_{2,12} \neq 0$, which is easily tested.

Subsequent sections rewrite the VAR(p) on an extended VAR(1) form, which facilitates

many computations.

15.9 VAR(1)

Both an AR(p) and a VAR(p) can be rewritten as larger VAR(1) systems (also called *companion form*)

$$x_t = Ax_{t-1} + u_t, \text{ with } \Omega = \text{Cov}(u_t), \quad (15.29)$$

and where u_t is vector of iid shocks, although possibly correlated in the cross-section. The details of this transformation is discussed in the subsequent section. Some of the elements in u_t might have zero variances, especially if this represents the companion form of a higher-order system.

The point of rewriting on companion form is that VAR(1) systems have simple formulas for assessing stationarity, making forecasting and calculating impulse response function as well as implied autocovariances.

15.9.1 VAR(1): Companion Form

Both an AR(p) and a VAR(p) can be rewritten as larger VAR(1) systems. This is helpful since calculations of forecasts and impulse response functions are straightforward for a VAR(1). As an example, an AR(3) can be rewritten on VAR(1) (or companion) form as

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix}, \quad (15.30)$$

where the 3-element vector (y_t, y_{t-1}, y_{t-2}) corresponds to x_t in (15.29) and the 3×3 matrix on the right hand side to A . This system incorporates the dynamics (and the shock) in the first row, while the remaining rows are identities (for instance $y_{t-1} = y_{t-1}$). Calculations of forecasts and impulse response functions will use the full system, but we are eventually only interested in the results for the first variable (x_t). The extension to higher order models is straightforward. Notice that this model has a singular variance-covariance matrix of the residuals, $\Omega = \text{Var}(u_t)$,

$$\Omega = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (15.31)$$

Similarly, a VAR(3) for the m -element y_t vector can be rewritten in a similar way (on companion form), except that we need matrices and vectors

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & A_3 \\ I_m & \mathbf{0}_{m \times m} & \mathbf{0}_{m \times m} \\ \mathbf{0}_{m \times m} & I_m & \mathbf{0}_{m \times m} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \mathbf{0}_{m \times 1} \\ \mathbf{0}_{m \times 1} \end{bmatrix}. \quad (15.32)$$

In this case, (y_t, y_{t-1}, y_{t-2}) is a $3m$ -element vector (there are m elements in y_t) which corresponds to x_t in (15.29) and the $3m \times 3m$ matrix on the right hand side to A . As above, calculations of forecasts and impulse response functions will use the full system, but we are eventually only interested in the results for the first m variable (those in the m -vector y_t). Also, the extension to higher order models is straightforward. The methods for assessing stationarity, making forecasts and calculating impulse response functions of a VAR(1) can thus be used for any AR(p) and a VAR(p) model.

Example 15.16 An AR(2) with $(a_1, a_2) = (0.85, -0.5)$ has the companion form

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 0.85 & -0.5 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}, \text{ with } \Omega = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

15.9.2 VAR(1): Stationarity

To assess the *stationarity* of a VAR(1) system, calculate the eigenvalues (λ_i) of the matrix A in (15.29). If $|\lambda_i| < 1$ for all i , then the model is stationary. This implies that the effect of a shock eventually dissipates, as will be seen by plotting the impulse response function.

This is helpful since forecasts and impulse response functions are straightforward to calculate for a VAR(1).

Example 15.17 With the VAR(1) matrix in Example 15.16, then we get $(|\lambda_1|, |\lambda_2|) \approx (0.707, 0.707)$. See Figure 15.3 for an illustration of the impulse response functions.

Example 15.18 See Figure 15.9 for an illustration of the impulse response functions for a VAR(1) of two variables. The system appears to be stationary, which is confirmed by the eigenvalues.

15.9.3 VAR(1): Forecasts

With the information available in t , that is, information about x_t , (15.29) can be used to forecast one- and two-step ahead as

$$\mathbb{E}_t x_{t+1} = Ax_t \text{ and } \mathbb{E}_t x_{t+2} = A\mathbb{E}_t x_{t+1} = A^2x_t, \quad (15.33)$$

where $A^2 = AA$ is the matrix product of A and A . The two-period forecast has the same form as the one-period forecast, but with other coefficients. More generally, the forecast and forecast error variance for s period ahead are

$$\mathbb{E}_t x_{t+s} = A^s x_t \quad (15.34)$$

$$\text{Var}(x_{t+s} - \mathbb{E}_t x_{t+s}) = \Sigma_{i=0}^{s-1} A^i \Omega A^{i'}, \quad (15.35)$$

where $\Omega = \text{Var}(u_t)$. (It is sometimes useful to notice that $A^{i'} = A^i$.)

Remark 15.19 (*Calculating $\text{Var}(x_{t+s} - \mathbb{E}_t x_{t+s})$) The variance-covariance matrix $\text{Var}(x_{t+s} - \mathbb{E}_t x_{t+s})$ Γ can be calculated by iterating over $\Gamma_i = A\Gamma_{i-1}A' + \Omega$ for $i = 1$ to s , starting at $\Gamma_0 = \mathbf{0}_{n \times n}$.

Example 15.20 (Forecasts from a VAR(1)) With the 2-variable VAR(1) in Example 15.18 and the initial values $x_t = (1, 2)$ we get

$$\begin{aligned} \mathbb{E}_t x_{t+1} &= \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.9 \\ -0.5 \end{bmatrix} \text{ and} \\ \mathbb{E}_t x_{t+2} &= \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix} \begin{bmatrix} 0.9 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0.35 \\ 0.24 \end{bmatrix}. \end{aligned}$$

See Figure 15.9 for an illustration.

Proof (of (15.34)–(15.35)) From the proof of (15.37)–(15.38), recall that $x_t = A^s x_{t-s} + \Sigma_{i=0}^{s-1} A^i u_{t-i}$, where the last term is $u_t + Au_{t-1} + \dots + A^{s-1}u_{t-s+1}$. Since $\mathbb{E}_{t-s} u_{t-s+i} = 0$ for $i \geq 1$, $\mathbb{E}_{t-s} x_t = A^s x_{t-s}$. Shift time subscripts to get (15.34). Also, the variance of the second term is $\Sigma_{i=0}^{s-1} A^i \Omega A^{i'}$. \square

Empirical Example 15.21 (AR(2) results for realized volatility of S&P 500 returns) Figure 15.10 shows results for monthly volatility of the S&P 500.

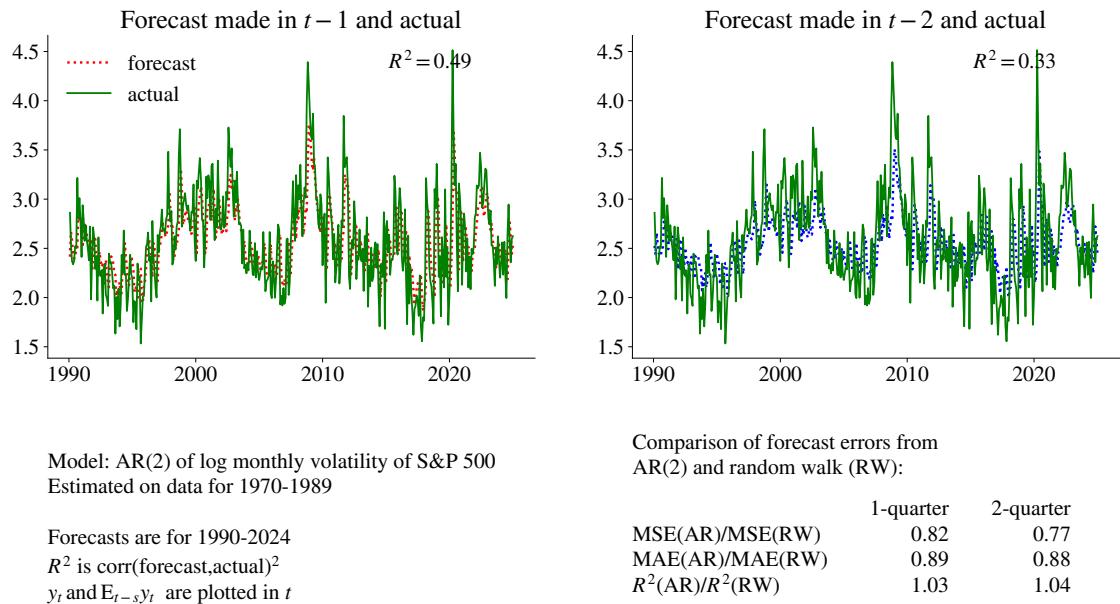


Figure 15.10: Forecasting with an AR(2)

Remark 15.22 (*Simplified forecast equations**) We could alternatively estimate a separate model for each forecasting horizon. In a large sample, this would give similar (but not the same) results as first estimate the VAR and then apply (15.34). For instance, suppose the correct model is the VAR(1) in (15.29) and that we want to forecast x one and two periods ahead. From (15.34) we see that the system of regression equations should be of the form

$$x_{t+1} = \delta x_t + u_{t+1}, \text{ and}$$

$$x_{t+2} = \gamma x_t + v_{t+2}.$$

Each equation is a system of n regressions: one for each element of the vector x_{t+1} , so δ is an $n \times n$ matrix of coefficients: one row per element in x_{t+1} . With estimated coefficients (OLS can be used), it is straightforward to calculate forecasts and forecast error variances. (When x_t is from a companion form as in (15.30) or (15.32), then some of these regressions are trivial, since the dependent variable is among the regressors.)

15.9.4 VAR(1): Impulse Response Function

The impulse response of the VAR(1) in (15.29) with respect to a shock in element j of u_t (u_{jt}) is

$$\frac{\partial x_{t+s}}{\partial u_{j,t}} = A^s e_j, \quad (15.36)$$

where e_j is a vector of zeros, except that element j equals one ($A^s e_j$ is thus the same as column j of A^s). That is, element (i, j) in A^s is the effect of shock $u_{j,t}$ on $x_{i,t+s}$. In case (15.29) represents a system written on companion form, then we are clearly only interested in the shocks to the first block of variables. Recall that the impulse response function is the same as the MA(∞) form of the model.

Example 15.23 Figure 15.3 shows (notice the markers) the impulse response function for the AR(2) in Example 15.16. Similarly, Figure 15.9 shows the result for the two-variable VAR(1) in Example 15.18.

Proof (of (15.36)) From the proof of (15.37)–(15.38), $x_t = \sum_{s=0}^{\infty} A^s u_{t-s}$. The response of y_t to $u_{t-s} = e_j$ (a vector of zeros, except that element j equals one) is thus $A^s e_j$, which is the same as column j of A^s . Shift subscripts to get (15.36). \square

Remark 15.24 (*Local projections**) Similar to the logic of different regressions for different forecasting horizons, we could also estimate the impulse response function by “local projections.” In practice, it means estimating

$$x_{t+s} = a + B_{1,s} x_t + \dots + B_{p,s} x_{t-s+1} + u_{t+s}$$

repeatedly (for $s = 1, 2, \dots$) and recording the $B_{1,s}$ estimates. The impulse response matrices are $B_{1,1}$, $B_{1,2}$, etc.

15.9.5 VAR(1): Implied Autocovariances

The variance-covariance matrix of x_t is (assuming the system is stationary)

$$\text{Var}(x_t) = \sum_{i=0}^{\infty} A^i \Omega A^{i'}, \text{ where } \Omega = \text{Var}(u_t), \quad (15.37)$$

$$\text{Cov}(x_t, x_{t-s}) = A^s \text{Var}(x_t). \quad (15.38)$$

As before, when A is a companion form then some rows and columns Ω are zeros. Typically, we are then also only interested in the upper left corner of $\text{Cov}(x_t, x_{t-s})$.

Remark 15.25 (*Calculating $\text{Var}(x_t)$) The variance-covariance matrix $\text{Var}(x_t) = \Gamma$ can be calculated by iterating over $\Gamma_i = A\Gamma_{i-1}A' + \Omega$ until $\Gamma_i \approx \Gamma_{i-1}$, starting at $\Gamma_0 = \mathbf{0}_{n \times n}$.

Example 15.26 With the AR(2) in Example 15.16, autocovariance 0 to 2 are $(1.964, 1.113, -0.036)$ so the autocorrelations are $(1, 0.567, -0.018)$. See Figure 15.2 for an illustration (notice the markers). Similarly, Figure 15.9 shows the autocovariances for the two-variable VAR(1) discussed in Example 15.18.

Proof (of (15.37)–(15.38)) Similar to the proof of (15.12)–(15.13), substitute for x_{t-1} in (15.29) using the same equation, but lagged once. Keep substituting (for y_{t-2}, \dots) to get $x_t = A^s x_{t-s} + \sum_{i=0}^{s-1} A^i u_{t-i}$. (The last term is $u_t + Au_{t-1} + \dots + A^{s-1}u_{t-s+1}$.) This shows directly that $\text{Cov}(x_t, x_{t-s}) = \text{Cov}(A^s x_{t-s}, x_{t-s}) = A^s \text{Var}(x_t)$, since x_{t-s} is uncorrelated with (u_t, \dots, u_{t-s+1}) . This gives (15.37). Also, assuming the VAR system is stationary and taking the limit as $s \rightarrow \infty$, gives $x_t = \sum_{i=0}^{\infty} A^i u_{t-i}$. Since u_t is iid, the variance is $\sum_{i=0}^{\infty} A^i \Omega A^{i'}$, where $\Omega = \text{Var}(u_t)$, which is (15.38). \square

15.10 Non-stationary Processes

15.10.1 Introduction

A *trend-stationary process* has a (deterministic) trend. The simplest example is

$$y_t = \mu + \beta t + \varepsilon_t \quad (15.39)$$

where ε_t is white noise. It can be made stationary by subtracting the linear trend. We can typically apply all standard econometric methods to such a model, provided we account for the trend. This can be done by either (a) first estimate the trend and then use $y_t - \hat{\beta}t$ in the subsequent analysis; or (b) work with the y_t series directly, but explicitly include a trend variable in the analysis (say, in the regression model).

A *unit root process* has a (random) trend. The simplest example is the *random walk* with drift

$$y_t = \mu + y_{t-1} + \varepsilon_t, \quad (15.40)$$

where ε_t is white noise. The name “unit root process” comes from the fact that the largest eigenvalue of the companion form (the VAR(1) form of the AR(p)) is one. Such a process is said to be integrated of order one, denoted I(1). Most standard statistical econometric methods fail on such data. However, the process can be made stationary by taking first

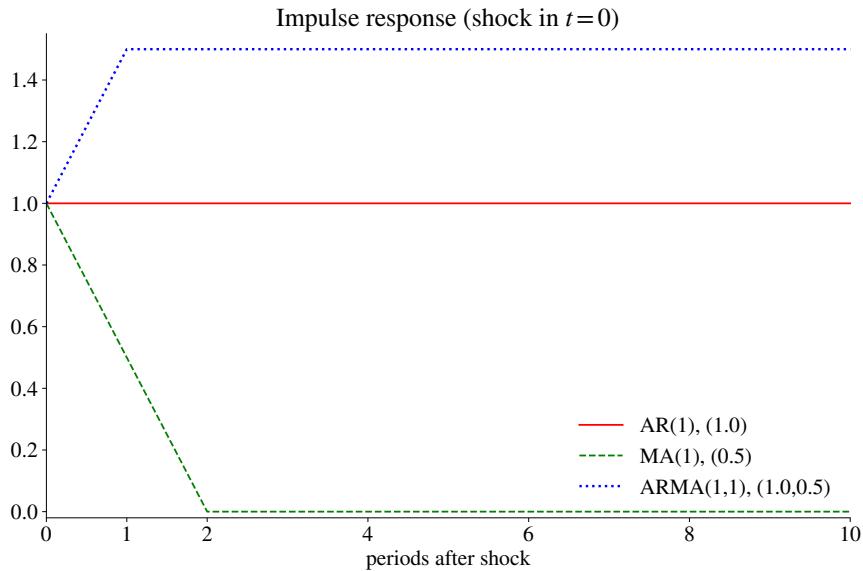


Figure 15.11: Impulse responses

differences

$$y_t - y_{t-1} = \mu + \varepsilon_t. \quad (15.41)$$

Standard methods can readily be applied to $y_t - y_{t-1}$. This is one (of several) reasons why financial econometrics study asset returns (not prices).

Example 15.27 (Non-stationary AR(2)) *The process $y_t = 1.5y_{t-1} - 0.5y_{t-2} + \varepsilon_t$ can be written*

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1.5 & -0.5 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix},$$

where the matrix has the eigenvalues 1 and 0.5: this process is non-stationary. Note that subtracting y_{t-1} from both sides gives $y_t - y_{t-1} = 0.5(y_{t-1} - y_{t-2}) + \varepsilon_t$, so the variable $z_t = y_t - y_{t-1}$ is stationary.

The distinguishing feature of unit root processes is that the effect of a shock never vanishes, that is, the impulse response function does not converge to zero. This is most easily seen for the random walk, where $A^s = 1$ for all s in (15.36). The effect of ε_t never dies out: a non-zero value of ε_t gives a permanent shift of the level of y_{t+s} . See Figure 15.11 for an illustration.

A consequence of the permanent effect of a shock is that the variance of the forecast error grows without bound as the forecasting horizon is extended. For instance, for the

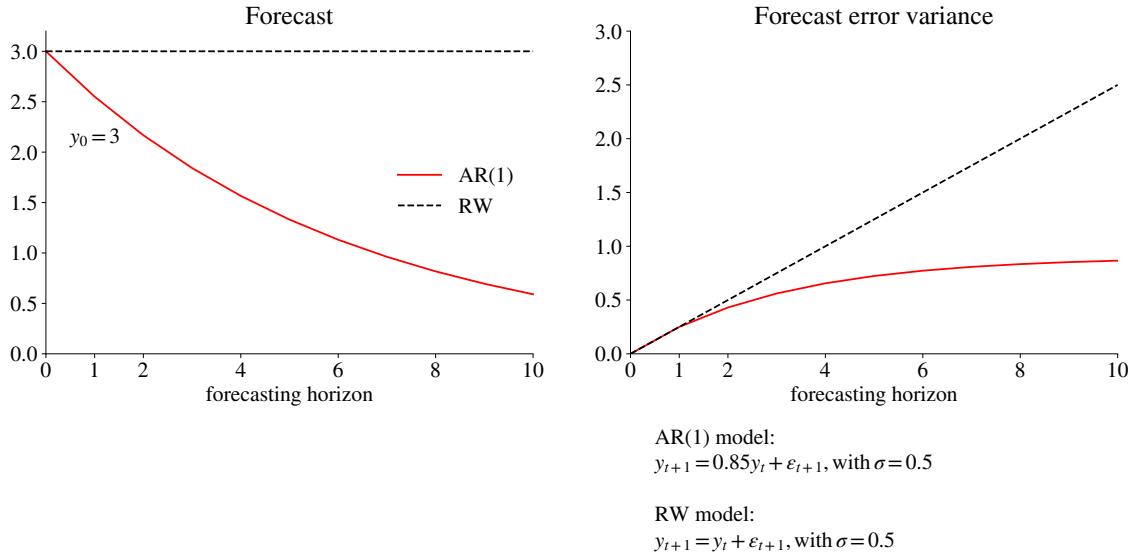


Figure 15.12: Properties of forecasts from random walk process

random walk (15.35), $\text{Var}(y_{t+s} - E_t y_{t+s}) = s\sigma^2$, where σ^2 is the variance of the shock. This means that the the forecast error variance grows linearly with the forecasting horizon. The unconditional variance is therefore infinite and standard results on inference are not applicable. See Figure 15.12.

A process could have r unit roots, that is, be integrated of order r : $I(r)$, or more. In this case, we need to difference r times to make it stationary. Also, a process can also be explosive, that is, have eigenvalues outside the unit circle. In this case, the impulse response function diverges—and this type of data is very difficult to analyse with traditional statistical methods.

Example 15.28 (*Two unit roots**) Suppose y_t in Example (15.27) is actually the first difference of some other series, $y_t = z_t - z_{t-1}$. We then have

$$\begin{aligned} z_t - z_{t-1} &= 1.5(z_{t-1} - z_{t-2}) - 0.5(z_{t-2} - z_{t-3}) + \varepsilon_t \\ z_t &= 2.5z_{t-1} - 2z_{t-2} + 0.5z_{t-3} + \varepsilon_t, \end{aligned}$$

which is an AR(3) with the following canonical form

$$\begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \end{bmatrix} = \begin{bmatrix} 2.5 & -2 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} z_{t-1} \\ z_{t-2} \\ z_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix}.$$

The eigenvalues are 1, 1, and 0.5, so z_t has two unit roots and needs to be differenced twice to become stationary.

Example 15.29 (Explosive AR(1).) Consider the process $y_t = 1.5y_{t-1} + \varepsilon_t$. The eigenvalue is then outside the unit circle, so the process is explosive. This means that the impulse response to a shock to ε_t diverges (it is 1.5^s for s periods ahead).

Remark 15.30 (Lag operator*) A common and convenient way of dealing with leads and lags is the lag operator, L . It is such that $L^s y_t = y_{t-s}$. For instance, the AR(1) model $y_t - \theta y_{t-1} = \varepsilon_t$ can be written as $(1 - \theta L)y_t = \varepsilon_t$. (Often, $\theta(L) = 1 - \theta L$ is called a lag polynomial.) Similarly, an ARMA(2,1) $y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} = \varepsilon_t + \alpha_1 \varepsilon_{t-1}$ can be written $(1 - \theta_1 L - \theta_2 L^2)y_t = (1 + \alpha_1 L)\varepsilon_t$.

15.10.2 Testing for a Unit Root*

Suppose we run an OLS regression of

$$y_t = ay_{t-1} + \varepsilon_t, \quad (15.42)$$

where the true value is $|a| < 1$. The asymptotic distribution of the LS estimator is

$$\sqrt{T}(\hat{a} - a) \sim N(0, 1 - a^2). \quad (15.43)$$

(The variance follows from the standard OLS formula where the variance of the estimator is $\sigma^2 (\Sigma_{t=1}^T x_t x'_t / T)^{-1}$. Here $\text{plim } \Sigma_{t=1}^T x_t x'_t / T = \text{Var}(y_t)$ which we know is $\sigma^2 / (1 - a^2)$).

It is well known (but not easy to show) that when $a = 1$, then \hat{a} is biased towards zero in small samples. In addition, the asymptotic distribution is no longer (15.43). In fact, there is a discontinuity in the limiting distribution as we move from a stationary to a non-stationary variable. This, together with the small sample bias means that we have to use simulated critical values for testing the null hypothesis of $a = 1$ based on the OLS estimate from (15.42).

In practice, the approach is to run the regression (15.42) with a constant (and perhaps even a time trend), calculate the test statistic

$$DF = \frac{\hat{a} - 1}{\text{Std}(\hat{a})}, \quad (15.44)$$

and reject the null of non-stationarity if DF is less than the critical values published by Dickey and Fuller (-2.86 at the 5% level if the regression has a constant, and -3.41 if the

regression includes a trend).

With more dynamics (to capture any serial correlation in ε_t in (15.42)), apply an *augmented Dickey-Fuller test* (ADF)

$$\begin{aligned} y_t &= \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_{2t}, \text{ or} \\ \Delta y_t &= \delta + (\theta_1 + \theta_2 - 1) y_{t-1} - \theta_2 \Delta y_{t-1} + \varepsilon_{2t}, \end{aligned} \quad (15.45)$$

and test the coefficient on y_{t-1} in (15.45) (which should equal $\theta_1 + \theta_2 - 1$) against the alternative that is less than zero. The critical values are as for the DF test. If ε_{2t} is autocorrelated, add further lags of Δy .

The *KPSS test* has stationarity as the null hypothesis (in contrast to the DF and ADF tests that have non-stationarity as the null hypothesis). It has three steps. First, regress

$$y_t = a + \varepsilon_t. \quad (15.46)$$

Second, define

$$S_t = \sum_{s=1}^t \hat{\varepsilon}_s \text{ for } t = 1, \dots, T \text{ and } \hat{\sigma}^2 = \text{Var}(\hat{\varepsilon}_t). \quad (15.47)$$

Third, the test statistic is

$$KPSS = \frac{1}{T^2} \sum_{t=1}^T S_t^2 / \hat{\sigma}^2 \quad (15.48)$$

Reject stationarity if $KPSS > 0.463$ (a 5% critical value). We could also include a linear trend in (15.46). The 5% critical value is then 0.146.

In practice, distinguishing between a stationary and a non-stationary series is very difficult (and impossible unless we restrict the class of processes, for instance, to an AR(2)), since any sample from a non-stationary process can be arbitrary well approximated by *some* stationary process et vice versa. The lesson to be learned, from a practical point of view, is that *strong persistence in the data generating process (stationary or not) invalidates the usual results on inference*. We are usually on safer ground to apply the unit root results in this case, even if the process is actually stationary.

15.10.3 Cointegration*

An exception to the “spurious regression” result: y_t and x_t are I(1) but share a common stochastic trend such that

$$y_t - \alpha - \beta x_t \text{ is I}(0). \quad (15.49)$$

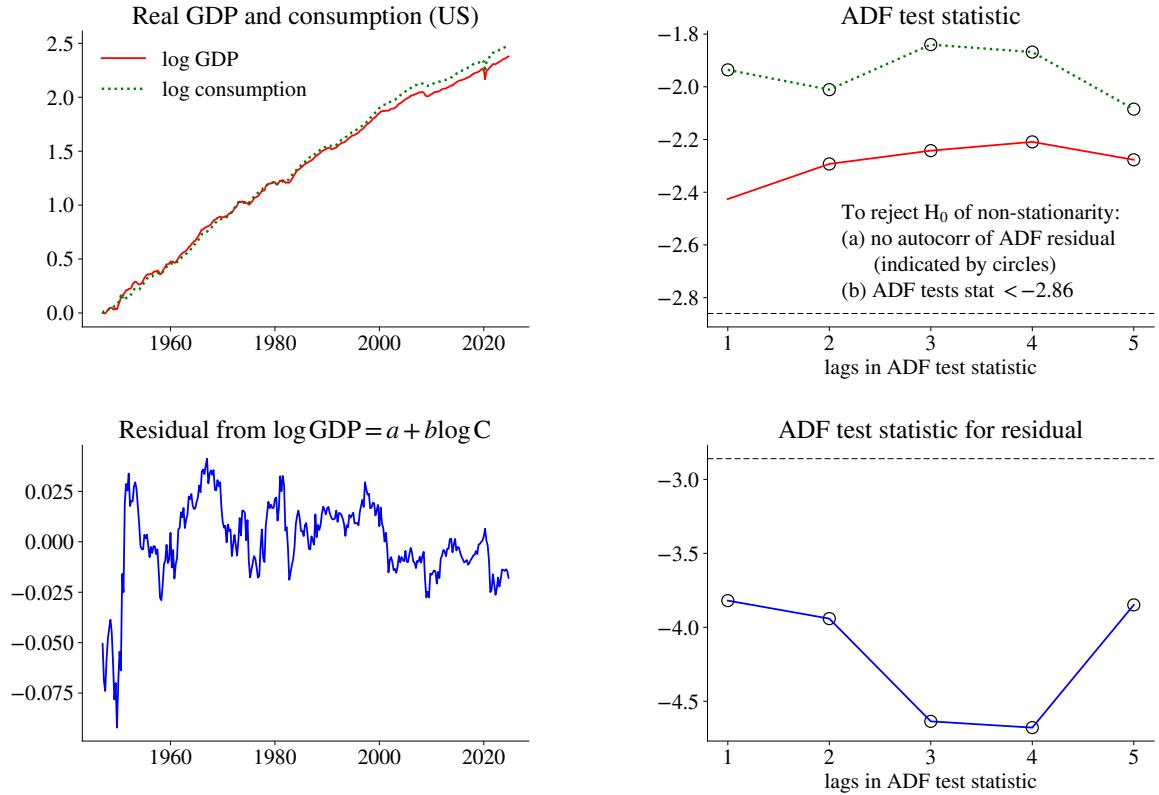


Figure 15.13: Unit root tests, US quarterly macro data

In this case, OLS works fine: it is actually very good (super consistent), $\hat{\beta}$ converges to true value β faster than in standard theory. The intuition is that if $\hat{\beta} \neq \beta$, then ε_t is I(1) and therefore has a high sample variance, so OLS will pick $\hat{\beta}$ close to β .

In (15.49), we call $(1, -\beta)$ the *cointegrating vector*, since

$$\begin{bmatrix} 1 & -\beta \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} \text{ is I}(0). \quad (15.50)$$

To test if y_t and x_t are cointegrated, we need to study three things. First, assess the plausibility: look at data and consider the (economic) theory. Second, validate that both x_t and y_t are I(1), by tests of stationarity. Third, run the regression

$$y_t = a + bx_t + \varepsilon_t, \quad (15.51)$$

and test that $\hat{\varepsilon}_t$ is I(0). This can be done with an ADF test on $\hat{\varepsilon}_t$, but using special critical values, for instance, -3.34 (at the 5% significance level).

One way to incorporate the cointegration in a model of the short-run dynamics is to use a *Error-Correction Model*, for instance,

$$\begin{aligned}\Delta y_t &= \delta - \gamma (y_{t-1} - \beta x_{t-1}) + \phi_1 \Delta x_{t-1} + \varepsilon_t \text{ or perhaps} \\ &= \delta - \gamma (y_{t-1} - \beta x_{t-1}) + \phi_1 \Delta x_{t-1} + \theta_1 \Delta y_{t-1} + \varepsilon_t\end{aligned}\quad (15.52)$$

Recall: (y_t, x_t) are I(1), but $y_{t-1} - \beta x_{t-1}$ is I(0), so all terms in (15.52) are I(0). We typically do not put the intercept into the cointegrating relation (as there is already another intercept in the equation). If $\gamma > 0$ (notice the sign convention in (15.52)), then the system is driven back to a stationary path for $y - \beta x$: the “error correction mechanism.” If ε_t is autocorrelated, then we add more lags of both Δy and Δx .

Estimation is straightforward (Engle-Granger’s 2-step method). First, estimate the cointegrating vector. Second, use it in (15.52) and estimate the rest of the parameters. Standard tests can be applied to them.

Empirical Example 15.31 (*Cointegration and error correction model of output and consumption*) See Figure 15.13 and Table 15.1.

	GDP growth
constant	0.01 (1.91)
Coint res _{t-1}	-0.08 (-2.25)
Δgdp_{t-1}	0.09 (1.01)
Δc_{t-1}	0.05 (0.20)
Δgdp_{t-2}	0.07 (0.83)
Δc_{t-2}	0.06 (0.52)
R^2	0.05
obs	309

Table 15.1: Error-correction model for real US GDP growth, 1947:01-2024:10. Numbers in parentheses are t-stats. The Coint res is the residual from regressing the log GDP level on the log consumption level.

Further Reading

See Verbeek (2017) 8-9, Greene (2018) 20, Hansen (2022a) 14-15, Pesaran (2015) and Hamilton (1994) for many more details.

Chapter 16

Predicting Asset Returns

16.1 Autocorrelations and Autoregression

16.1.1 Autocorrelation Coefficients

Let $\hat{\rho}_s$ be the sample autocorrelation for lag s and T the sample size. The results from the chapter on time series analysis show that (in large samples) $\sqrt{T}\hat{\rho}_s$ is a t-stat for the hypothesis $\rho_s = 0$ and that the Box-Pierce test, $T \sum_{s=1}^L \hat{\rho}_s^2$, is χ_L^2 -distributed under the hypothesis that all autocorrelations are zero.

Empirical Example 16.1 (*Autocorrelations for different lags, daily equity returns*) See Figure 16.1 for autocorrelations (different lags) of daily S&P 500 returns. The figure suggests little autocorrelation in returns (R_t^e), but considerable autocorrelation for the absolute value ($|R_t^e|$). Since, $R_t^e = \text{sign}(R_t^e)|R_t^e|$, this suggests that it is very difficult to predict the sign of the returns. Also, see Figure 16.2 for ten size-sorted equity portfolios which suggests that most size categories have more autocorrelations than large cap (which are fairly close to S&P 500).

16.1.2 Autoregressions

An alternative way of testing autocorrelations is to estimate an AR model

$$y_t = c + a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \varepsilon_t, \quad (16.1)$$

and then test if all slope coefficients (a_1, a_2, \dots, a_p) are zero. This is similar to the Box-Pierce test, since most stationary financial time series processes can be well approximated by an AR of relatively low order.

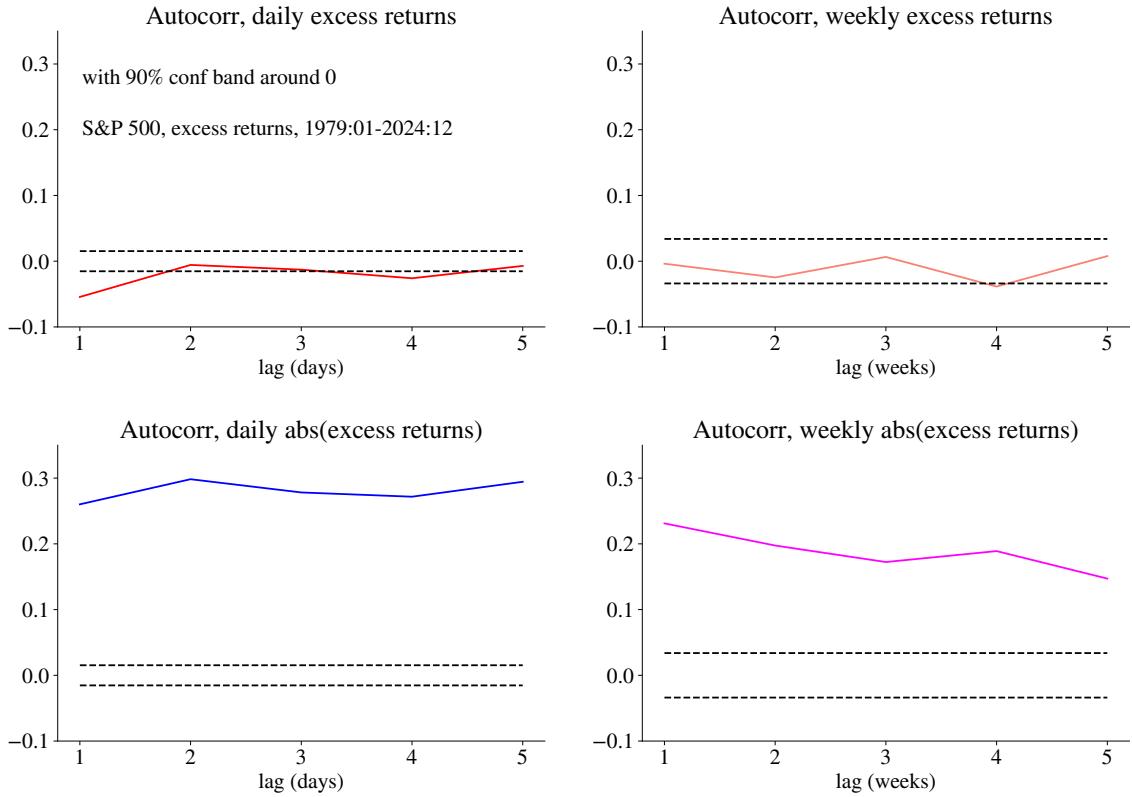


Figure 16.1: Predictability of daily US stock returns

Empirical Example 16.2 (*AR(1) for long run equity returns*) See Figure 16.4 for AR(1) results for different (long) investment horizons. The evidence suggests some negative autocorrelation (mean reversion in the price level) for multi-year return horizons.

The autoregression can also allow the coefficients to depend on the market situation. For instance, an AR(1) where the autoregression coefficient may be different depending on the sign of last period's return

$$y_t = \alpha + \beta \delta_{t-1} y_{t-1} + \gamma (1 - \delta_{t-1}) y_{t-1} + \varepsilon_t, \quad (16.2)$$

where $\delta_{t-1} = 1$ if $y_{t-1} < 0$ and 0 otherwise.

Empirical Example 16.3 (*Asymmetric AR(1) for daily S&P 500 returns*) See Figure 16.3 for an asymmetric AR(1).

If *overlapping returns*, for instance, daily data on 2-day returns, then (16.1) must be adjusted so as to not create a mechanical link. For instance, if r_t is the daily log return,

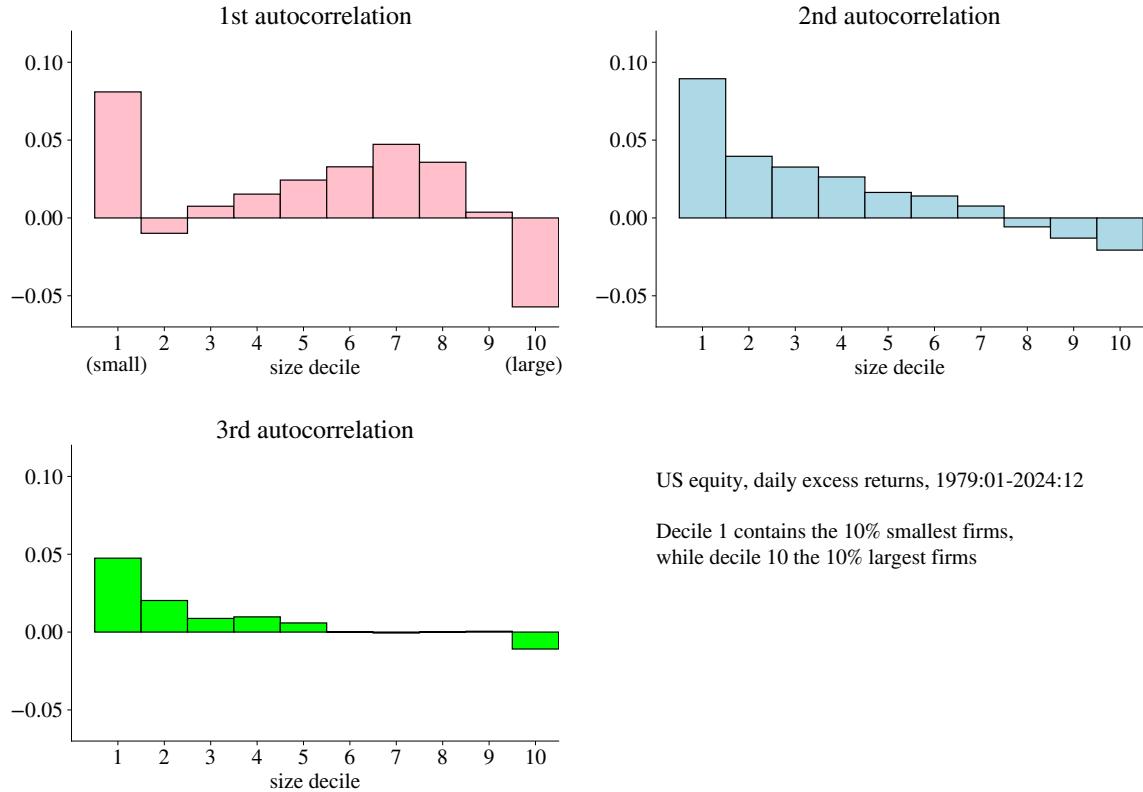


Figure 16.2: Predictability of daily US stock returns, size deciles

then $y_t = r_{t-1} + r_t$ is the daily 2-day return, so an appropriate “AR(1)” is

$$r_{t-1} + r_t = c + a(r_{t-3} + r_{t-2}) + u_t, \text{ that is, } y_t = c + ay_{t-2} + u_t. \quad (16.3)$$

More generally, with q -period returns, the predictor must be lagged at least $q + 1$ times. However, the residuals (u_t) may still be autocorrelated, even under the null hypothesis of iid returns (see the Remark below for an example). This can be handled by either using only non-overlapping returns (in our example, weekly data on weekly returns) or by applying an estimator of the variance-covariance matrix which is robust to autocorrelation, for instance, those of Newey-West or Hansen-Hodrick.

Remark 16.4 (*Overlapping returns**) Consider two successive observations of (16.3)

$$\begin{aligned} r_{t-1} + r_t &= c + a(r_{t-3} + r_{t-2}) + u_t \\ r_t + r_{t+1} &= c + a(r_{t-2} + r_{t-1}) + u_{t+1}. \end{aligned}$$

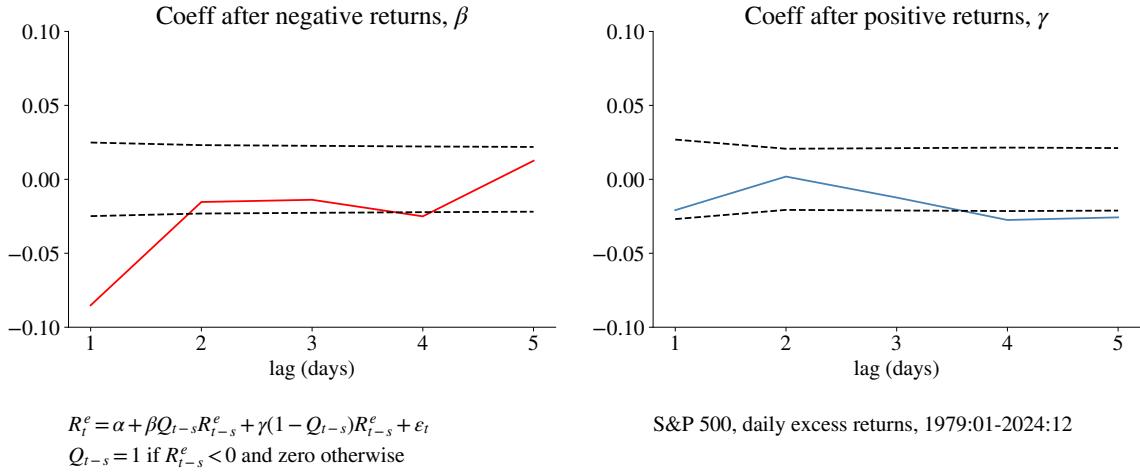


Figure 16.3: Predictability of daily US stock returns, symmetric and asymmetric AR(1)

If returns r_t are iid white noise processes, ε_t , then the true $b = 0$, and $u_t = \varepsilon_{t-1} + \varepsilon_t$ and $u_{t+1} = \varepsilon_t + \varepsilon_{t+1}$. Clearly, $\text{Cov}(u_t, u_{t+1}) = \text{Var}(\varepsilon_t)$, so the regression residuals are autocorrelated. Instead, with non-overlapping data, the second observation is instead

$$r_{t+1} + r_{t+2} = c + a(r_{t-1} + r_t) + u_{t+2}.$$

Since $u_{t+2} = \varepsilon_{t+1} + \varepsilon_{t+2}$, it is not correlated with u_t .

16.1.3 Variance Ratios

A variance ratio is another way to measure predictability. It is defined as the variance of a q -period return divided by q times the variance of a 1-period return

$$VR_q = \frac{\text{Var}(\sum_{s=0}^{q-1} y_{t-s})}{q \text{Var}(y_t)}. \quad (16.4)$$

This measure has considerable appeal in finance, since MV preferences, applied to the choice between a risky asset and risk-free asset, result in a decision rule that depends on the expected q -period return divided by the variance. If the expected returns are the same for different periods, a low variance ratio implies that a long-run investor should invest more in the risky asset (“safe in the long run”) than a one-period investor.

To see that this is related to predictability, notice that the 2-period variance ratio can

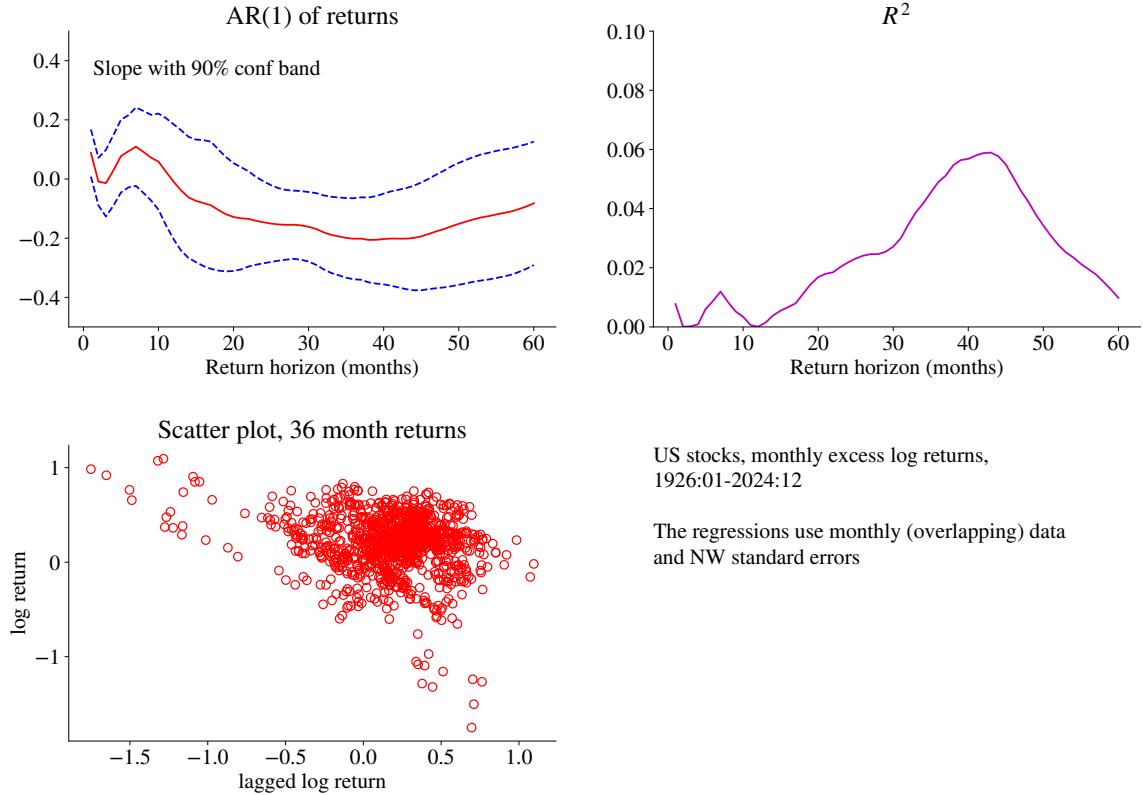


Figure 16.4: Predictability of long run US stock returns

be written as

$$VR_2 = \frac{\text{Var}(y_t + y_{t-1})}{2 \text{Var}(y_t)} = 1 + \rho_1. \quad (16.5)$$

(This follows directly from $\text{Var}(y_t + y_{t-1}) = 2 \text{Var}(y_t) + 2 \text{Cov}(y_t, y_{t-1})$.) It is clear from (16.5) that if y_t is not serially correlated, then the variance ratio is unity; a value above one indicates positive serial correlation and a value below one indicates negative serial correlation. The same applies to longer horizons.

There are two main ways of estimating a variance ratio. *First*, calculate the variance of non-overlapping q -period returns, then of one-period returns, and finally apply (16.4). *Second*, rewrite (16.4) as

$$\begin{aligned} VR_q &= \sum_{s=-(q-1)}^{q-1} (1 - |s|/q) \rho_s \\ &= 1 + 2 \sum_{s=1}^{q-1} (1 - s/q) \rho_s. \end{aligned} \quad (16.6)$$

In this approach, we estimate the autocorrelation coefficients (using all available data

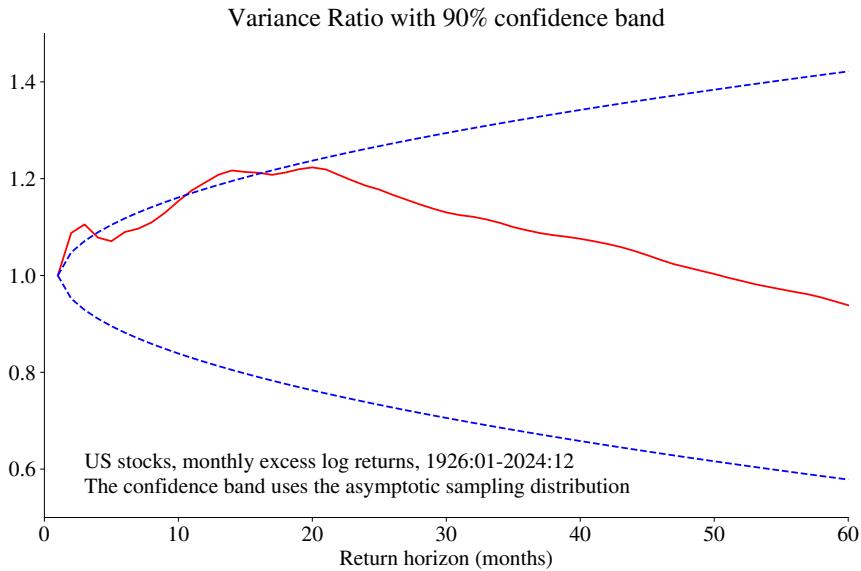


Figure 16.5: Variance ratios, long run US excess stock returns

points for each estimation) and then apply (16.6).

Empirical Example 16.5 (*Variance ratio for long run U.S. equity returns*) See Figure 16.5.

Remark 16.6 (*Sampling distribution of \widehat{VR}_q) Under the null hypothesis that there is no autocorrelation, $\sqrt{T}\hat{\rho}_i$ and $\sqrt{T}\hat{\rho}_j$ are independent $N(0, 1)$ variables. Together with (16.6) we get

$$\sqrt{T}(\widehat{VR}_q - 1) \xrightarrow{d} N(0, \Sigma_{s=1}^{q-1} 4(1 - s/q)^2).$$

For instance, for $q = 2$ the variance is 1 and for $q = 3$ it is $20/9$. (The variance can also be written as $(q - 1)(2q - 1)2/(3q)$.)

16.2 Other Predictors and Methods

Predictability and autocorrelation are not necessarily synonymous: although autocorrelation implies predictability, we can have predictability without autocorrelation, for instance, by using other predictors.

There are many other possible predictors of future stock returns. For instance, lagged short-run returns on other assets have been used to predict short-run returns, and both the dividend-price ratio and nominal interest rates have been used to predict long-run returns.

16.2.1 Lead-Lags

Stock indices have more positive autocorrelation than (most) individual stocks: there should therefore be cross-autocorrelations across individual stocks. (See Campbell, Lo, and MacKinlay (1997) 2.)

Empirical Example 16.7 (*Correlations of $R_{i,t}$ and $R_{j,t-s}$*) Figure 16.6 shows results from augmented AR(1) estimations for each of the ten size-sorted equity portfolios: the lagged daily return of the largest firms (decile 10) is added as a regressor. It appears important.

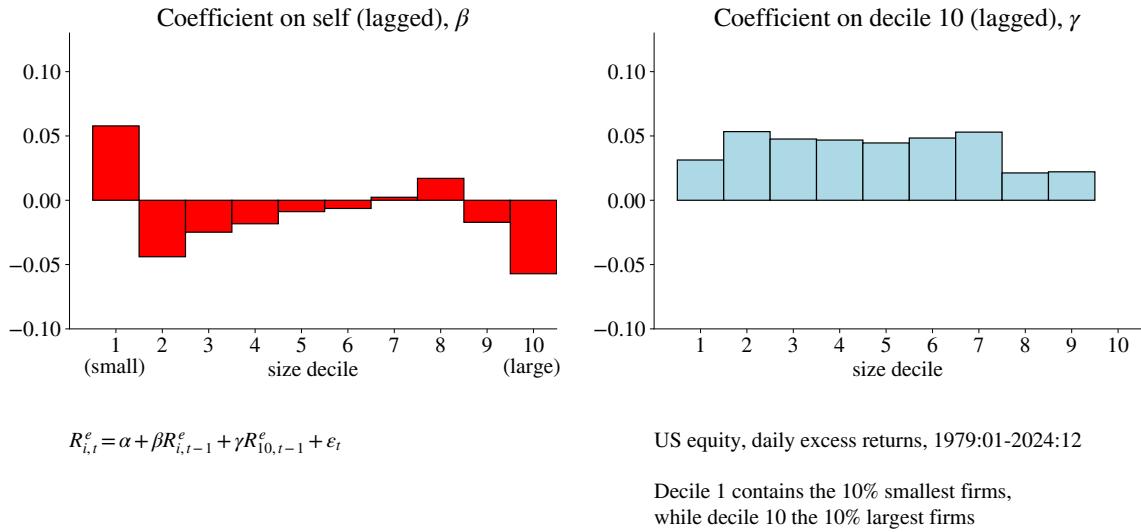


Figure 16.6: Spillover effects, daily data

16.2.2 Dividend-Price Ratio as a Predictor

One of the most successful attempts to forecast long-run returns is based on a current valuation ratio, for instance, the dividend-price ratio or the earnings-price ratio, for instance

$$R_{t+s} = \alpha + \beta_s \ln(E_t/P_t) + \varepsilon_{t+s}, \quad (16.7)$$

where R_{t+s} is the return from t to $t + s$.

Empirical Example 16.8 (*Predicting long run equity returns with E/P*) See Figure 16.7. The results show significant predictability for multi-year returns, but the R^2 values are low.

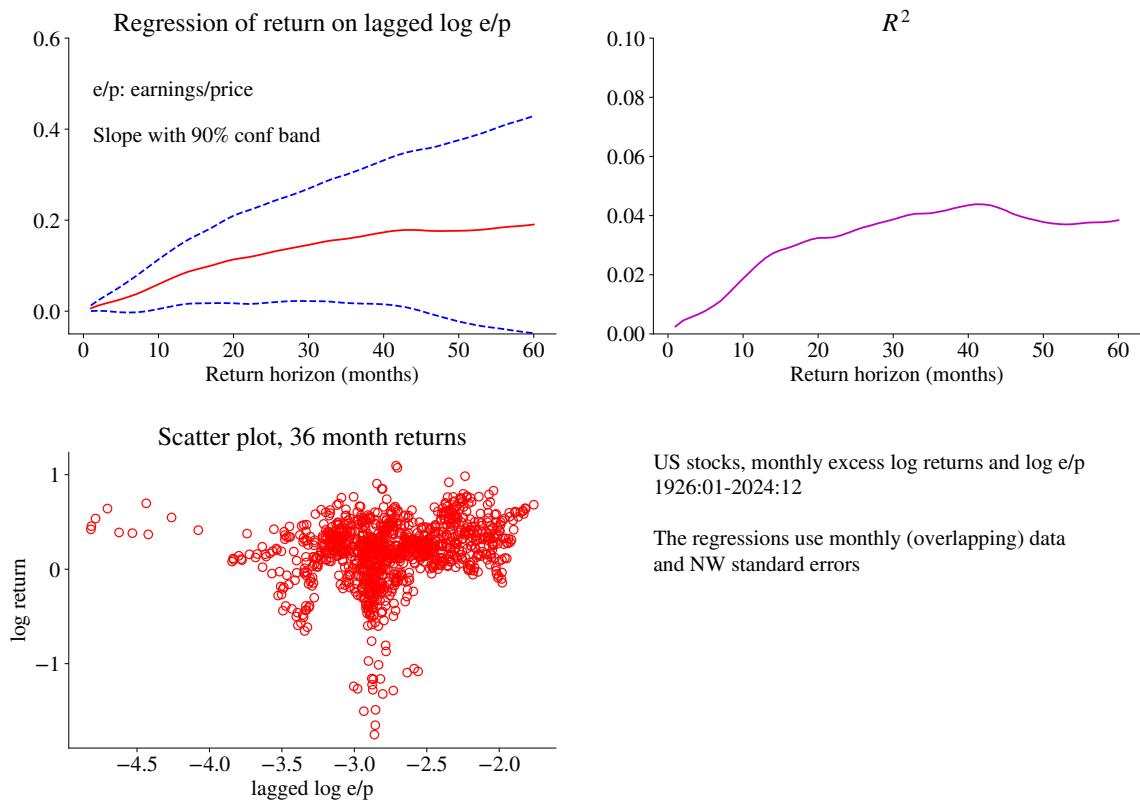


Figure 16.7: Forecast regressions, long run US stock returns

16.3 Out-of-Sample Forecasting Performance

16.3.1 In-Sample versus Out-of-Sample Forecasting

References: Goyal and Welch (2008), and Campbell and Thompson (2008)

In-sample evidence on predictability may suffer from several problems. First, the link between the predictor and future returns may be unstable (so the model has “breaks”). Second, if the estimated (in-sample) model includes many predictors, then it is likely to give poor predictions due to in-sample “overfitting.”

To gauge the out-of-sample predictability, we could estimate the prediction equation using data up to and including $t - 1$, and then make a forecast for period t . Repeating this for further periods gives a series of out-of-sample forecast errors. See Figure 16.8 for an illustration.

The out-of-sample forecasting performance is then compared with a benchmark prediction model, for instance, the historical average. Notice that this benchmark model is

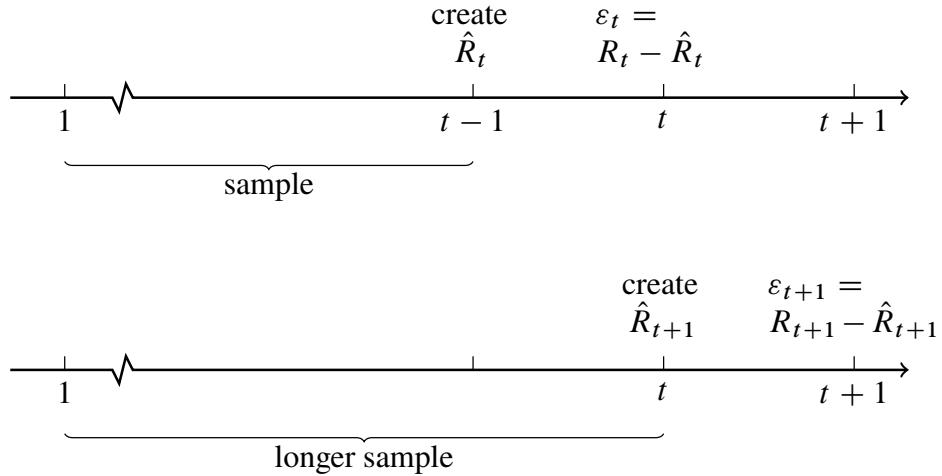


Figure 16.8: Out-of-sample forecasting

also estimated on data up to and including $t-1$, so it changes over time. Thus, this approach compares the forecast performance of two models estimated in a recursive way (increasing sample size).

It is common to make this comparison in terms of the RMSE and an “out-of-sample R^2 ”

$$R_{OS}^2 = 1 - \frac{1}{T} \sum_{t=s}^T (y_t - \hat{y}_t)^2 / \frac{1}{T} \sum_{t=s}^T (y_t - \tilde{y}_t)^2, \quad (16.8)$$

where s is the first period with an out-of-sample forecast, \hat{y}_t is the forecast based on the prediction model (estimated on data up to and including $t-1$) and \tilde{y}_t is the prediction from some benchmark model (also estimated on data up to and including $t-1$).

Example 16.9 (R_{OS}^2)

$$R_{OS}^2 = 1 - 0.4/0.5 = 0.2 \text{ (better than benchmark)}$$

$$R_{OS}^2 = 1 - 0.5/0.4 = -0.25 \text{ (worse than benchmark)}$$

Empirical Example 16.10 (Out-of-sample prediction of equity returns) Figure 16.9 shows results for daily size-sorted equity returns. There is some short-run predictability for small firm returns also out-of-sample. Instead, Figure 16.10 shows results on predicting long run equity returns with E/P. The evidence suggests that the in-sample long-run predictability vanishes out-of-sample.

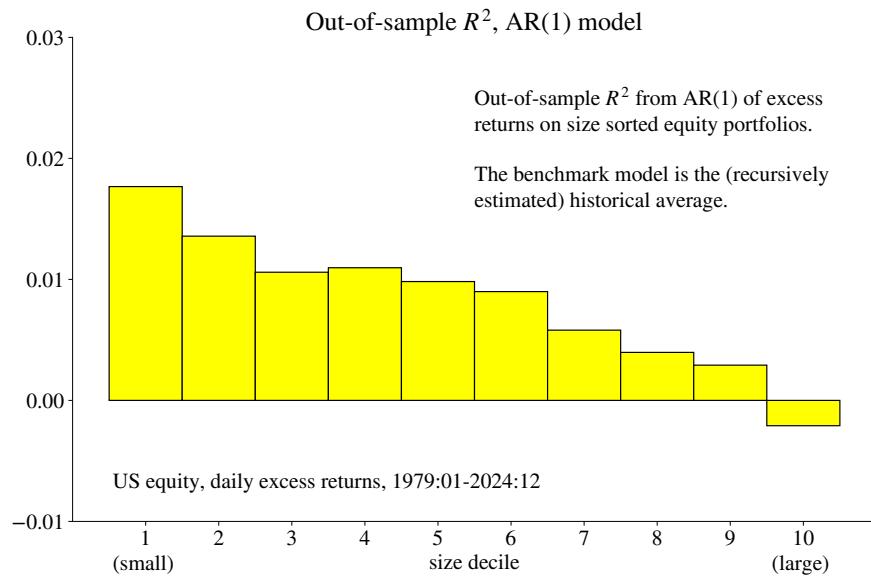


Figure 16.9: Predictability of daily US stock returns, out-of-sample

16.3.2 Trading Strategies

Another way to measure predictability, and to illustrate its economic importance, is to calculate the return of a *dynamic trading strategy*, and then measure the performance. The trading strategy should, of course, be based on the variable that is supposed to forecast returns.

The performance is often measured in terms of the average excess return, Sharpe ratio of the alpha from the factor model

$$R_{pt}^e = \alpha + \beta' R_{bt}^e + \varepsilon_t, \quad (16.9)$$

where R_{pt}^e is the excess return on the portfolio being studied and R_{bt}^e a vector of benchmark excess returns, for instance, just the market excess return, if we want to rely on CAPM. Neutral performance requires $\alpha = 0$, which can be tested with a t test.

Empirical Example 16.11 (*Momentum for daily returns on the 25 FF portfolios*) *Figure 16.11 suggests that there is considerable momentum in the cross-section of the 25 FF portfolios. Investing in past winners earns high returns.*

Empirical Example 16.12 (*Mean reversion of daily S&P 500 returns*) *Figure 16.12 shows that extreme S&P 500 returns are followed by mean-reverting movements the*

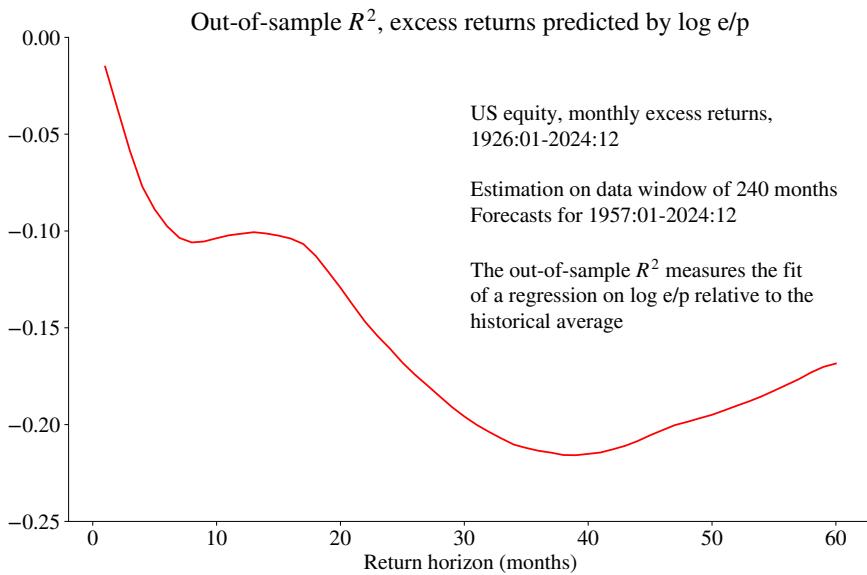


Figure 16.10: Predictability of long run US stock returns, out-of-sample

following day (negative autocorrelation)—which suggests that a trading strategy should sell after a high return and buy after a low return. Compare with Figure 16.2.

Empirical Example 16.13 (*Long run S&P 500 after different p/e values*) Figure 16.13 shows average one-year return on S&P 500 for different bins of the p/e ratio (at the beginning of the year). The figure illustrates that buying when the market is undervalued (low p/e) might be a winning strategy. Compare with Figure 16.7 (in-sample, but there we used e/p , not p/e) but also Figure 16.10 (out-of-sample evaluation of regression results).

16.3.3 Technical Analysis

Main reference: Bodie, Kane, and Marcus (2002) 12.2; Reilly and Brown (2012) 16; Brock, Lakonishok, and LeBaron (1992); Lo, Mamaysky, and Wang (2000)

General Idea of Technical Analysis

Technical analysis is typically a data mining exercise which looks for local trends or systematic non-linear patterns. The basic idea is that markets are not instantaneously efficient: prices react somewhat slowly and predictably to news, perhaps because of market psychology. In practice, much of technical analysis earlier amounted to plotting

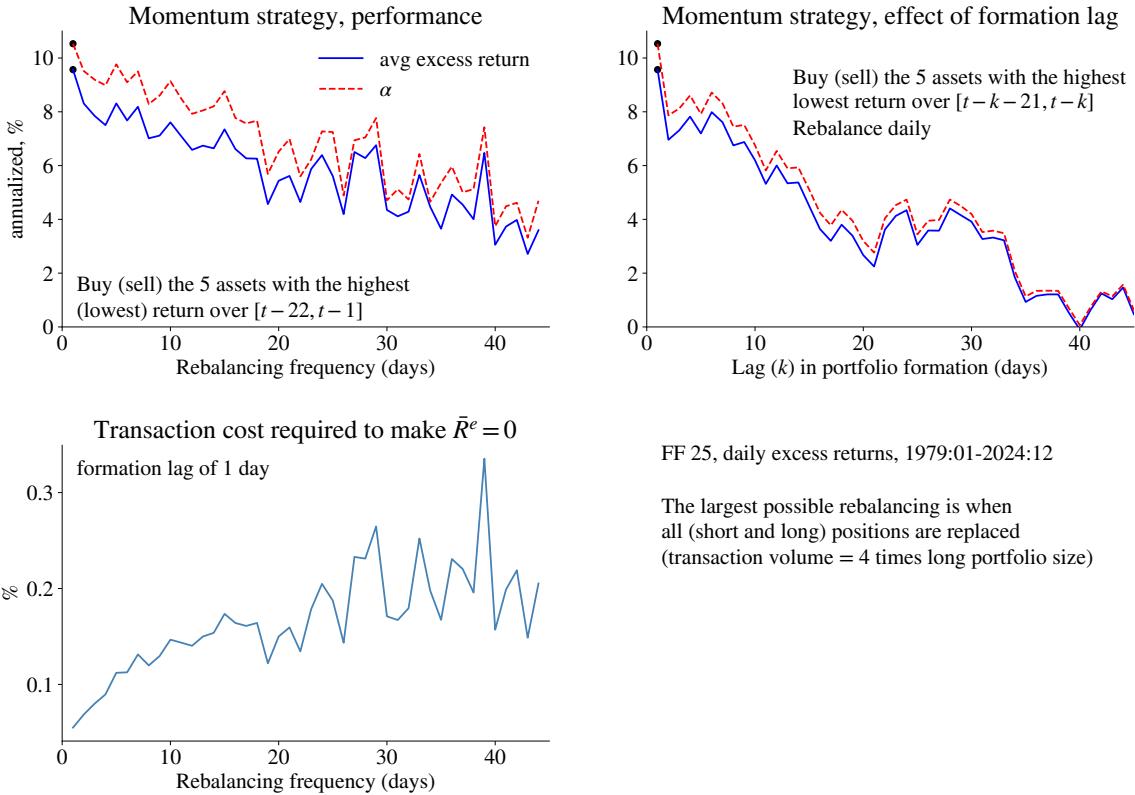


Figure 16.11: Predictability of daily US stock returns, momentum strategy

different transformations (for instance, a moving average) of prices, and to spot systematic patterns. More recently, elements of technical analysis have been incorporated into trading algorithms. It is also common to use information from other markets (for instance, a high CBOE put/call ratio is often interpreted as bearish sentiment), trading volume, or measures of market wide trends (the “breadth” of the market compares the number of assets with price increase/decrease).

Technical Analysis and Local Trends

Many trading rules rely on some kind of local trend which can be thought of as positive autocorrelation in price movements, also called momentum. (In physics, momentum equals the mass times speed.)

A *moving average rule* is to buy if a short moving average (equally weighted or exponentially weighted) goes above a long moving average. The idea is that the event signals a new upward trend. Let $S(L)$ be the lag order of a short (long) moving average,

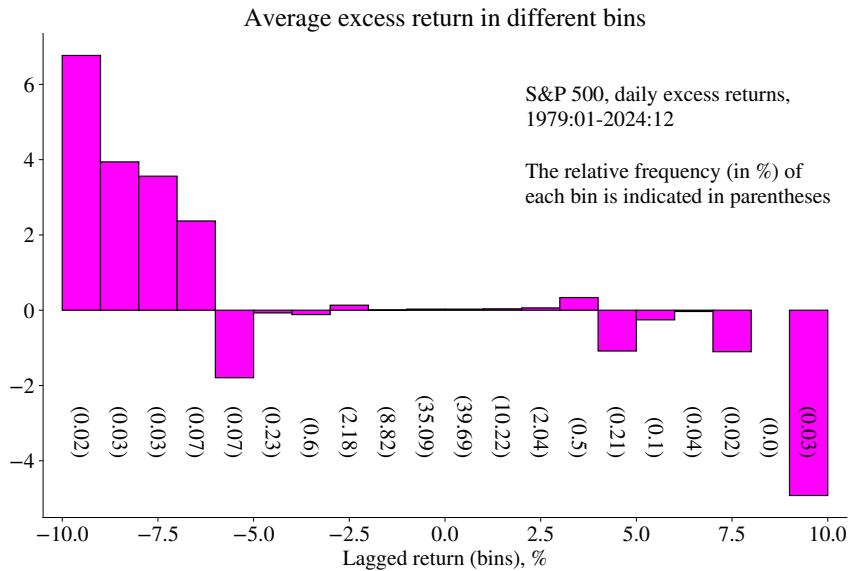


Figure 16.12: Predictability of daily US stock returns, out-of-sample

with $S < L$ and let b be a bandwidth (perhaps 0.01). Then, a MA rule for period t could be

$$\begin{aligned} \text{buy in } t \text{ if } & MA_{t-1}(S) > MA_{t-1}(L)(1 + b) \\ \text{sell in } t \text{ if } & MA_{t-1}(S) < MA_{t-1}(L)(1 - b) , \\ \text{no change} & \quad \text{otherwise} \end{aligned} \quad (16.10)$$

where $MA(K)$ is a unweighted or exponential moving average. (In the latter case, $MA_t = (1 - \lambda)P_t + \lambda MA_{t-1}$, with $\lambda = (L - 1)/(L + 1)$.) If we instead believe in *mean reversion* of the prices, then we can essentially reverse the previous trading rule.

The difference between the two moving averages is called an *oscillator*

$$\text{oscillator}_t = MA_t(S) - MA_t(L), \quad (16.11)$$

(or sometimes, moving average convergence divergence, or MACD) and the sign is often taken as a trading signal (this is the same as a moving average crossing, MAC). A version of the moving average oscillator is the *relative strength index*, which is the ratio of average price level (or returns) on “up” days to the average price (or returns) on “down” days—during the last z (14 perhaps) days. (Not to be confused with relative strength, which typically refers to the ratio of two different asset prices, for instance, an equity compared to the market.) Yet another version is to compare the oscillator_t to an moving average of the oscillator_t (also called a signal line).

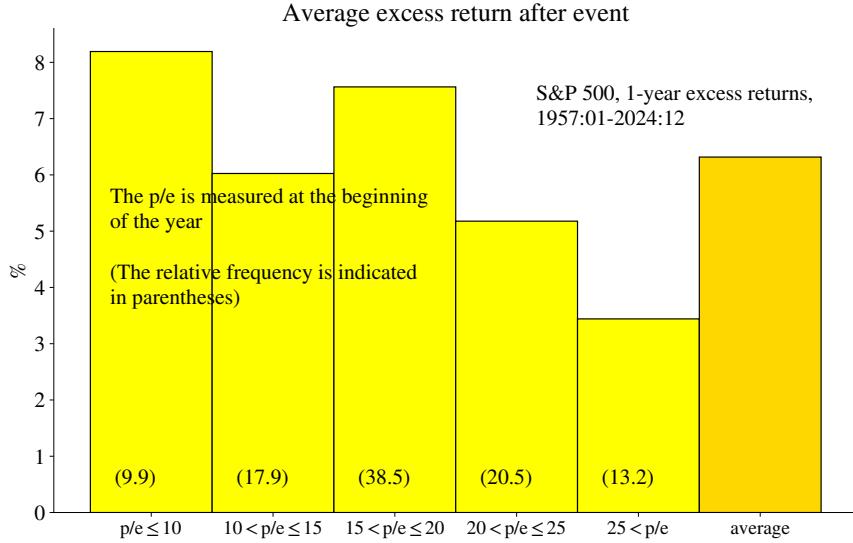


Figure 16.13: Predictability of annual US stock returns, out-of-sample

The *trading range break-out rule* typically amounts to buying when the price rises above a previous peak (local maximum). The idea is that a previous peak is a *resistance level* in the sense that some investors are willing to sell when the price reaches that value. Round numbers often play the role as resistance levels. Once this resistance level has been broken, the price can possibly rise substantially and thus motivate a buy. On the downside, a *support level* plays the same role: some investors are willing to buy when the price reaches that value. To implement this, it is common to let the resistance/support levels be approximated by minimum and maximum values over a data window of length L . With a bandwidth b (perhaps 0.01), the rule for period t could be

$$\begin{aligned} \text{buy in } t \text{ if } & P_t > (1 + b) \max(p_{t-1}, \dots, p_{t-S}) \\ \text{sell in } t \text{ if } & P_t < (1 - b) \min(p_{t-1}, \dots, p_{t-S}) \\ \text{no change} & \text{ otherwise.} \end{aligned} \tag{16.12}$$

This rule essentially bets on mean reversion between the resistance and support levels, followed by momentum once these levels have been passed.

When the price is already trending up, then the trading range break-out rule may be replaced by a *channel rule*, which works as follows. First, draw a *trend line* through previous lows and a *channel line* through previous peaks. Extend these lines. If the price moves above the channel (band) defined by these lines, then buy. A version of this is to define the channel by a *Bollinger band*, which is ± 2 standard deviations from a moving

data window around a moving average.

A *head and shoulder* pattern is a sequence of three peaks (left shoulder, head, right shoulder), where the middle one (the head) is the highest, with two local lows in between on approximately the same level (neck line). (Easier to draw than to explain in words.) If the price subsequently goes below the neckline, then it is thought that a negative trend has been initiated. (An inverse head and shoulder has the inverse pattern.)

Clearly, we can replace “buy” in the previous rules with something more aggressive, for instance, replace a short position with a long.

The trading volume is also often taken into account. If the trading volume of assets with declining prices is high relative to the trading volume of assets with increasing prices, then this is interpreted as a market with selling pressure. (The possible problem with this interpretation is that there is a buyer for every seller.)

Empirical Example 16.14 (*Daily trading strategy for S&P 500 and risk-free*) See Figure 16.14 for an illustration (over a short subsample) of an inverted MA strategy: buy when the recent index values leave the band on the downside et vice versa. The performance (over a longer sample) is reported in Table 16.1, which suggests that buy and sell signals indeed contain information. However, a daily rebalancing might be too costly (transaction costs).

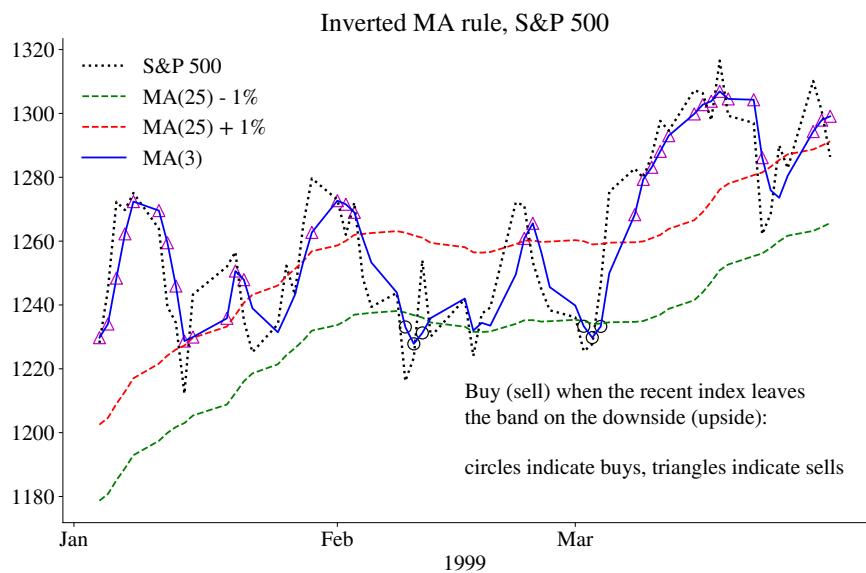


Figure 16.14: Example of a daily trading rule

	Mean	Std
All days	7.2	18.0
After buy signal	16.2	27.3
After neutral signal	4.8	14.7
After sell signal	4.4	13.5
Strategy	9.1	27.5
Transaction cost	0.1	

Table 16.1: Excess returns (annualized, in %) from technical trading rule (Inverted MA rule). Daily S&P 500 data 1990:01-2024:12. The trading strategy involves (a) on everay day: hold one unit of the index and short the riskfree; (b) on days after a buy signal: double the position in (a); (c) on days after a sell signal: short sell the position in (a), effectively having a zero investment. The transaction costs shows the cost (in %) of the trade volume that the strategy can pay and still perform as well as the static holding of (a).

16.4 Forecast Averaging

Reference: Elliot and Timmermann (2016) 14

Averaging across forecasts has often proved to be a good way of producing a better forecast. There are two main cases: (1) when we have access to the forecasts only; and (2) when we have access to the forecasts and also the data/model that produced them.

Suppose we have access to K different forecasts made in $t - h$, \hat{R}_t^i for $i = 1$ to K , of the return R_t . We could form a weighted average as

$$R_t^* = \sum_{i=1}^K w_i \hat{R}_t^i, \text{ with } \sum_{i=1}^K w_i = 1. \quad (16.13)$$

For instance, the weights w_i may be chosen as to minimize the forecast error variance or the MSE over the sample up to and including $t - h$. In practice, it seems difficult to beat an unweighted average or an unweighted average after having pruned the most extreme forecasts (“trimmed mean”).

Instead, suppose we have access also to the models and data that produce the different forecasts. It can then be argued that the proper way to proceed is to pool all the data and apply traditional model selection techniques. However, the average discussed above (16.13) is often a hard to outperform.

Remark 16.15 (*Minimising the MSE) Let Σ be the variance-covariance matrix of the forecast errors from K different models. If the forecasts are unbiased (so the forecast errors have zero means), then the MSE of a combined forecast is $w' \Sigma w$. Therefore, minimize

$w' \Sigma w / 2 + \lambda(1 - \mathbf{1}'w)$ with respect to w to get the first order conditions $\Sigma w = \mathbf{1}\lambda$ and $1 = \mathbf{1}'w$, which together imply $w = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$.

16.5 Evaluating Forecasting Performance

Further reading: Diebold (2001) 11; Stekler (1991); Diebold and Mariano (1995); Clark and West (2007)

To evaluate forecasting performance, we need a sample (history) of the forecasts and the resulting forecast errors. A few periods are not enough, since the results could be driven by luck.

To test forecasting performance, let ε_t be the forecast error in period t

$$\varepsilon_t = R_t - \hat{R}_t, \quad (16.14)$$

where \hat{R}_t is the forecast (made in $t - h$) and R_t the actual outcome. (Warning: some authors prefer to work with $\hat{R}_t - R_t$ as the forecast error instead.)

The forecast is often compared to a benchmark forecast, such as a ‘no-change’ model, a random walk, or the historical average. The idea of such a comparison is to study if the resources employed in creating the forecast really bring value added compared to a very simple (and inexpensive) forecast.

Ultimately, the ranking of forecasting methods by assessing the benefits/costs of forecast errors, which may differ between organizations. For instance, a forecasting agency has a reputation (and eventually customers) to lose, while an investor has more immediate pecuniary concerns. Unless the relation between the forecast error and the losses are immediately understood, the ranking of two forecast methods is typically done based on a number of standard criteria. Several of those criteria are inspired by basic statistics.

Most statistical forecasting methods are based on the idea of minimizing the sum of squared forecast errors, $\sum_{t=1}^T \varepsilon_t^2$. For instance, the least squares (LS) method picks the regression coefficient in

$$R_t = \beta_0 + \beta_1 x_{t-h} + u_t \quad (16.15)$$

to minimize the sum of squared residuals. This will, among other things, yield fitted residuals with a zero mean, and also a zero correlation with the regressor. As usual, rational forecasts should have forecast errors that cannot be predicted (by past regressors or forecast errors).

The evaluation of a forecast often involves extending these ideas, irrespective of

whether a LS regression has been used or not. This means studying the following properties of the forecast errors ε_t in (16.14): (i) whether it has a zero mean; (ii) the mean squared (or absolute value); (iii) the frequency the square (or absolute value) is lower than some threshold; and (iv) whether it is correlated with past information.

Consider two different forecast errors: from the model we want to evaluate (e_t), and from a benchmark (\tilde{R}_t). Comparing the two models according to the above critera could be done by defining either of the following moment conditions

$$g_t = e_t - \varepsilon_t \quad (16.16)$$

$$g_t = e_t^2 - \varepsilon_t^2 \text{ or } g_t = |e_t| - |\varepsilon_t|, \quad (16.17)$$

$$g_t = \delta[\text{sign}(\tilde{R}_t) \neq \text{sign}(R_t)] - \delta[\text{sign}(\hat{R}_t) \neq \text{sign}(R_t)] \quad (16.18)$$

$$g_t = e_t x_{t-h} - \varepsilon_t x_{t-h}, \quad (16.19)$$

where \tilde{R}_t is the forecast from the benchmark model and $\delta(x) = 1$ if x is true and zero otherwise.

The different moment conditions correspond to the different aspects of the forecasts discussed above. For instance, (16.16) is for testing if the two methods have the same average forecast error, while (16.17) tests the MSE or MAD. In contrast, (16.18) tests if the e model forecasts the wrong sign of the return more often than the ε model does, and (16.19) whether the forecast errors are correlated with past information.

From the usual properties of a sample average and the assumption that g_t is not autocorrelated, we typically have that

$$\bar{g} \xrightarrow{d} N(0, \text{Var}(g_t)/T), \quad (16.20)$$

where $\bar{g} = \sum_{t=1}^T g_t / T$ is the average. Alternatively, the variance could be estimated by a Newey-West approach. This is the Diebold and Mariano (1995) test.

However, when the models behind e and ε are *nested* (say, e is generated by a special case of the model that generates ε), then the asymptotic distribution is non-normal so an adjustment must be applied (see Clark and McCracken (2001) and Clark and West (2007)), which typically means studying $g_t = 2e_t(e_t - \varepsilon_t)$ instead of $g_t = e_t^2 - \varepsilon_t^2$.

Table 16.2 shows Monte Carlo simulations of the t-stat of equal predictive performance. The data is iid and the smaller model is the historical average, while the larger model is a regression on a constant and an irrelevant variable. The results indicate that the distribution of the Diebold-Mariano t-stats is shifted to the left and also has too low variance (it should

ideally be an $N(0, 1)$), irrespective of whether we consider a recursive (longer and longer sample) or a moving data window. The Clark-West approach improves, but there is still a small shift to the left.

	DM recursive	CW recursive	DM windowed	CW windowed
mean	-0.94	-0.34	-1.30	-0.19
std	0.76	0.97	0.76	0.99
percdentile 5	-2.16	-1.88	-2.57	-1.82
percdentile 95	0.36	1.32	-0.06	1.44

Table 16.2: Simulation results for the Diebold-Mariano (DM) and Clark-West (CW) t-stats, using recursive estimation or moving windows. Sample size 300, first estimation starts at obs 50. In the application with moving windows, each data window has 50 observations. Data is iid noise, the smaller model uses a regression on a constant and the larger model a regression on a constant and an irrelevant variable.

	AR(1)		E/P		Combination	
	mean	t-stat	mean	t-stat	mean	t-stat
MSE in-sample	294.63		286.04			
R^2_{oos}	-0.04		-0.07		-0.03	
$e - \varepsilon$	0.23	2.05	-1.53	-1.62	-0.65	-1.35
$e^2 - \varepsilon^2$	-11.33	-1.44	-20.40	-0.93	-8.11	-0.76
$ e - \varepsilon $	-0.17	-1.24	-0.82	-1.23	-0.32	-0.94
$2e(e - \varepsilon)$	-10.40	-1.36	9.34	0.42	-0.53	-0.05

Table 16.3: Mariano-Diebold (and Clark-West) tests of forecasting 1-year S&P returns with different models. The total sample is 1946–2024, but the forecasts are made for 1971–2024. The e forecasts are the historical average returns while the ε forecasts are out-of-sample and based on the different regressions. Estimation is done on an expanding data window. The std use a NW approach with 1 lag (year).

Empirical Example 16.16 (*Empirical results on predicting annual S&P 500 returns*)
Table 16.3 summarizes results for predicting 1-year S&P 500 returns. The combined model seems to do slightly better than the two individual models, AR(1) and E/P regression.

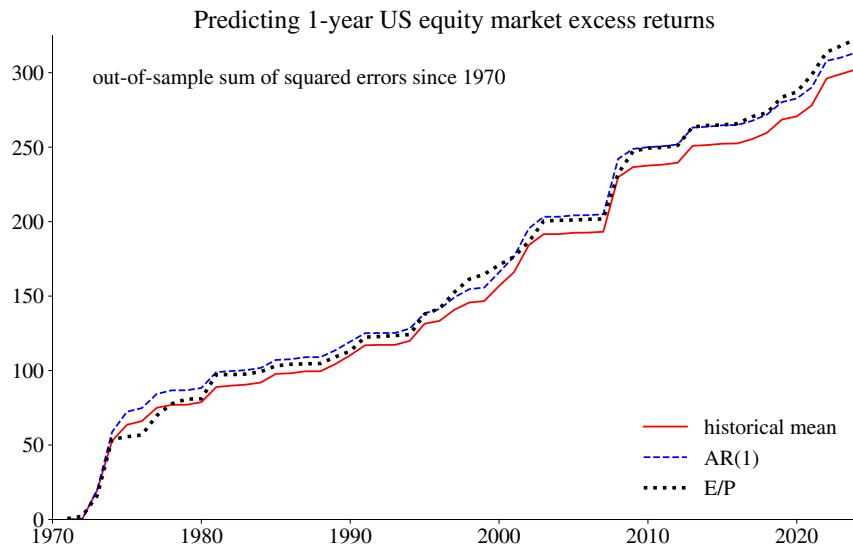


Figure 16.15: Accumulation of the (oos) MSE from three different forecasting models

16.6 Security Analysts

16.6.1 Evidence on Analysts' Performance

Makridakis, Wheelwright, and Hyndman (1998) 10.1 shows that there is little evidence that the average stock analyst beats (on average) the market (a passive index portfolio). In fact, less than half of the analysts beat the market. However, there are analysts which seem to outperform the market for some time, but the autocorrelation in over-performance is weak. The evidence from mutual funds is similar. For them it is typically also found that their portfolio weights do not anticipate price movements.

It should be remembered that many analysts also are sales persons: either of a stock (for instance, since the bank is underwriting an offering) or of trading services. It could well be that their objective function is quite different from minimizing the squared forecast errors—or whatever we typically use in order to evaluate their performance.

16.6.2 Do Security Analysts Overreact?

The paper by Bondt and Thaler (1990) compares the (semi-annual) forecasts (one- and two-year time horizons) with actual changes in earnings per share (1976-1984) for several hundred companies. The paper has regressions like

$$\text{Actual change} = \alpha + \beta(\text{forecasted change}) + \text{residual},$$

and then studies the estimates of the α and β coefficients. With rational expectations (and a long enough sample), we should have $\alpha = 0$ (no constant bias in forecasts) and $\beta = 1$ (proportionality).

The main result is that $0 < \beta < 1$, so that the forecasted change tends to be too wild in a systematic way: a forecasted change of 1% is (on average) followed by a less than 1% actual change in the same direction. This means that analysts in this sample tended to be too extreme—to exaggerate both positive and negative news.

16.6.3 High-Frequency Trading Based on Recommendations from Stock Analysts

Barber, Leavy, McNichols, and Trueman (2001) give a somewhat different picture. They focus on the profitability of a trading strategy based on the recommendations of analysts. They use a huge data set (some 360,000 recommendations, US stocks) for the period 1985-1996. They sort stocks into five portfolios depending on the consensus (average) recommendation—and redo the sorting every day (if a new recommendation is published). They find that such a daily trading strategy yields an annual abnormal return of 4% on the portfolio of the most highly recommended stocks, and an annual -5% abnormal return on the least favourably recommended stocks.

This strategy requires a lot of trading (a turnover of 400% annually), so trading costs would typically reduce the abnormal return on the best portfolio to almost zero. A less frequent rebalancing (weekly, monthly) gives a very small abnormal return for the best stocks, but still a negative abnormal return for the worst stocks. Chance and Hemler (2001) obtain similar results when studying the investment advice by 30 professional “market timers.”

16.6.4 Economic Experts

Several papers, for instance, Bondt (1991) and Söderlind (2010), have studied whether economic experts can predict the broad stock markets. The results suggests that they cannot. For instance, Söderlind (2010) shows that the economic experts that participate in the semi-annual Livingston survey (mostly bank economists) forecast the S&P worse than the historical average (recursively estimated), and that their forecasts are strongly correlated with recent market data (which in itself, cannot predict future returns).

16.6.5 Bond Rating Agencies versus Stock Analysts

Ederington and Goh (1998) use data on all corporate bond rating changes by Moody's between 1984 and 1990 and the corresponding earnings forecasts (by various stock analysts).

The idea of the paper by Ederington and Goh (1998) is to see if bond ratings drive earnings forecasts (or vice versa), and if they affect stock returns (prices).

1. To see if stock returns are affected by rating changes, they first construct a “normal” return by a market model:

$$\text{normal stock return}_t = \alpha + \beta \times \text{return on stock index}_t,$$

where α and β are estimated on a normal time period (not including the rating change). The abnormal return is then calculated as the actual return minus the normal return. They then study how such abnormal returns behave, on average, around the dates of rating changes. Note that “time” is then measured, individually for each stock, as the distance from the day of rating change. The result is that there are significant negative abnormal returns following downgrades, but zero abnormal returns following upgrades.

2. They next turn to the question of whether bond ratings drive earnings forecasts or vice versa. To do that, they first note that there are some predictable patterns in revisions of earnings forecasts. They therefore fit a simple autoregressive model of earnings forecasts, and construct a measure of earnings forecast revisions (surprises) from the model. They then relate this surprise variable to the bond ratings. In short, the results are the following:
 - (a) both earnings forecasts and ratings react to the same information, but there is also a direct effect of rating changes, which differs between downgrades and upgrades.
 - (b) downgrades: the ratings have a strong negative direct effect on the earnings forecasts; the returns react even quicker than analysts
 - (c) upgrades: the ratings have a small positive direct effect on the earnings forecasts; there is no effect on the returns

A possible reason for why bond ratings could drive earnings forecasts and prices is that bond rating firms typically have access to more inside information about firms than stock analysts and investors.

A possible reason for the observed asymmetric response of returns to ratings is that firms are quite happy to release positive news, but perhaps more reluctant to release bad news. If so, then the information advantage of bond rating firms may be particularly large after bad news. A downgrading would then reveal more new information than an upgrade.

The different reactions of the earnings forecasts and the returns are hard to reconcile.

16.6.6 International Differences in Analyst Forecast Properties

Ang and Ciccone (2001) study earnings forecasts for many firms in 42 countries over the period 1988 to 1997. Some differences are found across countries: forecasters disagree more and the forecast errors are larger in countries with low GDP growth, less accounting disclosure, and less transparent family ownership structure.

However, the most robust finding is that forecasts for firms with losses are special: forecasters disagree more, are more uncertain, and are more overoptimistic about such firms.

16.6.7 Analysts and Industries

Boni and Womack (2006) study data on some 170,000 recommendations for a very large number of U.S. companies for the period 1996–2002. Focusing on revisions of recommendations, the papers shows that analysts are better at ranking firms within an industry than ranking industries.

16.6.8 Insiders

Corporate insiders *used to* earn superior returns, mostly driven by selling off stocks before negative returns. (There is little/no systematic evidence of insiders gaining by buying before high returns.) Actually, investors who followed the insider's registered transactions (in the U.S., these are made public six weeks after the reporting period), also used to earn some superior returns. It seems as if these patterns have more or less disappeared.

Further Reading

Elton, Gruber, Brown, and Goetzmann (2014) 17 and 27, Campbell, Lo, and MacKinlay (1997) 2 and 7, Cochrane (2005) 20.1, Campbell (2018) 5 and Pesaran (2015) 17 summarise a large number of studies of return predictability.

For a discussion of security analysts, see, among others, Makridakis, Wheelwright, and Hyndman (1998) 10.1 and Elton, Gruber, Brown, and Goetzmann (2014) 27.

Chapter 17

Event Studies

17.1 Basic Structure of Event Studies

The idea of an event study is to study the effect (on stock prices or returns) of a special type of event by using a cross-section of such events. For instance, what is the effect of a stock split announcement on the share price? Other events could be debt issues, mergers and acquisitions, earnings announcements, or monetary policy moves.

The event is typically assumed to be a binary variable. For instance, it could be a merger, or if the monetary policy surprise was positive (lower interest rates than expected). The basic approach is then to study what happens to the returns of those assets that have such an event.

Only news should move the asset price, so it is often necessary to explicitly model the previous expectations to define the event. For earnings, the event is typically taken to be the earnings announcement minus the average of analysts' forecasts. Similarly, for monetary policy moves, the event could be specified as the interest rate decision minus previous forward rates.

Similarly, we often study the *abnormal return* around such events, defined as (for asset i in period t)

$$u_{it} = R_{it} - R_{it}^{normal}, \quad (17.1)$$

where R_{it} is the actual return and the last term is the normal return, which may differ across assets and time. The definition of the normal return is discussed below. The idea of the “abnormal return” is that the event is likely to change the return relative to what it would otherwise have been on the same date, that is the normal return. However, event studies are also applied to other variables than returns: for instance, volatility, trade volume, bid-ask spreads, disagreement among forecasters, etc. Most of the methods discussed below apply

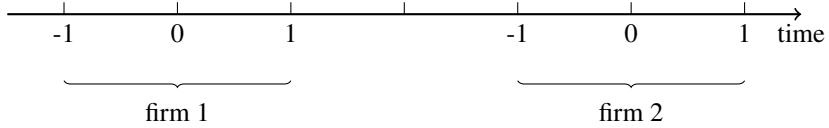


Figure 17.1: Event days and windows

in these cases as well, although the subsequent text refers specifically to “returns.”

Suppose we have a sample of n such events (“assets”). To maintain simple notation, we normalize the time so period 0 is the time of the event. Clearly the actual calendar time of the events for assets i and j are likely to differ, but we shift the time line for each asset individually so the time of the event is normalized to zero for every asset. See Figure 17.1 for an illustration.

To study information leakage and slow price adjustment, the abnormal return is often calculated for some time before and after the event: the *event window* (often ± 20 days or so). For day s (that is, s days after the event time 0), the cross sectional average abnormal return is

$$\bar{u}_s = \sum_{i=1}^n u_{is} / n. \quad (17.2)$$

For instance, \bar{u}_2 is the average abnormal return two days after the event, and \bar{u}_{-1} is for one day before the event.

The cumulative abnormal return (CAR) of asset i is simply the sum of the abnormal return in (17.1) over some period around the event. It is often calculated from the beginning of the event window (or from day 0), but this can be generalised. For instance, the cumulated return from s_1 to s_2 is

$$\text{car}_{i,[s_1,s_2]} = \sum_{\tau=s_1}^{s_2} u_{i\tau}. \quad (17.3)$$

The cross sectional average is clearly

$$\overline{\text{car}}_{i,[s_1,s_2]} = \sum_{i=1}^n \text{car}_{i,[s_1,s_2]} / n, \quad (17.4)$$

or, similarly, $\sum_{\tau=s_1}^{s_2} \bar{u}_{\tau}$ (as the order of summation does not matter).

Remark 17.1 (*Normalized cumulative abnormal returns**) *Normalized cumulative abnormal returns are sometimes reported. They are the cumulative returns from the start of*

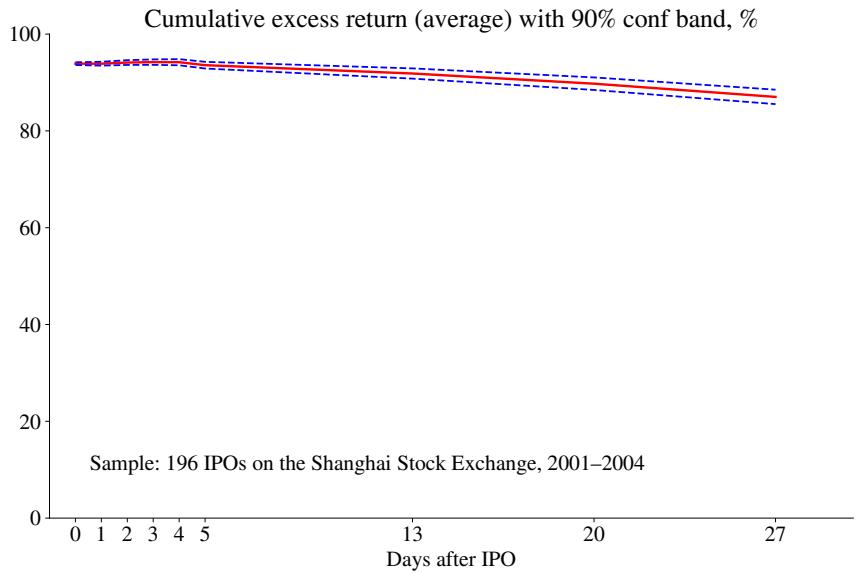


Figure 17.2: Event study of IPOs in Shanghai 2001–2004. (Data from Nou Lai.)

the event window until period s , minus the cumulative abnormal return up to and including the event day ($s = 0$).

Example 17.2 (Abnormal returns for ± 1 day around the event, two firms) Suppose there are two firms and the event window contains ± 1 day around the event day, and that the abnormal returns (in percent) are as follows

s	u_{1s}	u_{2s}	$car_{1,[-1,s]}$	$car_{2,[-1,s]}$	\bar{u}_s	$\overline{car}_{[-1,s]}$
-1	0.2	-0.1	0.2	-0.1	0.05	0.05
0	1.0	2.0	1.2	1.9	1.5	1.55
1	0.1	0.3	1.3	2.2	0.2	1.75

Empirical Example 17.3 (Cumulative abnormal returns) See Figure 17.2 for results on IPOs on the Shanghai stock exchange.

17.2 Models of Normal Returns

This section summarizes the most common ways of calculating the normal return in (17.1). The parameters in these models are typically estimated on a recent sample, the *estimation window*, which ends before the event window. See Figure 17.3 for an illustration. (When

there is no return data before the event window for instance, when the event is an IPO, then the estimation window can be after the event window.)

In this way, the estimated behaviour of the normal return should be unaffected by the event. It is almost always assumed that the event is exogenous in the sense that it is not due to the movements of the asset price during either the estimation window or the event window.

The *constant mean return model* assumes that the return of asset i fluctuates randomly around some mean μ_i

$$R_{it} = \mu_i + \varepsilon_{it}. \quad (17.5)$$

This mean is estimated by the sample average (during the estimation window). The normal return in (17.1) is then the estimated mean. $\hat{\mu}_i$. During the event window, we calculate the abnormal return as

$$u_{it} = R_{it} - \hat{\mu}_i. \quad (17.6)$$

The standard error is estimated by the standard error of $\hat{\varepsilon}_{it}$ in the estimation window. This means that we disregard the sampling uncertainty of the estimated mean, which makes sense if the estimation window is not very short (recall that the variance of a sample average is $\text{Var}(R_{it})/T$, where T is the number of data points in the estimation window). The same applies to the other models discussed below.

The *market model* is a linear regression of the return of asset i on the market return

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it}. \quad (17.7)$$

Notice that we typically do not impose the CAPM restrictions on the intercept in (17.7). The normal return in (17.1) is then calculated by combining the regression coefficients with the actual market return as $\hat{\alpha}_i + \hat{\beta}_i R_{mt}$. For the event window we calculate the abnormal return as

$$u_{it} = R_{it} - \hat{\alpha}_i - \hat{\beta}_i R_{mt}. \quad (17.8)$$

The standard error of this is estimated by the standard error of $\hat{\varepsilon}_{it}$ in the estimation window.

The market model has increasingly been replaced by a multi-factor model, for instance, the Fama and French (1993) model.

When we restrict $\alpha_i = 0$ and $\beta_i = 1$ in (17.7), then this approach is called the *market-adjusted-return model*. This is a particularly useful approach when there is no return data before the event, for instance, with an IPO. For the event window we calculate

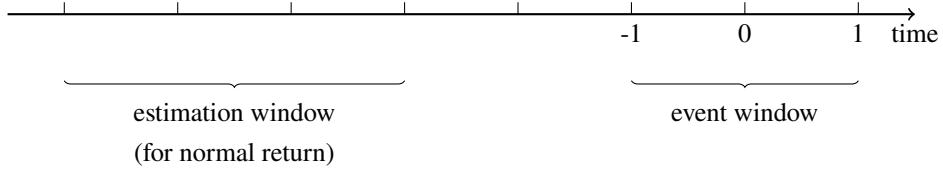


Figure 17.3: Event and estimation windows

the abnormal return as

$$u_{it} = R_{it} - R_{mt} \quad (17.9)$$

and the standard error of it is estimated by $\text{Std}(R_{it} - R_{mt})$ in the estimation window.

Finally, another approach is to construct the normal return as the actual return on assets which are similar to the asset with an event. For instance, if asset i is a small manufacturing firm with an event, then the normal return could be calculated as the average (in the cross-section, but same day) return for other small manufacturing firms without events, the *matching portfolio*. For the event window we calculate the abnormal return as

$$u_{it} = R_{it} - R_{pt}, \quad (17.10)$$

where R_{pt} is the return of the matching portfolio. The standard error of it is estimated by $\text{Std}(R_{it} - R_{pt})$ in the estimation window.

All the methods discussed here try to take into account the risk premium on the asset. It is captured by the mean in the constant mean model, the beta in the market model, and by the way the matching portfolio is constructed.

Apart from accounting for the risk premium, does the choice of the model of the normal return matter? Yes, but only if the model produces a higher coefficient of determination (R^2) than competing models. In that case, the variance of the abnormal return is smaller which makes the tests more precise. For instance, consider the market model (17.7). Under the null hypothesis that the event has no effect on the return, the abnormal return would be just the residual in the regression (17.7). It has the variance (assuming we know the model parameters)

$$\text{Var}(u_{it}) = \text{Var}(\varepsilon_{it}) = (1 - R^2) \text{Var}(R_{it}), \quad (17.11)$$

where R^2 is the coefficient of determination of the regression (17.7).

Proof (of (17.11)) Recall that R^2 of the regression (17.7) is defined as

$$R^2 = 1 - \text{Var}(\varepsilon_{it}) / \text{Var}(R_{it}).$$

Rearrange to get (17.11). \square

This variance is crucial for testing the hypothesis of no abnormal returns: the smaller is the variance, the greater the power to reject a false null hypothesis. The constant mean model has $R^2 = 0$, so the market model could potentially give a much smaller variance. If the market model has $R^2 = 0.75$, then the standard deviation of the abnormal return is only half that of the constant mean model (since $\sqrt{1 - 0.75} = 0.5$). Experience with multi-factor models also suggest that they give relatively small improvements of the R^2 compared to the market model. For these reasons, and for reasons of convenience, the market model is still the dominating model of normal returns.

High frequency data can be very helpful, provided the time of the event is known. High frequency data effectively allows us to decrease the volatility of the abnormal return since it filters out irrelevant (for the event study) shocks to the return while still capturing the effect of the event.

17.3 Testing the Abnormal Return

In testing if the abnormal return is different from zero, there are two sources of sampling uncertainty. *First*, the parameters of the normal return model are uncertain. *Second*, even if we knew the normal return for sure, the returns are random variables, and they will always deviate from their population means in any finite sample. The first source of uncertainty is likely to be much smaller than the second, provided the estimation window is much longer than the event window. This is the typical situation, so the rest of the discussion will focus on the second source of uncertainty.

It is typically assumed that the abnormal returns are uncorrelated both across time and across assets. The first assumption is motivated by the very low autocorrelation of returns. The second assumption makes a lot of sense if the events are not overlapping in time, so that the event of assets i and j happen at different (calendar) times. In contrast, if the events happen at the same time, the cross-correlation must be handled somehow (see below). This is, for instance, the case if the events are macroeconomic announcements or monetary policy moves. For the rest of this section we assume no autocorrelation or cross correlation.

Let $\sigma_i^2 = \text{Var}(u_{it})$ be the variance of the abnormal return of asset i . The *variance of the cross-sectional average* (across the n assets) *average*, \bar{u}_s in (17.2), is then

$$\text{Var}(\bar{u}_s) = \sum_{i=1}^n \sigma_i^2 / n^2, \quad (17.12)$$

since all covariances are assumed to be zero. In a large sample, we can therefore use a t -test

$$\bar{u}_s / \text{Std}(\bar{u}_s) \xrightarrow{d} N(0, 1). \quad (17.13)$$

In most applications the σ_i^2 values used in (17.12) are from the estimation window, as discussed in Section 17.2.

Remark 17.4 (*An alternative way to test \bar{u}_s , using the cross-sectional standard deviation of u_{is}) An alternative and intuitive approach to calculate $\text{Var}(\bar{u}_s)$ is to use σ_{CR}^2 / n , where σ_{CR}^2 is the cross-sectional variance of u_{is} . In fact, it can be shown that the cross-sectional variance σ_{CR}^2 is an unbiased estimate of $\sum_{i=1}^n \sigma_i^2 / n$. This suggests that using (17.12) and σ_{CR}^2 / n are conceptually the same. (The extension to cumulative returns period is straightforward.) However, σ_{CR}^2 suffers from the drawback that it is an estimate based on only n data points, so it is likely to be a noisy estimate when the number of events (n) is small. This supports the use of (17.12) and a large estimation window for each of the assets.

The *cumulative abnormal return* can also be tested with a t -test. Since the returns are assumed to have no autocorrelation the variance of car is

$$\text{Var}(\text{car}_{i,[s_1,s_2]}) = (s_2 - s_1 + 1)\sigma_i^2. \quad (17.14)$$

This variance is increasing in the number of return periods, since we are considering cumulative returns (not the time average of returns). Also, the variance of the *cross-sectional average car* is

$$\text{Var}(\overline{\text{car}}_{[s_1,s_2]}) = (s_2 - s_1 + 1) \text{Var}(\bar{u}_s), \quad (17.15)$$

where $\text{Var}(\bar{u}_s)$ is defined in (17.12).

Empirical Example 17.5 (Testing CAR) See Figure 17.2 for results on IPOs.

Example 17.6 (Variances of abnormal returns) If the standard deviations of the daily abnormal returns of the two firms in Example 17.2 are $\sigma_1 = 0.1$ and $\sigma_2 = 0.2$, then

we have the following variances

s	u_{1s}	u_{2s}	$car_{1,[-1,s]}$	$car_{2,[-1,s]}$	\bar{u}_s	$\overline{car}_{[-1,s]}$
s	0.1^2	0.2^2	$(s+2)0.1^2$	$(s+2)0.2^2$	$\frac{0.1^2+0.2^2}{4}$	$(s+2)\frac{0.1^2+0.2^2}{4}$

Example 17.7 (*Tests of abnormal returns*) By dividing the numbers in Example 17.2 by the square root of the numbers in Example 17.6 (that is, the standard deviations for $s = -1, 0, 1$) we get the t -statistic for the abnormal returns

s	u_{1s}	u_{2s}	$car_{1,[-1,s]}$	$car_{2,[-1,s]}$	\bar{u}_s	$\overline{car}_{[-1,s]}$
-1	2	-0.5	2	-0.5	0.4	0.4
0	10	10	8.5	6.7	13.4	9.8
1	1	1.5	7.5	6.4	1.8	9.0

17.3.1 When Events are Clustered

When events (for different assets or firms) occur on the same days (are clustered), then we may have to consider the possibility that the abnormal returns are correlated, especially when we test the cross-sectional average abnormal return.

The easiest approach to handle clustered events is demonstrate that they are still unlikely to be correlated. For instance, if we use the market model for normal returns, then the abnormal returns for firms in different industries might be uncorrelated. The second easiest approach is to create a portfolio which includes all assets (firms) with an event—and use that as the (only) test asset in an event study. Alternatively, we keep the cross-sectional data on the abnormal returns, but explicitly handle the cross-sectional correlations. To illustrate that, suppose there are only two assets, so the cross-sectional average is

$$\bar{u}_s = (u_{1s} + u_{2s})/2, \quad (17.16)$$

where u_{is} is the abnormal return on asset i on day s . In general, the variance of this average is

$$\text{Var}(\bar{u}_s) = (\sigma_1^2 + \sigma_2^2 + 2\sigma_{12})/4, \quad (17.17)$$

where σ_{12} is the covariance of u_{1s} and u_{2s} . Previously, we assumed that $\sigma_{12} = 0$ since the abnormal returns of the two assets referred to different calendar days. Here, we instead have to estimate σ_{12} in the estimation window and include it in the calculations. Notice

that we can write this in matrix form as

$$\text{Var}(\bar{u}_s) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}' \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} / 4, \quad (17.18)$$

where the matrix in the middle is the variance-covariance matrix of the two returns. (It is symmetric so $\sigma_{12} = \sigma_{21}$.) With n assets, the formula becomes

$$\text{Var}(\bar{u}_s) = \mathbf{1}' \Sigma \mathbf{1} / n^2, \quad (17.19)$$

where Σ is the variance-covariance matrix of all n assets and $\mathbf{1}$ is a column (n rows) vectors filled with ones. Clearly, this is the same as summing all the elements of the variance-covariance matrix and then dividing by n^2 . We then replace $\text{Var}(\bar{u}_s)$ in (17.12) and (17.15) with (17.19). The rest of the analysis is unchanged.

17.4 Quantitative Events

Some events are not easily classified as binary variables. For instance, the effect of positive earnings surprise is likely to depend on how large the surprise is, not just if there was a positive surprise. This can be studied by regressing the abnormal return (typically the cumulative abnormal return) on the value of the event (x_i)

$$\text{car}_{i,[s_1,s_2]} = a + bx_i + \zeta_i. \quad (17.20)$$

The slope coefficient is then a measure of how much the cumulative abnormal return reacts to a change of one unit of x_i .

Further Reading

See also Elton, Gruber, Brown, and Goetzmann (2014) 17 and Campbell, Lo, and MacKinlay (1997) 4.

Chapter 18

Distributions and Option Pricing

18.1 Estimating and Testing Distributions

This chapter presents methods for studying distributions. This can be applied to data, residuals, or some other transformation, depending on the aim.

18.1.1 Histograms and Averaged Shifted Histograms

Histograms, such as the one in Figure 18.1, are very useful for gauging the shape of the distribution. The histogram either shows the number of occurrences or a normalised version where the area under the histogram integrates to one (similar to a pdf).

Empirical Example 18.1 *Figure 18.1 shows a histogram of daily equity returns. Zooming in suggests the presence of fat tails and possibly some skewness.*

The formal definition of a normalised histogram is: at any value x in the bin $c \pm h/2$ (for instance, for $x = 0.4$ which is in the bin 0 ± 0.5) the value is

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^T \delta(c - h/2 < x_t \leq c + h/2) \text{ for } x \in (c - h/2, c + h/2], \quad (18.1)$$

where $\delta(q) = 1$ if q is true, and zero otherwise. The sum clearly counts the number of data points that are inside the bin $c \pm h/2$. This is a normalised histogram, since we are dividing by the bin width h . Multiply by h to get a more traditional histogram which simply counts the relative frequencies of the different bins. In any case, $\hat{f}(x)$ is a step function since it is the same for all x values within the same bin. (Also, in principle, bin widths could vary across bins.)

To calculate the histogram, we need to make two choices: the width of the bins and their location. These choices are often made to facilitate interpretation of the results (for

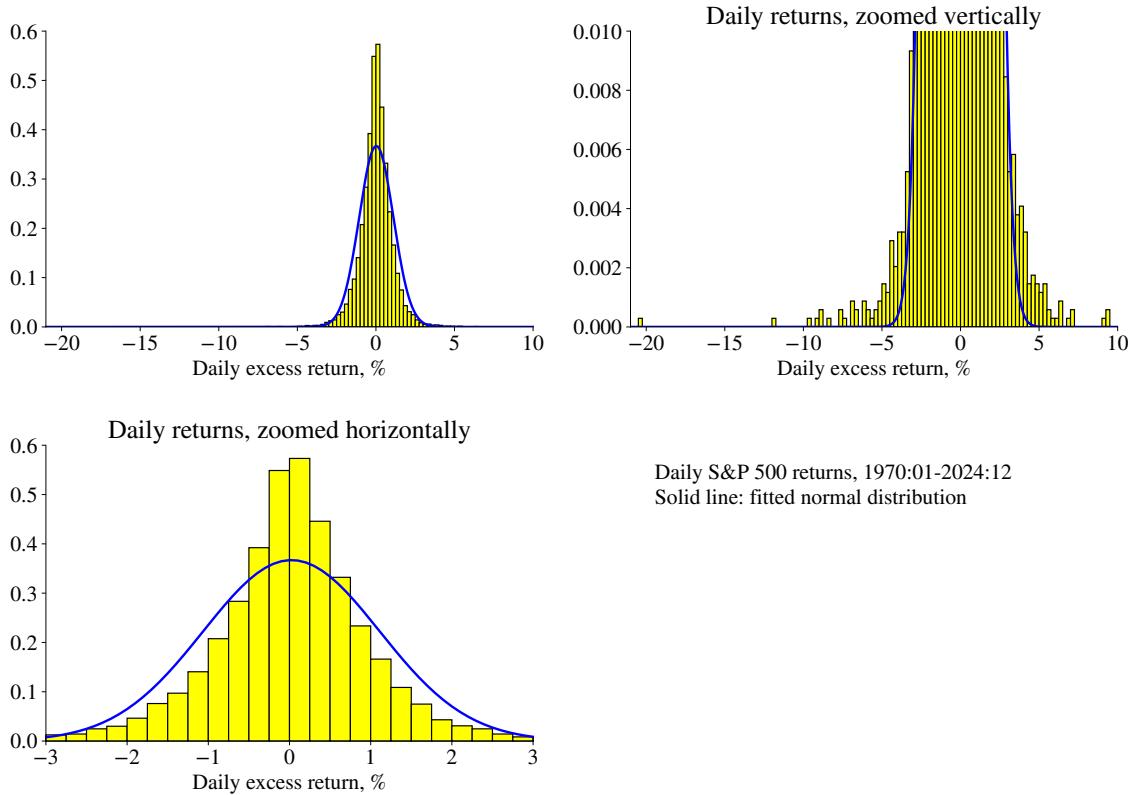


Figure 18.1: Distribution of daily S&P returns

instance, bins for returns might be $(0\%, 1\%]$, $(1\%, 2\%]$, etc), but the properties of the data set should also be considered: there are useful recommendations for how to set the bin widths as functions of the volatility of the data and the sample length (see Remark).

Remark 18.2 (*Histogram bin width**) *It is often recommended to use histogram bins that are $h = 3.5\sigma / T^{1/3}$ wide (alternatively, $h = 2 \text{ IQR} / T^{1/3}$ where IQR is the interquartile range, $\text{quantile}(0.75) - \text{quantile}(0.25)$). Clearly, if c is the bin midpoint, then the bin is $c \pm h/2$.*

However, the choice of the location (equivalently, the bin midpoints) can also be tricky. The upper panel of Figure 18.2 illustrates that.

The *averaged shifted histogram* (ASH, Scott (1985)) is a way to reduce the importance of the location choice. The idea is to calculate several histograms using different bin locations (c), and then average across them (at the same x value). Figure 18.2 shows the average of two sets of histograms: the original set and one where the bins have been shifted

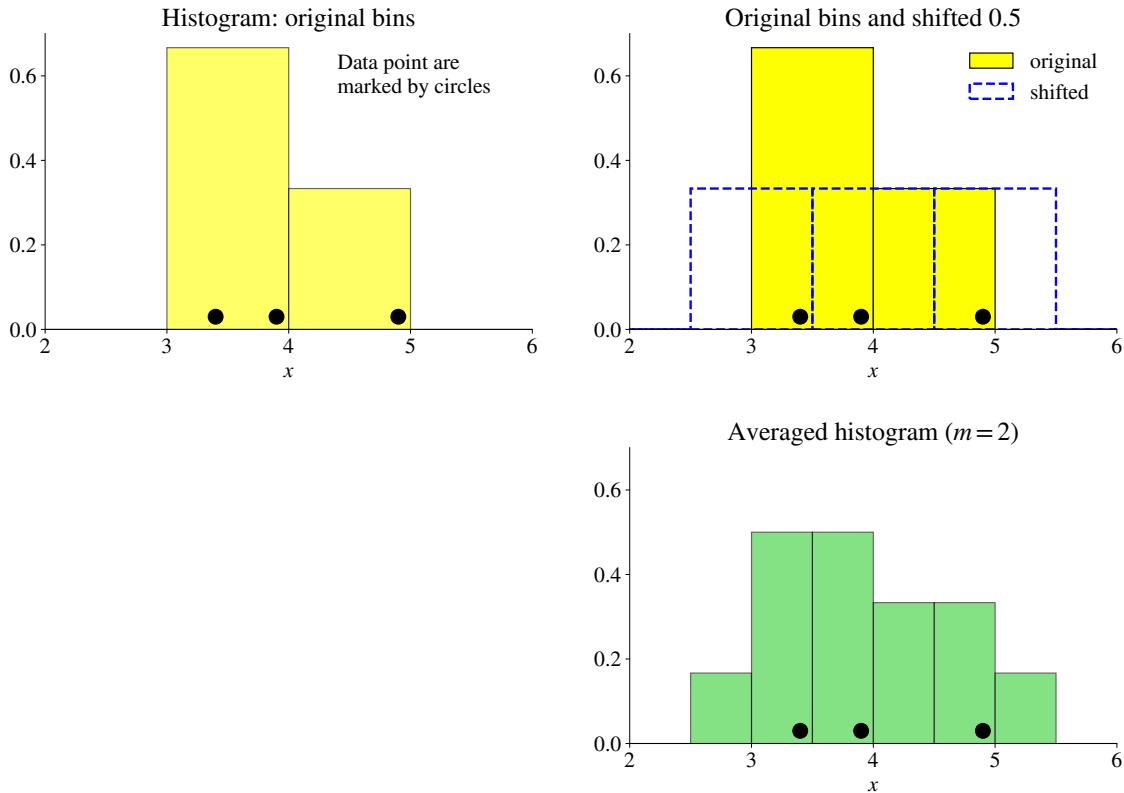


Figure 18.2: Example of histograms on a very small sample

up ($h/2$). Note that we also add one more bin on the left (not all statistical packages do that).

Empirical Example 18.3 (*Averaged shifted histogram*) See Figure 18.10.

With $m - 1$ shifts, the formal definition of the ASH at any value x is

$$\hat{f}_{ash}(x) = \sum_{j=1}^m \hat{f}_j(x)/m \quad (18.2)$$

where $\hat{f}_j(x)$ is the value (at x) of the histogram from (18.1) where each bin center c is shifted up $(j - 1)h/m$. This approach excludes part of the original histogram's range (to the left of the lowest of the new bins). For that reason, we may add an extra bin, h below the new set of bins. Also, notice that for $j = 1$, we recover the original (non-shifted) histogram.

Example 18.4 (*How the bins are shifted**) For $j = 0$, we have the original histogram since $(j - 1)h/m = 0$. For $j \geq 1$, we add one bin to the left and shift the other bins to

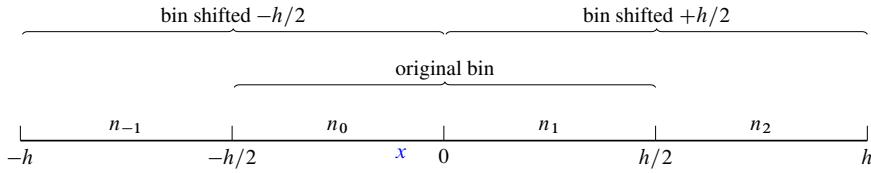


Figure 18.3: Calculation of averaged shifted histogram

the right. For instance, for $m = 2$ and $m = 3$ we have the following shifts

j	$\underline{m = 2}$	$\underline{m = 3}$
2	$(j-1)h/m$	$(j-1)h/m$
3	$h/2$	$h/3$

Remark 18.5 (*Calculating the ASH**) The ASH can be calculated from the definition (18.1)–(18.2). Alternatively, the following equivalent approach might be quicker. Figure 18.3 shows the original bin $0 \pm h/2$ and also bins that are shifted $h/2$. This implicitly defines a set of smaller bins $(-h, -h/2]$, $(-h/2, 0]$ etc. Let n_i indicate the number of data points in each of smaller bins. Notice that $f_1(x) = (n_0 + n_1)/(hT)$ and $f_2(x) = (n_{-1} + n_0)/(hT)$, so the average is $(n_0 + n_1 + n_{-1} + n_0)/(2hT)$. It can be noticed that the more general expression with $m - 1$ shifts (the figure shows $m = 2$) is

$$\hat{f}_{ash}(x) = \sum_{i=1-m}^{m-1} (1 - |i|/m) n_i / (Th).$$

This shows strong similarities with a kernel density estimator (see below).

18.1.2 QQ Plots

A QQ plot is a way to assess if the empirical distribution conforms reasonably well to a prespecified theoretical distribution, for instance, a normal distribution where the mean and variance have been estimated from the data.

In essence, a QQ plot is a scatter plot where each point shows a specific percentile (quantile) according to the empirical distribution as well as the theoretical distribution. For instance, if the 2nd percentile (0.02 quantile) is at -10 in the empirical distribution, but at only -3 in the theoretical distribution, then this indicates that the two distributions have very different left tails.

Empirical Example 18.6 (*QQ plots of equity and FX returns*) See Figures 18.4 –18.5 for illustrations. The evidence suggests that daily equity returns have strongly non-normal distributions with fat tails and some skewness, while monthly returns look more normally distributed, perhaps with some skewness left.

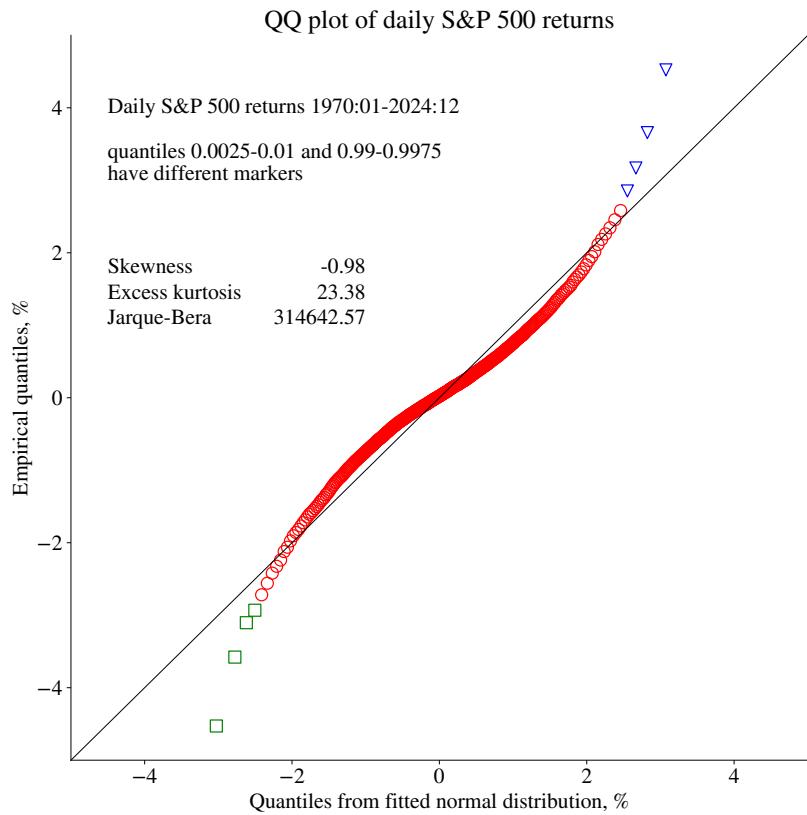


Figure 18.4: Quantiles of daily S&P returns

18.1.3 Parametric Tests of a Normal Distribution

The skewness, kurtosis, and Jarque-Bera tests for normality are useful diagnostic tools. First, let $z_t = (x_t - \mu)/\sigma$ and then note

	Test statistic	Distribution	
skewness	$\sum_{t=1}^T z_t^3 / T$	$N(0, 6/T)$	(18.3)
kurtosis	$\sum_{t=1}^T z_t^4 / T$	$N(3, 24/T)$	
Jarque-Bera	$(T/6)\text{skewness}^2 + (T/24)(\text{kurtosis} - 3)^2$	χ_2^2	

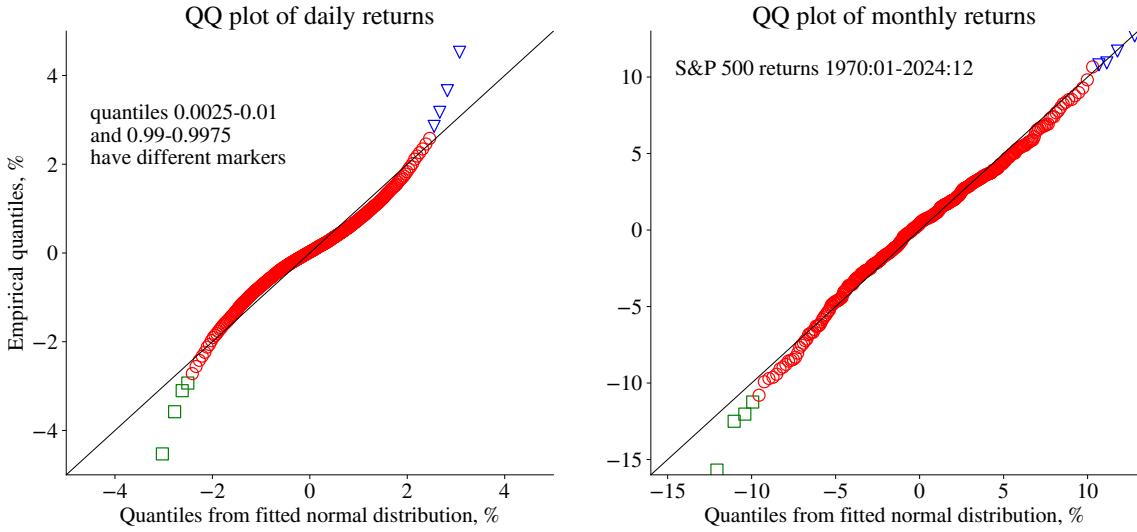


Figure 18.5: Distribution of S&P returns (different horizons)

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (18.3) are under the null hypothesis that x_t is iid $N(\mu, \sigma^2)$. The “excess kurtosis” is defined as the kurtosis minus 3. See Figure 18.4 for an example.

The intuition for the χ^2_2 distribution of the Jarque-Bera test is that both the skewness and kurtosis are, if properly scaled, $N(0, 1)$ variables. It can also be shown that they, under the null hypothesis, are uncorrelated. The Jarque-Bera test statistic is therefore a sum of the square of two uncorrelated $N(0, 1)$ variables, which has a χ^2_2 distribution.

18.1.4 Nonparametric Tests of General Distributions

The *Kolmogorov-Smirnov* test is designed to test if an empirical distribution function, $EDF(x)$, conforms with a theoretical cumulative distribution function, $F(x)$. The empirical distribution function is defined as the fraction of observations which are less than or equal to x , that is,

$$EDF(x) = \sum_{t=1}^T \delta(x_t \leq x) / T, \quad (18.4)$$

where $\delta(q) = 1$ if q is true, and zero otherwise. The $EDF(x_t)$ and $F(x_t)$ are often plotted against the sorted sample of x_t values. See Figure 18.6 for an illustration.

A test of whether the empirical distribution differs from the theoretical can be done in

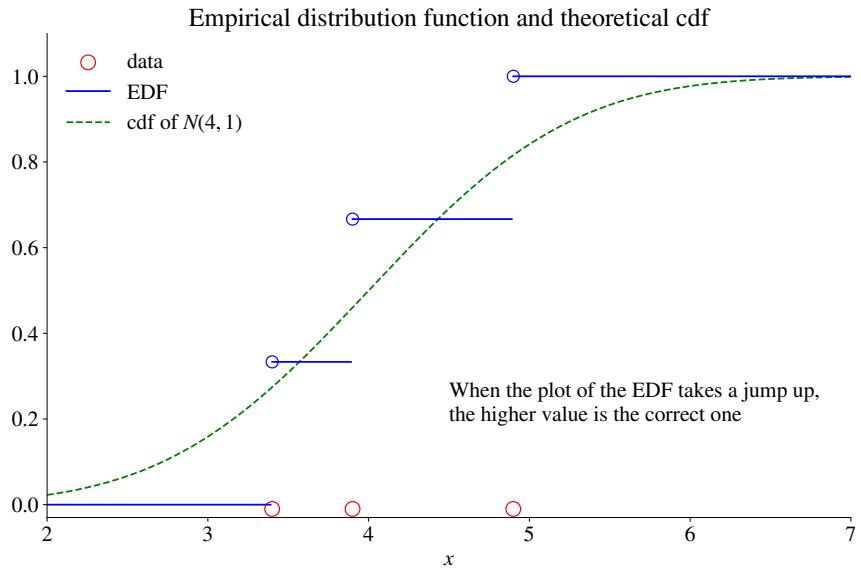


Figure 18.6: Example of empirical distribution function

terms of the absolute value of the maximum distance

$$D = \max_x |\text{EDF}(x) - F(x)|. \quad (18.5)$$

This maximum is across all possible x values (not just the observed data), but there is a simple way to identify the maximum (see the Remark below). See Figure 18.7 for an illustration.

Remark 18.7 (*Calculating the D statistic**) *The maximum in (18.5) must be at or just before a data point (where the EDF jumps up), since the theoretical distribution is weakly increasing. Let (z_1, \dots, z_T) be the sorted sample of data. First, at any data point, calculate $|\text{EDF}(z_i) - F(z_i)|$ as $|i/T - F(z_i)|$. Second, just before a data point, the theoretical cdf is (almost) $F(z_i)$ and the EDF is $(i-1)/T$, so calculate $|(i-1)/T - F(z_i)|$. Find the largest of the two numbers and then across the sample.*

We reject the null hypothesis that the two distributions are the same if $\sqrt{T}D > c$, where c is a critical value, (1.36, 1.48, 1.63) on the (5%, 2.5%, 1%) significance levels, see Mittelhammer (1996) 10. Instead, if the $F(x_t)$ distribution is $N(\mu, \sigma^2)$ with estimated parameters, then the critical values should be changed to (0.80, 0.89, 1.03).

Remark 18.8 (*K-S test for comparing two EDFs**) *Let $\text{EDF}_1(x)$ be the edf for variable 1 (from a sample with T_1 observations) and $\text{EDF}_2(x)$ for variable 2 (from a sample with T_2*

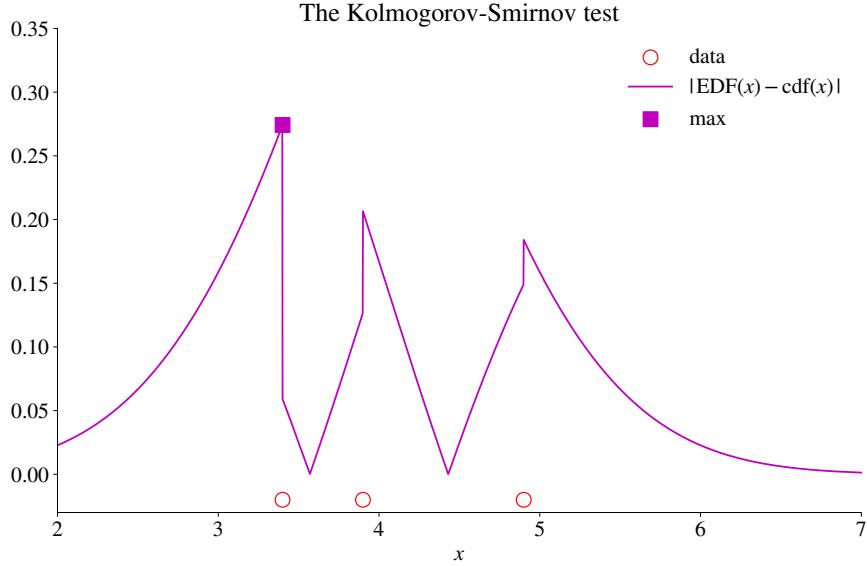


Figure 18.7: K-S test

observations). Define $D = \max_x |EDF_1(x) - EDF_2(x)|$ and reject the null hypothesis that they are from the same distribution if $\sqrt{T_1 T_2 / (T_1 + T_2)} D > c$ where c is the same critical value as before. To find the maximum, try all $x = \text{data}_i$ values from the union of the two samples.

Pearson's χ^2 test does the same thing as the K-S test but for a discrete distribution. Suppose you have K categories with N_i values in category i . The theoretical distribution predicts that the fraction p_i should be in category i , with $\sum_{i=1}^K p_i = 1$. Then

$$\sum_{i=1}^K (N_i - Tp_i)^2 / (Tp_i) \sim \chi^2_{K-1}. \quad (18.6)$$

There is a corresponding test for comparing two empirical distributions.

18.1.5 Kernel Density Estimation

Reference: Silverman (1986)

The normalised histogram at the point x (say, $x = 2.3$) can actually be defined as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{h} \delta \left(\left| \frac{x_t - x}{h} \right| \leq 1/2 \right), \quad (18.7)$$

where $\delta(q) = 1$ if q is true, and zero otherwise. In this case, the intervals ("bins") are h wide around a point x : $x - h/2$ to $x + h/2$.

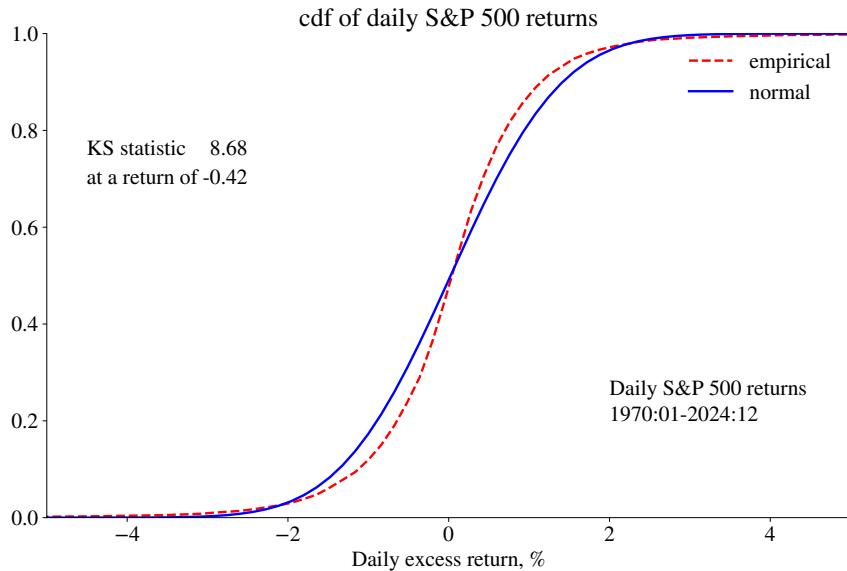


Figure 18.8: K-S test

In fact, that $\delta(|x_t - x| \leq h/2)/h$ is the pdf value of a uniformly distributed variable (over the interval $x - h/2$ to $x + h/2$). This shows that our estimate of the pdf (18.7) can be thought of as an average of hypothetical pdf values of the data close to x .

However, we can gain efficiency and get a smoother (across x values) estimate by using a density function other than the uniform. In particular, using a density function that tapers off continuously instead of suddenly dropping to zero (as the uniform density does) improves the properties. In fact, a normal pdf is often used. The kernel density estimator of the pdf at some point x is then

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{h} \phi\left(\frac{x_t - x}{h}\right), \quad (18.8)$$

where $\phi(z)$ is the pdf of a $N(0, 1)$ variables, so the summand is the pdf of a $N(x, h^2)$ variable evaluated at x_t . See Figure 18.9 for an example of the weights in the calculation.

The value $h = 1.06 \text{Std}(x_t) T^{-1/5}$ is sometimes recommended, since it can be shown to be the optimal choice (in MSE sense) if data is normally distributed and the gaussian kernel is used. A more robust choice is to replace $\text{Std}(x_t)$ in the formula by $\min(\text{Std}(x_t), \text{IQR}/1.34)$, where IQR is the inter-quartile range.

Empirical Example 18.9 (*Kernel density estimate of equity returns*) See Figure 18.10.

It can be shown that (with iid data and a Gaussian kernel) the asymptotic distribution

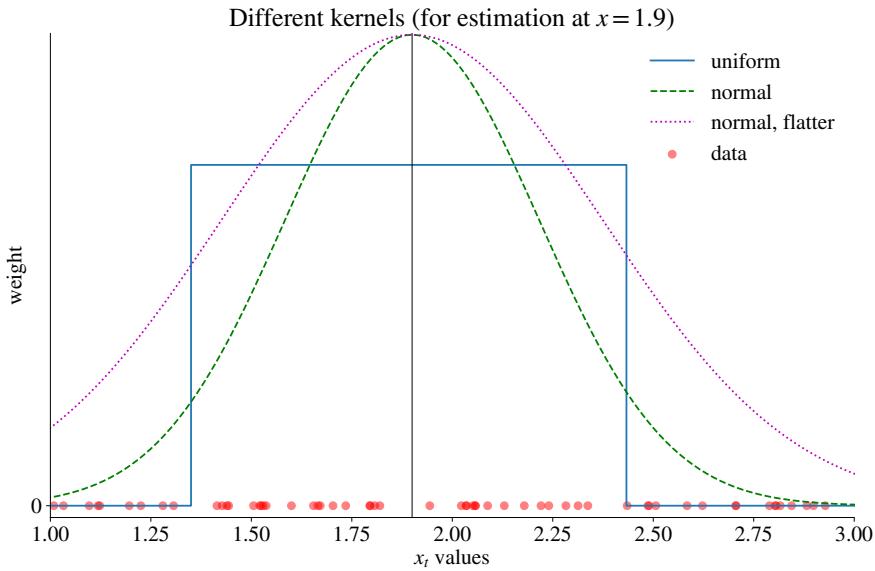


Figure 18.9: Different weighting functions for kernel density estimate

of the estimate is

$$\sqrt{Th}[\hat{f}(x) - f(x)] \rightarrow^d N \left[0, f(x)/(2\sqrt{\pi}) \right]. \quad (18.9)$$

We can also estimate multivariate pdfs. Let x_t be a $d \times 1$ vector. We can then estimate the pdf at a point x by using a multivariate Gaussian kernel as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{(2\pi)^{d/2} |H|^{1/2}} \exp \left[-\frac{1}{2} (x_t - x)' H^{-1} (x_t - x) \right], \quad (18.10)$$

where H is a bandwidth matrix. There are several common choices of H : (a) $h^2 I$, (b) $\text{diag}(h_1^2, \dots, h_d^2)$ which creates a diagonal matrix with zeros for off-diagonal elements, and (c) $h^2 \hat{\Omega}$ where $\hat{\Omega}$ is the estimated covariance matrix of x_t . Let $\lambda = \{4/[T(d+2)]\}^{1/(d+4)}$. For option (c), a common recommendation is $h = \lambda$; for option (b) $h_i = \lambda \text{ Std}(x_{it})$; and for option (a) $h_i = \lambda$ times the average standard deviation.

In the bivariate case, it might be useful to plot the contours of the estimated pdf rather than all scatter points, especially for very large data set.

Remark 18.10 (H^*) *With just one variable ($d = 1$), $H = h^2 \hat{\Omega}$, and the recommended bandwidth, $H^{1/2} \approx 1.06 \text{ Std}(x_t) T^{-1/5}$, which is the same as discussed for the univariate case (18.8). With two variables ($d = 2$), $\lambda = T^{-1/6}$.*

Empirical Example 18.11 (*Kernel estimate of a bivariate pdf*) See Figure 18.11 for a

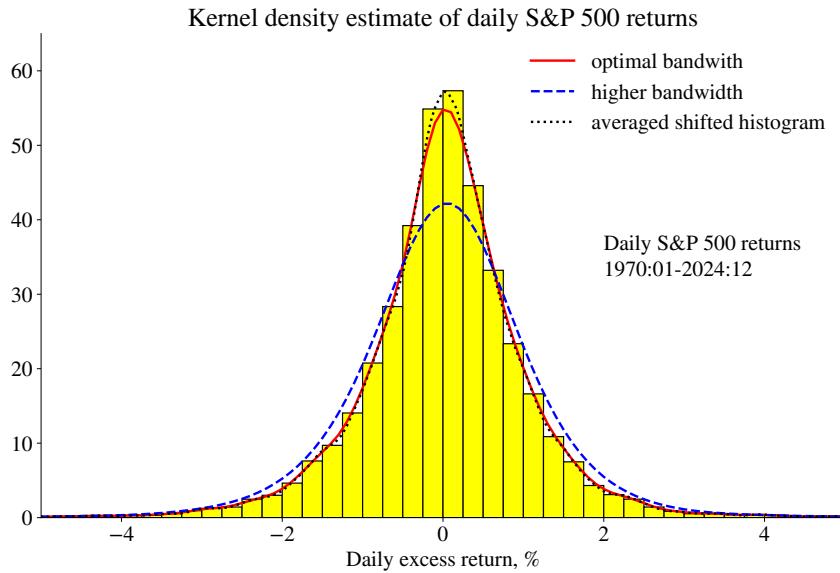


Figure 18.10: Distribution of stock returns

scatter plot and the contours from a an estimated bivariate density function. The negative relation between carry trade returns and market volatility is considerable.

18.2 Option Pricing

18.2.1 The Black-Scholes Option Price Model

This section discusses how to estimate the inputs to the Black-Scholes option pricing formula and also how to evaluate it.

A European call option contract traded today stipulates that the buyer of the contract has the right (not the obligation) to buy one unit of the underlying asset (from the issuer of the option) in m periods at the strike price K .

The standard option pricing formula is the Black-Scholes (B-S) model (see ? and ?). The basic model assumption is that the log price of the underlying asset is normally distributed. The formula for a call option price (C_t) on an underlying asset without dividends (until expiration of the option) is

$$C_t = S_t \Phi(d_1) - e^{-y_t m} K \Phi(d_2), \text{ where} \quad (18.11)$$

$$d_1 = \frac{\ln(S_t/K) + (y_t + \sigma^2/2)m}{\sigma \sqrt{m}} \text{ and } d_2 = d_1 - \sigma \sqrt{m},$$

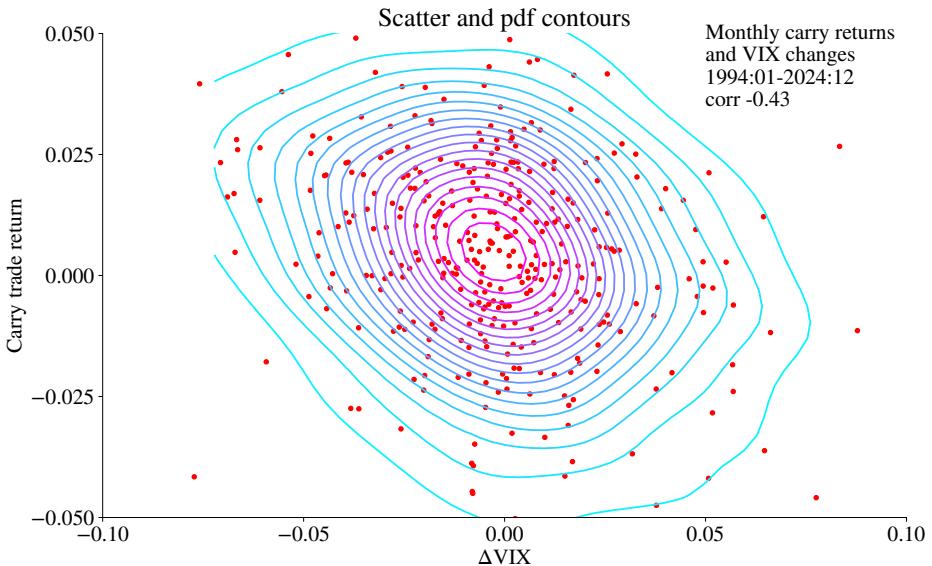


Figure 18.11: Scatter and contours of estimated bivariate pdf

where S_t is the price of the underlying asset today, y_t is the continuously compounded interest rate, and σ^2 is the perceived annualised variance of the return of the underlying asset between now (t) and expiration of the option (in $t + m$). Also, $\Phi()$ is the cumulative distribution function of a standard normal, $N(0, 1)$, variable.

It can be shown that the call option price is increasing in the volatility and decreasing in the strike price. (The formula for a put option, which gives the right to sell the underlying for K , is $P = e^{-ym} K \Phi(-d_2) - S \Phi(-d_1)$.)

The B-S formula can be derived from several stochastic processes of the underlying asset price, but they all imply that the distribution of the change in log asset price between t and $t + m$ is normal with some mean α (not important for the option price) and the variance $m\sigma^2$

$$\ln S_{t+m} - \ln S_t \sim N(\alpha, m\sigma^2). \quad (18.12)$$

18.2.2 Estimation of the Parameters in Black-Scholes*

The only *unknown parameter* in (18.11) is the annualized variance of the return of the underlying asset, σ^2 . This should reflect the beliefs about the uncertainty between the day the option is traded and its expiration.

One way of formulating such a belief is to estimate the volatility on a historical sample. If so, two considerations are important. *First*, the Black-Scholes model is not consistent

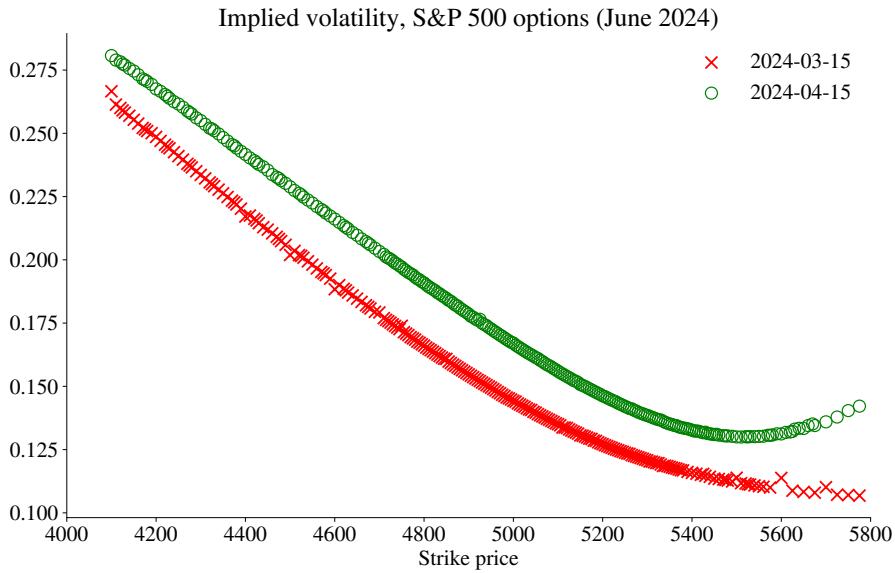


Figure 18.12: Implied volatilities of S&P 500 options, selected dates

with stochastic volatility: it is derived under the assumption that the variance is known. For that reason, a GARCH model would require another option pricing model. Anyhow, Black-Scholes is sometimes still used as an approximation in such cases. *Second*, there is a choice of data frequency in estimating the variance. This is discussed below.

Assume that we have return data for observations in $\tau = 1, 2, \dots, n$. This could be 5-minute intervals, days, weeks or something else. Let the time between τ and $\tau + 1$ be v years. For instance, with daily data $v = 1/252$ (if only trading days are counted), but with weekly data $v = 1/52$. Calculate the traditional sample variance, here denoted \hat{s}^2 , and then annualise as

$$\hat{\sigma}^2 = \hat{s}^2/v. \quad (18.13)$$

If the returns were iid normally distributed, then traditional results give $\text{Var}(\hat{s}^2) = 2s^4/n$, which implies that $\text{Var}(\hat{\sigma}^2) = 2\sigma^4/n$. This suggests using as *many* observations as possible (high frequency data). However, this result relies heavily on the iid assumption. In particular, high-frequency returns are often characterised by heavy tails, autocorrelation (bid-ask bounce, among other things) and heteroskedasticity, which need to be filtered out before estimation. Unless that is done, daily returns might be a safer bet.

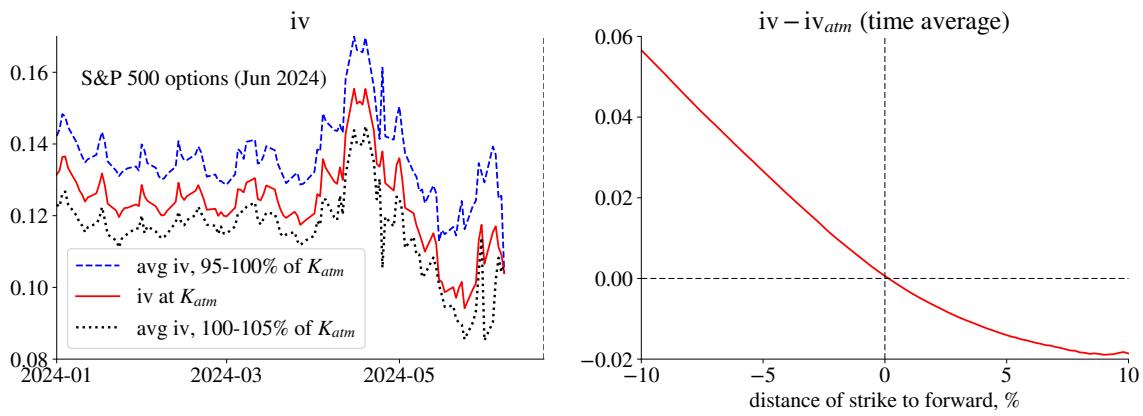


Figure 18.13: Implied volatilities

18.2.3 Testing Black-Scholes: Implied Volatility

The pricing formula (18.11) contains only one unknown parameter: the standard deviation σ in the distribution of $\ln S_{t+m}$, see (18.12). With data on the option price, spot price, the interest rate, and the strike price, we can solve for standard deviation: the *implied volatility*. This can be used to *test* the model.

We can solve for one implied volatility for each available strike price. If the Black-Scholes formula is correct, that is, if the assumption in (18.12) is correct, then these volatilities should be *the same across strike prices and time to expiration*, and it should also be *constant over time*. If not, data on option prices are driven by another model.

Empirical Example 18.12 (*Implied volatility for S&P 500 options*) See Figure 18.12 for volatility smiles for a few trading days, and Figure 18.13 for a summary of the dynamics over time (S&P 500 options).

In contrast, it is often found that the implied volatility is a “smirk” (equity markets) or “smile” (FX markets) shaped function of the strike price. One possible explanation for a smirk shape is that market participants assign a higher probability to a dramatic drop in share prices than a normal distribution suggests. A possible explanation for a smile shape is that the (perceived) distribution of the future asset price has more probability mass in the tails (“fat tails”) than a normal distribution has. In addition, the implied volatilities seems to vary considerably over time. Together, these findings question some of the key assumptions in the Black-Scholes model. For instance, Heston and Nandi (2000) and Duan (1995) have suggested models to accommodate this.

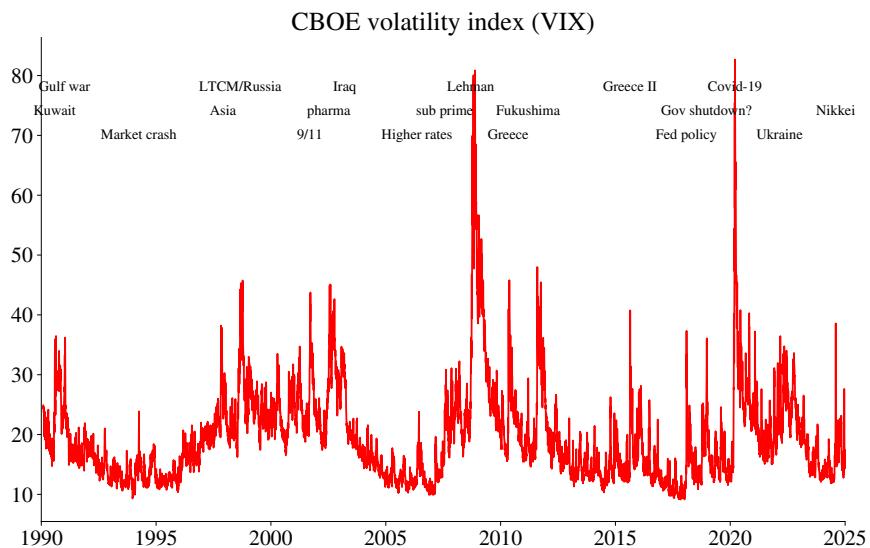


Figure 18.14: CBOE VIX, summary measure of implied volatilities (30 days) on US stock markets

Empirical Example 18.13 (VIX) See Figure 18.14 for an illustration of how VIX (an index of 1-month S&P 500 implied volatility) has changed over time.

Further Reading

For the estimation and test of distributions, see Harvey (1989), Davidson and MacKinnon (1993), Silverman (1986), Mittelhammer (1996), DeGroot (1986) and Hansen (2022b) 17.

For option pricing, see Hull (2022) 20, Taylor (2005) 13–14, Campbell, Lo, and MacKinlay (1997) 9, and Gourieroux and Jasiak (2001) 12–13.

Chapter 19

Maximum Likelihood Estimation

19.1 Maximum Likelihood

There are many cases when OLS and IV are *not* well suited for the estimation problem. An earlier chapter discussed how GMM might help. This chapter, in contrast, will focus on Maximum likelihood estimation (MLE), which might also be useful—and it has a long tradition in econometrics.

The key steps in MLE are to (1) specify a “likelihood function”; (2) maximise it by choosing the parameter values.

The intuition of MLE is to choose model parameters to *make the data most likely* under the model. For instance, to estimate the mean and variance of a normally distributed random variable, MLE picks parameters to make the pdf values of the data as high as possible: see Figure 19.1 (left subfigure) for an example, where one of the parameter choices assigns higher pdf values for most data points.

MLE has good properties, provided the basic distributional assumptions are correct; that is, if we maximize the correct likelihood function. In that case, MLE gives the most efficient/precise estimators, at least in large samples. The standard errors for MLE are based on asymptotic results related to the Delta method.

19.1.1 Basic Approach

For a random variable y_t with a continuous distribution, we specify the *likelihood* for the sample (y_1, y_2, \dots, y_T) as the joint probability density function

$$L(\beta) = \text{pdf}(y_1, \dots, y_T; \beta), \quad (19.1)$$

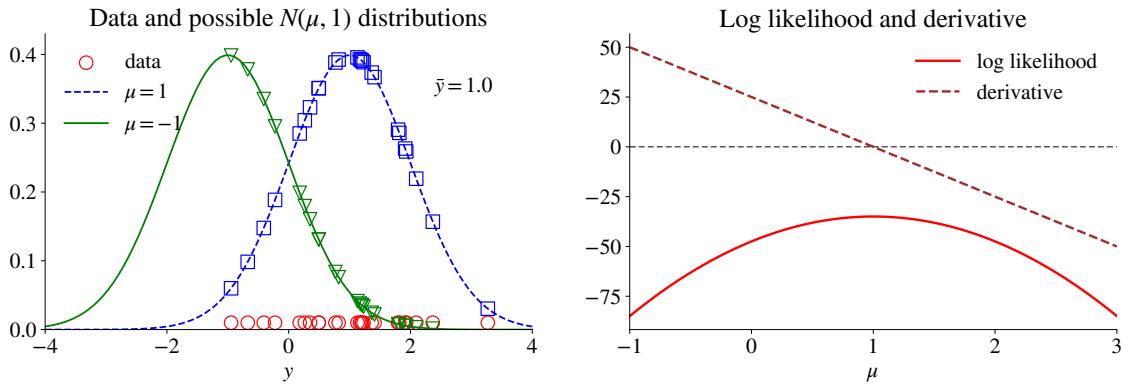


Figure 19.1: Likelihood function for estimating a sample mean

where β denotes the true value of the parameter vector, which we want to estimate by *maximizing* the value of $L(\beta)$. For instance, β could be the mean and variance of y_t . Note that the family of the distribution, for instance, normal, is typically predetermined. The importance of this choice, and what happens when it is wrong, will be discussed in later sections.

Define the *likelihood contribution* of observation t as

$$L_t(\beta) = \text{pdf}(y_t; \beta), \quad (19.2)$$

where the notation is meant to show that the likelihood depends on the parameter vector β and on data for period t . We typically work with the case where the data is *identically distributed*, so the pdf is the same for each observation.

Remark 19.1 (*Normal distribution*) *The pdf of a $N(\mu, \sigma^2)$ distribution is*

$$\text{pdf}(y; \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}u^2/\sigma^2), \text{ where } u = y - \mu,$$

and where $\beta = (\mu, \sigma^2)$. We could also write this as

$$\text{pdf}(y) = \phi(u_t/\sigma)/\sigma,$$

where $\phi(z)$ is the pdf of an $N(0, 1)$ variable z . For the logarithm, we get $\ln \phi(u_t/\sigma) - \ln \sigma$, which is $-\ln(2\pi)/2 - u^2/2 - \ln \sigma$.

If the data points (y_1, y_2, \dots, y_T) are *independent*, then the joint pdf is the product of the

pdfs for different observations

$$L(\beta) = L_1(\beta)L_2(\beta)\dots L_T(\beta), \quad (19.3)$$

where $L_t(\beta)$ is defined in (19.2). When data is not independent, we typically rewrite the likelihood function in terms of a random variable that is, for instance, the residuals rather than y_t . This is discussed in section 19.1.4. Also, when data is not identically distributed, then we model the time variation and add its parameters to the β vector, for instance, when the variance follows a GARCH model. This is discussed in section 19.1.5.

We get the same result if we instead maximize $\ln L(\beta)$, since the logarithmic function is increasing and monotone. Since this might be easier (especially with a normal distribution), we define the *log likelihood function* as

$$\ln L(\beta) = \sum_{t=1}^T \ln L_t(\beta), \quad (19.4)$$

which is the sum of the log likelihood contributions for different observations.

In many cases, implementation of MLE amounts to applying a numerical optimization algorithm to find the β vector that maximizes (19.4). For standard errors (discussed in later sections), derivatives are needed, which often can be calculated numerically as well. However, to understand in more detail how MLE works, it is useful to work through a few pen and paper examples, which is done in the next sections.

Remark 19.2 (*MLE as GMM**) To maximize (19.4), the first order conditions are $\partial \ln L(\hat{\beta})/\partial \beta = \mathbf{0}$, which is the same as requiring the sample average of $\partial \ln L_t(\hat{\beta})/\partial \beta$ to be zero. Thus, $\partial \ln L_t(\hat{\beta})/\partial \beta$ could be used as moment conditions in GMM.

19.1.2 Estimating the Mean with ML

Suppose we know y_t is iid $N(\mu, \sigma^2)$, but we don't know the value of μ . Since we assume that y_t is independent across observations, the log likelihood function is (19.4) with the following contribution from observation t

$$\ln L_t(\beta) = -\ln(2\pi)/2 - \ln(\sigma^2)/2 - (y_t - \mu)^2/(2\sigma^2), \quad (19.5)$$

where $\beta = (\mu, \sigma^2)$. The derivative with respect to μ is $(y_t - \mu)/\sigma^2$, so for the sum (19.4) we have

$$\frac{\partial \ln L(\beta)}{\partial \mu} = \sum_{t=1}^T (y_t - \mu)/\sigma^2. \quad (19.6)$$

The first-order condition for maximizing the log likelihood function is that this should equal zero. Clearly, the value of μ that satisfies this is the traditional sample average

$$\hat{\mu} = \sum_{t=1}^T y_t / T. \quad (19.7)$$

(This result is the same whether you assume or estimate the value of σ^2 .) See Figure 19.1 (right subfigure) for an illustration of how the log likelihood peaks, and thus has a zero derivative, when μ equals the sample average.

However, with another assumption about the distribution of the residuals, MLE might differ from the sample mean. For instance, when the data have a Laplace distribution, $\text{pdf}(y_t - \mu) = \exp(-|y_t - \mu|/\sigma)/(2\sigma)$, then MLE involves minimizing the sum of absolute errors $|y_t - \mu|$, not the squared errors, so the MLE is the sample median.

Example 19.3 (*MLE of the variance*) To instead estimate the variance, assuming you know the mean μ , use (19.5) and note that $\partial \ln L_t / \partial \sigma^2 = -1/(2\sigma^2) + (y_t - \mu)^2 / (2(\sigma^2)^2)$. $\sum_{t=1}^T \partial \ln L_t / \partial \sigma^2$ is then zero at $\hat{\sigma}^2 = \sum_{t=1}^T (y_t - \mu)^2 / T$, which is the traditional formula for a sample variance, albeit divided by T rather than $T - 1$.

19.1.3 MLE of a Regression

Consider a multiple regression model

$$y_t = b'x_t + u_t, \quad (19.8)$$

and assume that u_t is iid $N(0, \sigma^2)$. In this cases the “mean” of y_t is $b'x_t$ (residuals have zero means). That is, the pdf of y_t conditional on x_t , often denoted $\text{pdf}(y_t | x_t)$, is a normal distribution with mean $b'x_t$ and variance σ^2 . MLE estimates $\beta = (b, \sigma^2)$ to fit the data as well as possible. The log likelihood contribution in t is then as in (19.5), but with mean $\mu = b'x_t$

$$\ln L_t(\beta) = -\ln(2\pi)/2 - \ln(\sigma^2)/2 - (y_t - b'x_t)^2 / (2\sigma^2). \quad (19.9)$$

Using this in (19.4) shows that maximum likelihood solves the same problem as OLS: minimize the sum of fitted residuals (the value of σ^2 does not matter for this).

However, with another assumption about the distribution of the residuals, MLE might differ from OLS. Similar to before, when the errors have a Laplace distribution, then MLE gives the least absolute deviations (LAD) estimator, which will be discussed in another chapter.

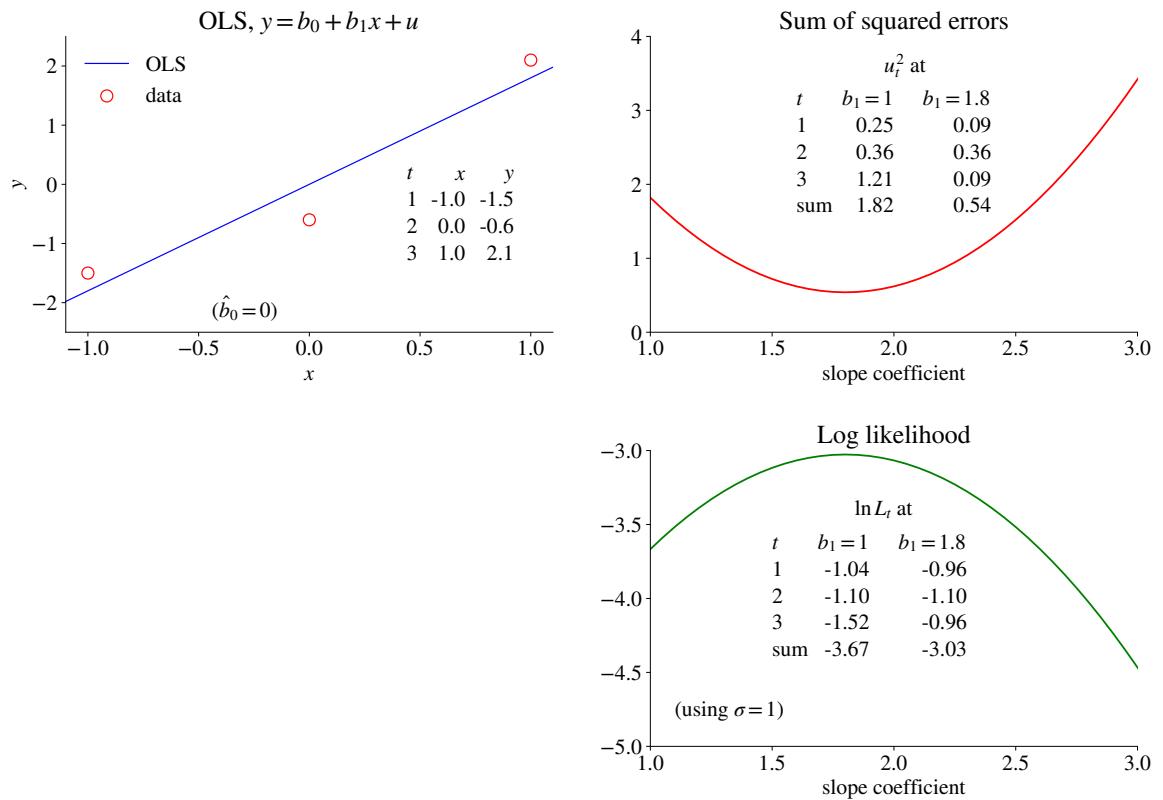


Figure 19.2: Example of OLS and ML estimation

Example 19.4 Figure 19.2 shows for a small sample for a simple regression, where $b = 1.8$ has the highest log likelihood value, since it has the lowest $\sum_{t=1}^T \hat{u}_t^2$.

19.1.4 Correlated Data: Estimating an MA(1) Model

Suppose y_t is autocorrelated, so the assumption of independently distributed data is invalid. However, conditional on past information, the randomness left in y_t could still be iid. For instance, if y_t is an MA(1) process, $y_t = \delta + \varepsilon_t + \theta \varepsilon_{t-1}$, then the pdf of y_t conditional on ε_{t-1} would have a mean of $\delta + \theta \varepsilon_{t-1}$ and a variance equal to $\sigma^2 = \text{Var}(\varepsilon_t)$. If ε_t is normally distributed, then the log likelihood contribution is again as in (19.5), but with $\mu = \delta + \theta \varepsilon_{t-1}$. The β vector now includes $(\delta, \sigma^2, \theta)$, which means that we need to model/estimate also the mechanism behind the autocorrelation.

19.1.5 MLE of a Regression with GARCH(1,1) Errors

Again, consider a regression model like (19.8), but where the variance of the residuals follows a GARCH(1,1) process

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \gamma \sigma_{t-1}^2. \quad (19.10)$$

(It is assumed that $\omega > 0$; $\alpha, \gamma \geq 0$; and $\alpha + \gamma < 1$.) Note that we here use γ to denote the coefficient of σ_{t-1}^2 .

If u_t is $N(0, \sigma_t^2)$, then the log likelihood contribution is again as in (19.5), but with $\mu = b'x_t$ and a time-varying σ^2 according to (19.10). To estimate the model parameters $(b, \omega, \alpha, \gamma)$, we need a numerical optimization routine.

Empirical Example 19.5 (*MLE of a GARCH model*) See Figure 19.3 for a MLE of a GARCH(1,1) model.

19.2 Key Properties of MLE

The properties of MLE depend on whether the likelihood function is correctly specified, that is, whether data actually follows the assumed model and distribution. This section assumes that is the case, but a later section will revisit the topic by studying what happens otherwise.

There are few general results on the small-sample properties of MLE: (1) it may or may not be biased; but (2) the distribution of the estimates should be (approximately) inherited from the distribution assumed in the likelihood function. As an example, estimating an AR(1) with MLE can (when treating the first observation as fixed, not random) amounts to doing OLS, and it is well known that OLS is biased in this case.

In contrast, MLE has very favourable asymptotic (large-sample) properties. In this case, (1) MLE is consistent; (2) MLE is the most efficient/precise estimator; and (3) MLE estimates ($\hat{\beta}$) are normally distributed as in

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V), \quad (19.11)$$

$$V = I(\beta)^{-1} \text{ with } I(\beta) = -E \frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta'} \frac{1}{T}. \quad (19.12)$$

$I(\beta)$ is called the *information matrix* (and should not be confused with the identity matrix). This result assumes that the log likelihood function is indeed twice differentiable (and for

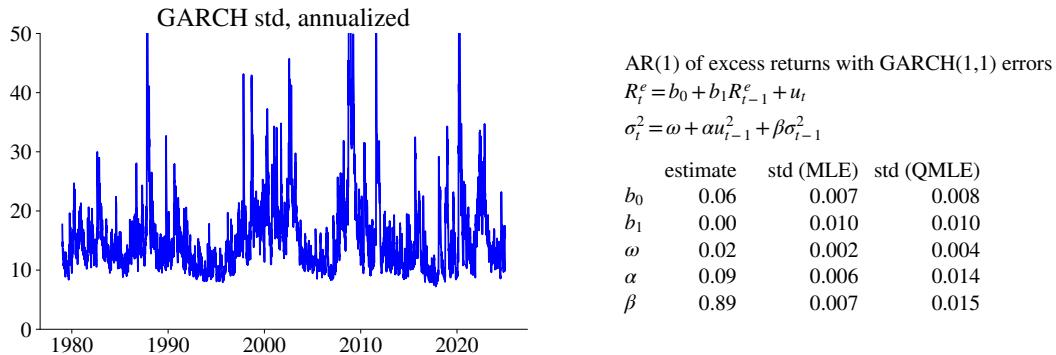


Figure 19.3: GARCH estimates

some results there are even stronger requirements). We can also rewrite the result in (19.11) as

$$\hat{\beta} \xrightarrow{a} N(\beta, V/T), \quad (19.13)$$

where \xrightarrow{a} means “is asymptotically distributed as.” A proof of the distribution is given below.

In practice, the information matrix is calculated (estimated) by using the point estimates and replacing the expectations with sample averages. Often, numerical derivatives are used.

Remark 19.6 ($L(\beta)$ or $L_t(\beta)$ in $I(\beta)$?) *$I(\beta)$ in (19.12) is the expected values of the second derivative of $(1/T) \ln L(\beta)$. Clearly, this could equally well be expressed as the expected value of the derivative of $\ln L_t(\beta)$, so also $I(\beta) = -E \frac{\partial^2 \ln L_t(\beta)}{\partial \beta \partial \beta'}$ is correct. Both expressions are used in the literature.*

Proof (of (19.12)) By the mean value theorem, $\partial \ln L(\hat{\beta}) / \partial \beta / T$ can be expressed as

$$\frac{1}{T} \frac{\partial \ln L(\hat{\beta})}{\partial \beta} = \frac{1}{T} \frac{\partial \ln L(\tilde{\beta})}{\partial \beta} + \frac{1}{T} \frac{\partial^2 \ln L(\tilde{\beta})}{\partial \beta \partial \beta'} (\hat{\beta} - \beta),$$

for some values $\tilde{\beta}$ between $\hat{\beta}$ and β . By the first order conditions, the left hand side is zero. Multiply by \sqrt{T} and solve as

$$\sqrt{T}(\hat{\beta} - \beta) = \left[-\frac{1}{T} \frac{\partial^2 \ln L(\tilde{\beta})}{\partial \beta \partial \beta'} \right]^{-1} \sqrt{T} \frac{1}{T} \frac{\partial \ln L(\tilde{\beta})}{\partial \beta}.$$

In the limit $\hat{\beta}$ (and thus $\tilde{\beta}$) converges to β (if MLE is consistent), so the term in parenthesis equals the information matrix $I(\beta)$ (calculated at the true values β). We thus get

$$\sqrt{T}(\hat{\beta} - \beta) = I(\beta)^{-1} \sqrt{T} \frac{1}{T} \frac{\partial \ln L(\beta)}{\partial \beta}.$$

By a central limit theorem, the last term has a normal distribution. It will have a zero mean, so its variance equals its second moment. We thus have that $\sqrt{T}(\hat{\beta} - \beta) \rightarrow^d N(0, V)$, where $V = I(\beta)^{-1} E[\frac{\partial \ln L_t(\beta)}{\partial \beta} (\frac{\partial \ln L_t(\beta)}{\partial \beta})'] I(\beta)^{-1}$. It can be shown that this simplifies to (19.12) when the likelihood function is correctly specified (“the information matrix equality”, see, for instance, Greene (2018) 14). When the likelihood function is misspecified, but the probability limit of the first order condition defines the true parameters (so MLE is consistent), then this is QMLE (discussed below). \square

The *asymptotic efficiency* means that, asymptotically, there is no other consistent and normally distributed estimator with a lower variance than $I(\beta)^{-1}$ in (19.12) (often called the Cramér-Rao lower bound), and that MLE achieves it. (See Greene (2018) 14 for a detailed discussion.) However, in finite samples there may be better estimators.

The *consistency* of MLE essentially relies on using the right likelihood function (correct distribution and model specification), but this might not be strictly needed, as discussed in section 19.3. As a simple example of when MLE is inconsistent, consider the case of the lagged dependent variable as regressor combined with autocorrelated residuals, for instance, an ARMA(1,1). Applying the approach in (19.9) then gives OLS, but it is well known that OLS is inconsistent in this case.

19.2.1 Examples of the Information Matrix

The first-order condition for estimating the mean is (19.6). Differentiate this with respect to μ to get

$$\frac{\partial^2 \ln L(\beta)}{\partial \mu \partial \mu} = -\frac{T}{\sigma^2}. \quad (19.14)$$

The information matrix is then

$$I(\beta) = -E \frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta'} \frac{1}{T} = \frac{1}{\sigma^2}, \quad (19.15)$$

which we combine with (19.11)–(19.13) to get

$$\sqrt{T}(\hat{\mu} - \mu) \rightarrow^d N(0, \sigma^2) \text{ or } \hat{\mu} \stackrel{a}{\sim} N(\mu, \sigma^2/T). \quad (19.16)$$

This is the standard expression for the distribution of a sample average. In practice, we replace σ^2 by the sample variance.

Example 19.7 (*Information matrix for estimating the variance**) The second derivative of the likelihood contribution in Example 19.3 is $\partial^2 \ln L_t / \partial \sigma^2 \partial \sigma^2 = \gamma^{-2}/2 - \gamma^{-3} (y_t - \mu)^2$,

where we use γ to denote σ^2 (to help remembering that we differentiate wrt. $\gamma = \sigma^2$, not σ). Take expectations and recall that $E(y_t - \mu)^2 = \gamma$. Simplify to get $E \partial^2 \ln L_t / \partial \sigma^2 \partial \sigma^2 = -\gamma^{-2}/2$. It follows that $V/T = 2\gamma^2 = 2\sigma^4$.

Example 19.8 (Information matrix of a regression*) The first derivative of (19.9) is $\partial \ln L_t(\beta)/\partial b = x_t(y_t - b'x_t)/\sigma^2$, so the second derivative is $\partial^2 \ln L_t(\beta)/\partial b \partial b' = -x_t x_t'/\sigma^2$. This gives $V/T = \sigma^2 (\sum_{t=1}^T x_t x_t')^{-1}$, which is the traditional result with iid residuals.

19.3 QMLE

A MLE based on the wrong likelihood function *may* still be useful. Suppose we use the likelihood function L and get estimates $\hat{\beta}$ by solving the first-order conditions

$$\frac{\partial \ln L(\hat{\beta})}{\partial \beta} = \mathbf{0} \quad (19.17)$$

If L is wrong, then we are maximizing the wrong thing. For instance, we might have constructed L by assuming that the regression residuals are normally distributed, while they are, in fact, following another distribution.

Example 19.9 (LS and QMLE) In a linear regression, $y_t = x_t' \beta + \varepsilon_t$, the first order conditions for MLE based on the assumption that $\varepsilon_t \sim N(0, \sigma^2)$ are $\sum_{t=1}^T x_t(y_t - x_t' \hat{\beta}) = \mathbf{0}$. This has an expected value of zero (at the true parameters), even if the shocks have a, say, t -distribution.

The example suggests that if the first order conditions hold (in expectation or probability limit) at the true parameter values (β)

$$E \frac{\partial \ln L(\beta)}{\partial \beta} = \mathbf{0}, \quad (19.18)$$

then the estimates are still consistent. In this case, $\sqrt{T}(\hat{\beta} - \beta) \rightarrow^d N(0, V)$, but with the “sandwich” variance-covariance matrix

$$V = I(\beta)^{-1} E \left[\frac{\partial \ln L_t}{\partial \beta} \left(\frac{\partial \ln L_t}{\partial \beta} \right)' \right] I(\beta)^{-1}. \quad (19.19)$$

The proof is the same as that of (19.12). This is often referred to as *quasi-MLE* (or pseudo-MLE).

It is difficult to tell whether we have the right likelihood function. For that reason, the practical implication of (19.19) is that it provides a more robust way of calculating standard errors. Also, differences between standard errors based on the information matrix and from the sandwich approach can be an indicator of issues with the likelihood function. This solves part of the issue with specifying the likelihood function, although it does not guarantee that the estimator is actually consistent.

Empirical Example 19.10 (*MLE of a GARCH model*) See Figure 19.3 for a MLE of a GARCH(1,1) model. For some of the parameters, the robust standard errors are distinctly higher.

Further Reading

See Verbeek (2017) 6, and Greene (2018) and Hansen (2022b) 10 for more detailed analysis.

Chapter 20

ARCH and GARCH

20.1 Heteroskedasticity

20.1.1 Descriptive Statistics of Heteroskedasticity

Time-variation in volatility (heteroskedasticity) is a common feature of financial data. The perhaps most straightforward way to gauge heteroskedasticity is to estimate a time series of variances on “rolling samples.” For a zero-mean variable, u_t , this could be

$$\sigma_t^2 = (u_{t-1}^2 + u_{t-2}^2 + \dots + u_{t-q}^2)/q. \quad (20.1)$$

Notice that σ_t^2 depends on lagged information, and could therefore be considered the prediction (made in $t - 1$) of the volatility in t . This method can be used for detecting time variation in volatility—and the estimates (for instance, over a month) are sometimes called *realized volatility*. A similar method can also be used to measure seasonality in volatility by estimating the variance for each “season”—for instance, Mondays.

Empirical Example 20.1 (*Time-varying equity volatility*) See Figure 20.1 for a comparison of realized 22-day S&P 500 volatility with an option based measure (VIX).

Empirical Example 20.2 (*Time-varying FX volatility*) See Figure 20.2 for an example of seasonality in FX returns.

Empirical Example 20.3 (*Predicting volatility*) Volatility and correlations are often predictable, at least for horizons up to a couple of months. See Table 20.1 for examples of very simple, but rather effective, prediction equations.

Unfortunately, the approach in (20.1) can produce quite abrupt changes in the estimate. An alternative is to apply an exponentially weighted moving average (EWMA) estimator,

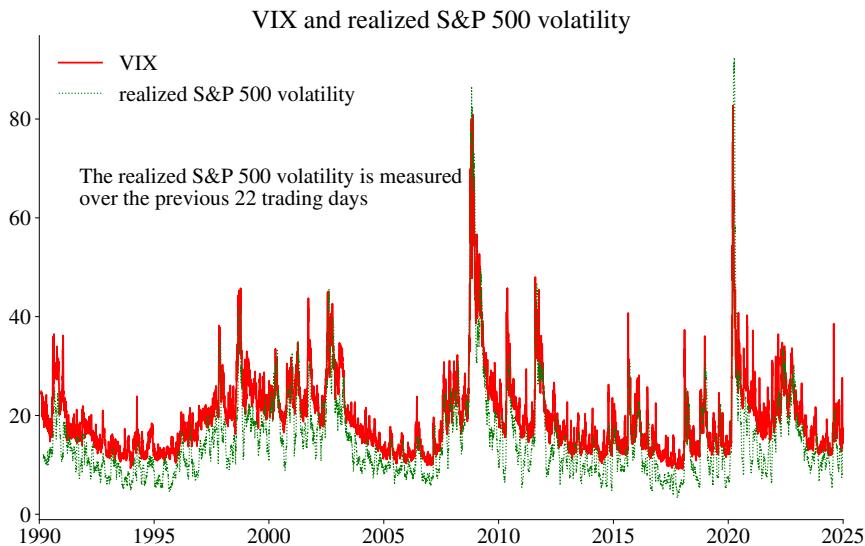


Figure 20.1: VIX and realized volatility (variance)

which uses all data points since the beginning of the sample, but where recent observations carry larger weights. The weight for lag s is $(1 - \lambda)\lambda^s$ where $0 < \lambda < 1$, so the weights sum to one

$$\sigma_t^2 = (1 - \lambda)(u_{t-1}^2 + \lambda u_{t-2}^2 + \lambda^2 u_{t-3}^2 + \dots). \quad (20.2)$$

See Figure 20.3 for an illustration of the weights. This can also be calculated recursively

$$\sigma_t^2 = (1 - \lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2. \quad (20.3)$$

The initial value (before the sample) could be assumed to be zero or (perhaps better) the unconditional variance in a historical sample. This method is commonly used by practitioners. For instance, the RiskMetrics (see JP Morgan (1996)) approach uses this method with $\lambda \approx 0.94$ on daily data. Alternatively, λ can be chosen to minimize some criterion function.

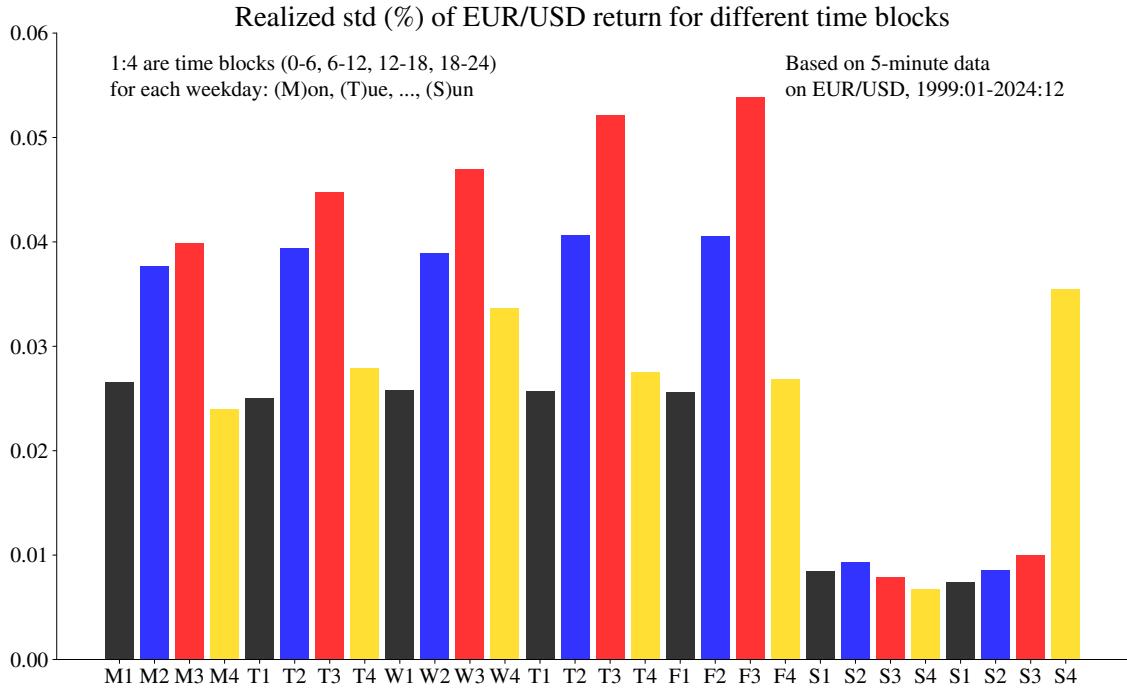


Figure 20.2: Standard deviation

Remark 20.4 (*EWMA on a moving data window*) To implement the EWMA method on a moving data window over q data points ($t - q$ to $t - 1$) use

$$\sigma_t^2 = (u_{t-1}^2 + \lambda u_{t-2}^2 + \dots + \lambda^{q-1} u_{t-q}^2)(1 - \lambda)/(1 - \lambda^q).$$

This makes sure that the weights sum to one. The recursive form is $\sigma_t^2 = (u_{t-1}^2 - \lambda^q u_{t-q-1}^2)(1 - \lambda)/(1 - \lambda^q) + \lambda \sigma_{t-1}^2$.

	(1)	(2)	(3)
log RV ($t - 22$)	0.67 (20.01)	0.10 (2.09)	
log VIX ($t - 22$)		1.05 (26.59)	0.94 (15.36)
constant	0.85 (9.90)	-0.48 (-4.15)	-0.40 (-3.51)
R^2	0.45	0.56	0.57
obs	8736	8757	8736

Table 20.1: Regression of 22-day log realized S&P return volatility 1990:02-2024:12. All daily observations are used, so the residuals are likely to be autocorrelated. Numbers in parentheses are t-stats, based on Newey-West with 30 lags.

20.1.2 Heteroskedastic Residuals in a Regression

Suppose we have a regression model

$$y_t = x_t' \beta + u_t, \quad (20.4)$$

where $E u_t = 0$ and $\text{Cov}(x_{it}, u_t) = 0$. In the standard case, we assume that u_t is iid (independently and identically distributed), which rules out heteroskedasticity.

OLS is still a useful estimator when the residuals are heteroskedastic. It is consistent and it is reasonably efficient (in terms of the variance of the estimates), although not the most efficient (MLE is). However, the standard expression for the standard errors of the coefficients is potentially incorrect (in particular, when the volatility is related to the regressors).

Alternatively, we could combine the regression model (20.4) with a (G)ARCH structure of the residual. This provides forecast of volatility, which can be useful for risk control or asset pricing. Also, as a by-product we get the correct standard errors.

20.1.3 Autoregressive Conditional Heteroskedasticity (ARCH)

Autoregressive heteroskedasticity is a special form of heteroskedasticity, and it is often found in financial data which shows volatility clustering.

To test for ARCH features, *Engle's ARCH test* is perhaps the most straightforward. Let u_t be a zero-mean variable, for instance, the fitted residuals from OLS. Then, estimate an

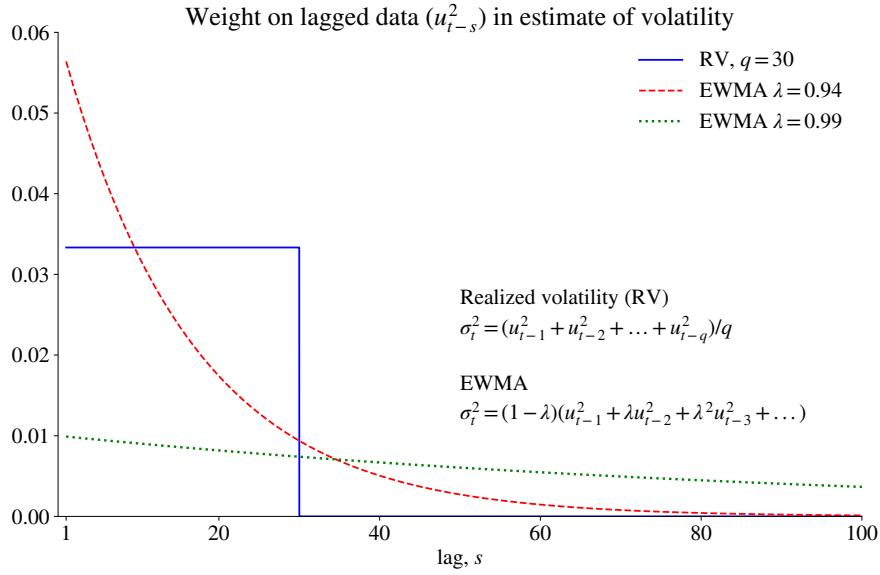


Figure 20.3: Weights on old data in the EMA approach to estimate volatility

AR(q) model for the squares

$$u_t^2 = \omega + a_1 u_{t-1}^2 + \dots + a_q u_{t-q}^2 + v_t. \quad (20.5)$$

Under the null hypothesis of no ARCH effects, all slope coefficients are zero. This can be tested by noting that, under the null hypothesis, $TR^2/(1 - R^2) \sim \chi_q^2$. Notice, however, that such ARCH effects may or may not invalidate the OLS standard errors for (20.4). Rather, use White's test to assess that.

20.2 ARCH Models

This section discusses the Autoregressive Conditional Heteroskedasticity (ARCH) model, which assumes that volatility depends on past volatility.

20.2.1 Properties of ARCH(1)

An ARCH(1) model of the residual in the regression equation (20.4), or some other zero-mean variable, can be written

$$u_t \sim N(0, \sigma_t^2), \text{ with} \quad (20.6)$$

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2, \text{ with } \omega > 0 \text{ and } 0 \leq \alpha < 1. \quad (20.7)$$

Notice that σ_t^2 is the conditional variance of u_t , and it is known already in $t - 1$. (Warning: some authors use a different convention for the time subscripts.) The non-negativity restrictions on ω and α are needed in order to guarantee $\sigma_t^2 > 0$ and the upper bound $\alpha < 1$ is needed in order to make the conditional variance stationary (more later).

If we assume that u_t is iid $N(0, \sigma_t^2)$, then the distribution of u_t , conditional on the information in $t - 1$, is $N(0, \sigma_t^2)$, where σ_t^2 is known in $t - 1$. Therefore, the one-step ahead distribution is normal, which can be used for estimating the model with ML. To get a distribution with even fatter tails, we could alternatively assume $u_t = \epsilon_t \sigma_t$, where ϵ_t has a student's t_v -distribution with $v > 2$ degrees of freedom.

Remark 20.5 (*Interpreting the variances of a t_v variable) If $\epsilon_t \sim t_v$ with $v > 2$, then it has a variance of $v/(v - 2)$. The variance of $u_t = \epsilon_t \sigma_t$ is thus $\sigma_t^2 v/(v - 2)$.

However, the distribution of u_{t+1} (still conditional on the information in $t - 1$) is more complicated. In particular, its variance is $\sigma_{t+1}^2 = \omega + \alpha u_t^2$, where u_t contains a random element. This makes u_{t+1} have a non-normal distribution. In fact, it will have fatter tails (excess kurtosis) than a normal distribution with the same variance, which is a common feature of financial data.

It is straightforward to show that the ARCH(1) model implies that the forecast the future conditional variance in $t + s$ is

$$E_t \sigma_{t+s}^2 = \bar{\sigma}^2 + \alpha^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2), \text{ with } \bar{\sigma}^2 = \omega/(1 - \alpha), \quad (20.8)$$

where $\bar{\sigma}^2$ is the unconditional variance and where we recall that $\sigma_{t+1}^2 = \omega + \alpha u_t^2$ is known in t . The conditional volatility behaves like an AR(1), and $0 \leq \alpha < 1$ is necessary to keep it positive and stationary.

Empirical Example 20.6 (ARCH and GARCH models) See Figure 20.4 for an illustration of the fitted volatilities.

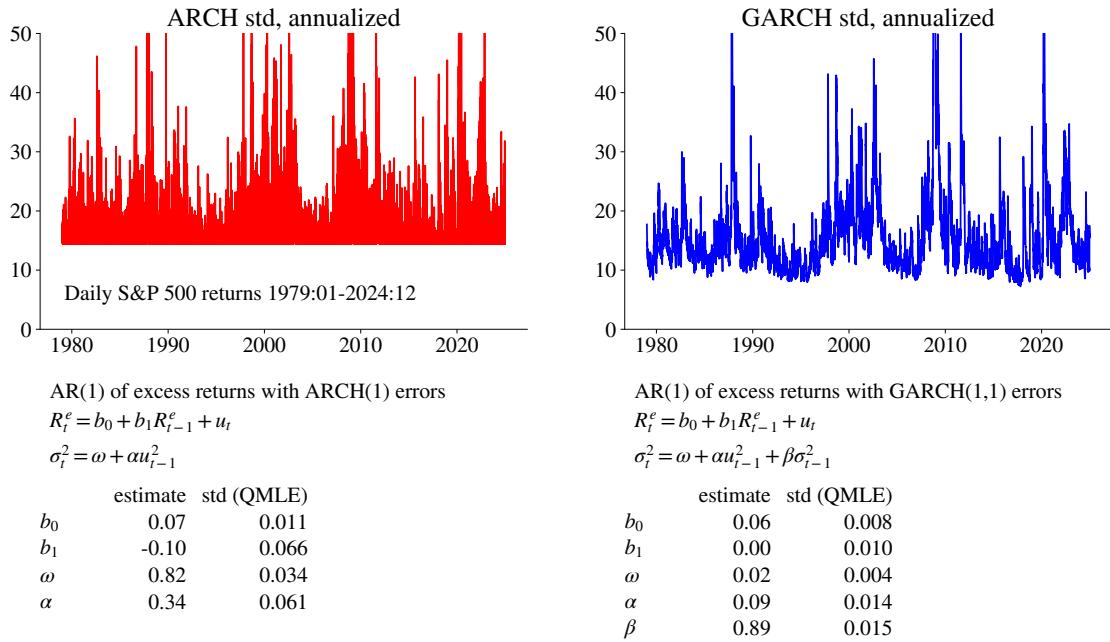


Figure 20.4: ARCH and GARCH estimates

20.2.2 Estimation of the ARCH(1) Model

The most common way to estimate the model is to assume that $u_t \sim N(0, \sigma_t^2)$, as we did in (20.6) and to set up the likelihood function as

$$\ln L = \sum_{t=1}^T \ln L_t, \text{ where } \ln L_t = \phi(u_t/\sigma_t) - \ln \sigma_t, \quad (20.9)$$

where $\phi()$ is the pdf of an $N(0, 1)$ variable. (L_t is the same as the log pdf of an $N(0, \sigma^2)$ variable.) The estimates are found by maximizing the likelihood function (choosing the parameters). This has to be done by a numerical optimization routine, which should preferably impose the constraints in (20.7).

Remark 20.7 (*Alternative expression for $\ln L_t$) We could equivalently use $\ln L_t = -\ln(2\pi)/2 - \ln(\sigma_t^2)/2 - u_t^2/(2\sigma_t^2)$.

If u_t is just a zero-mean variable (so we have no regression equation), then this maximisation amounts to choosing the parameters (ω and α) in (20.7). Instead, if u_t is a residual from a regression equation (20.4), then we instead need to choose both the regression coefficients (β) in (20.4) and the parameters (ω and α) in (20.7).

In either case, we need a starting value of $\sigma_1^2 = \omega + \alpha u_0^2$. This most common approach is to use the first observation as a “starting point,” that is, we actually have a sample from ($t = 0$) to T , but observation 0 is only used to construct a starting value of σ_1^2 , and only observations 1 to T are used in the calculation of the likelihood function value.

Notice that if we estimate a regression function and an ARCH model simultaneous with MLE, then we automatically get the right standard errors of the regression coefficients from either the information matrix or the sandwich approach.

Remark 20.8 (*Regression with ARCH(1) residuals*) *To estimate the full model (20.4) and (20.7) by MLE, we can do as follows.*

First, guess values of the parameters β , ω , and α . The guess of β can be taken from an LS estimation of (20.4), and the guess of ω and α from an LS estimation of $\hat{u}_t^2 = \omega + \alpha \hat{u}_{t-1}^2 + \varepsilon_t$ where \hat{u}_t are the fitted residuals from the LS estimation of (20.4).

Second, loop over the sample (first $t = 1$, then $t = 2$, etc.) and calculate u_t from (20.4) and σ_t^2 from (20.7). Plug in these numbers in (20.9) to find the likelihood value.

Third, make better guesses of the parameters and do the second step again. Repeat until the likelihood value converges (at a maximum).

Remark 20.9 (*Imposing parameter constraints on ARCH(1)**) *If the numerical optimization algorithm does not handle constraints, iterate over values of $(\beta, \tilde{\omega}, \tilde{\alpha})$ where $\omega = \tilde{\omega}^2$ and $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha})]$. This imposes the restrictions in (20.7).*

It is sometimes found that the *standardized residuals*, u_t/σ_t , still have too fat tails compared with $N(0, 1)$. This would violate the assumption about a normal distribution in (20.9). Estimation using other likelihood functions, for instance, for a t-distribution can then be used. Or the estimation can be interpreted as a quasi-MLE, which requires different a calculation of the variance-covariance matrix of the parameters.

It is straightforward to add more lags to (20.7). For instance, an ARCH(p) would be

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \dots + \alpha_p u_{t-p}^2. \quad (20.10)$$

The form of the likelihood function is the same except that we now need p starting values and that constraints are $0 \leq \alpha_j$ and $\sum_{j=1}^p \alpha_j \leq 1$.

20.3 GARCH Models

Instead of specifying an ARCH model with many lags, it is often more convenient to specify a low-order Generalized ARCH (GARCH) model. The GARCH(1,1) is a simple

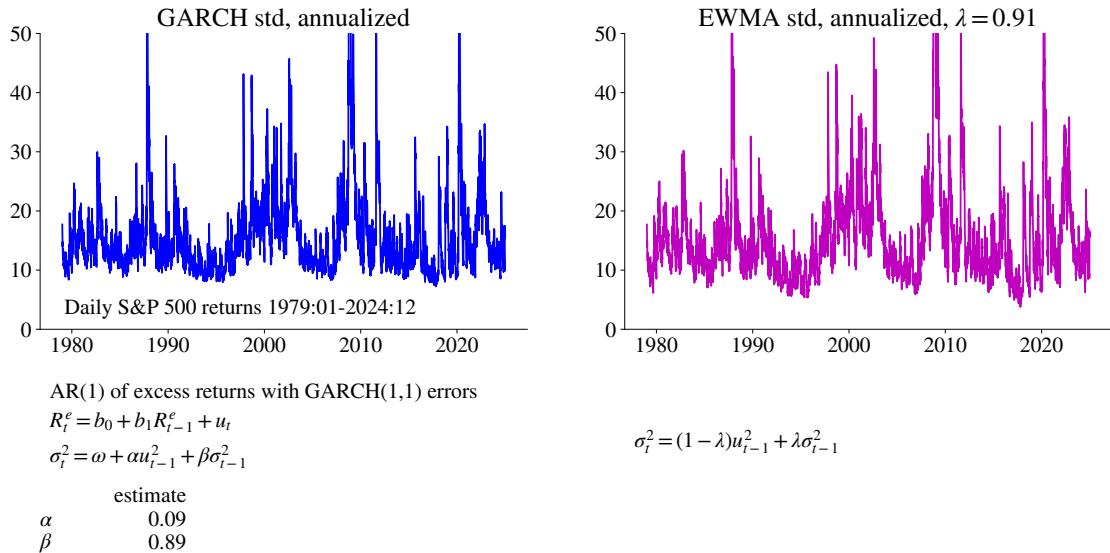


Figure 20.5: Conditional standard deviation, estimated by GARCH(1,1) model

and perhaps surprisingly general model, where the volatility follows

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2, \text{with} \quad (20.11)$$

$$\omega > 0; \alpha, \beta \geq 0; \text{ and } \alpha + \beta < 1.$$

The non-negativity restrictions are needed in order to guarantee that $\sigma_t^2 > 0$ in all periods. The upper bound $\alpha + \beta < 1$ is needed in order to make the σ_t^2 stationary and therefore the unconditional variance finite.

Remark 20.10 *The GARCH(1,1) has many similarities with the exponential moving average estimator of volatility (20.3). The main differences are that the exponential moving average does not have a constant and volatility is non-stationary (the coefficients sum to unity).*

Empirical Example 20.11 (GARCH(1,1) vs EWMA) See Figure 20.5 for an example.

The GARCH(1,1) corresponds to an ARCH(∞) with geometrically declining weights, which is seen by solving (20.11) recursively by substituting for σ_{t-1}^2 (and then $\sigma_{t-2}^2, \sigma_{t-3}^2, \dots$)

$$\sigma_t^2 = \omega / (1 - \beta) + \alpha \sum_{j=0}^{\infty} \beta^j u_{t-1-j}^2. \quad (20.12)$$

This suggests that a GARCH(1,1) might be a reasonable approximation of a high-order ARCH.

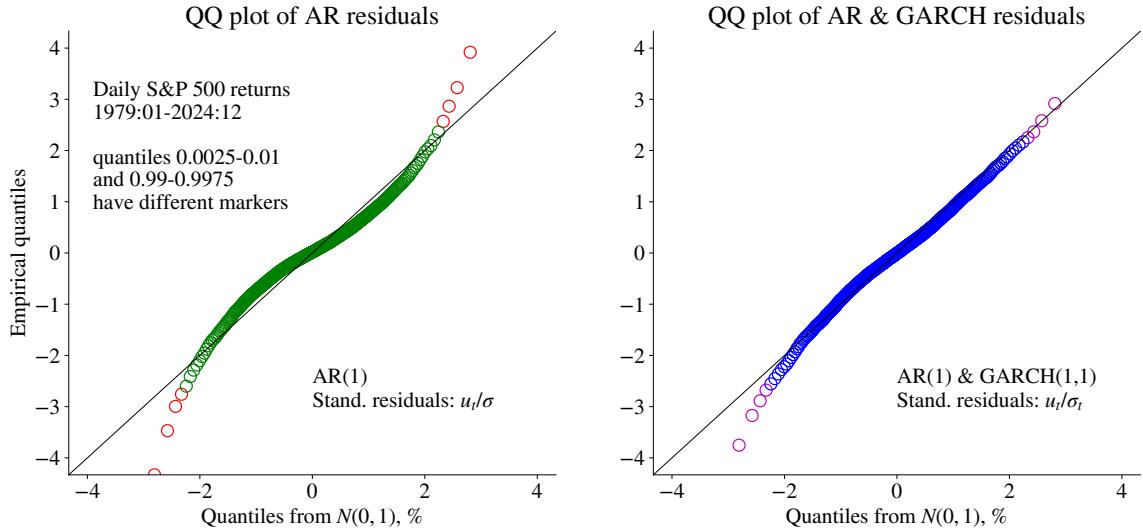


Figure 20.6: QQ-plot of residuals

Also, the GARCH(1,1) model implies that we can forecast the future conditional variance (σ_{t+s}^2) as

$$E_t \sigma_{t+s}^2 = \bar{\sigma}^2 + (\alpha + \beta)^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2), \text{ with } \bar{\sigma}^2 = \omega / (1 - \alpha - \beta), \quad (20.13)$$

which is of the same form as for the ARCH model (20.8), but where the sum of α and β is like an AR(1) parameter.

To estimate the model consisting of (20.4) and (20.11) we can still use the likelihood function (20.9) and do a MLE (but we now have to choose a value of β as well). We typically create the starting value of u_0^2 as in the ARCH(1) model, but this time we also need a starting value of σ_0^2 . It is often recommended to use $\sigma_0^2 = \text{Var}(u_t)$.

Remark 20.12 (*Imposing parameter constraints on GARCH(1,1)*) To impose the restrictions in (20.11), iterate over values of $(b, \tilde{\omega}, \tilde{\alpha}, \tilde{\beta})$ and let $\omega = \tilde{\omega}^2$, $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$, and $\beta = \exp(\tilde{\beta})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$.

Empirical Example 20.13 (*Distribution of normalised residuals*) See Figure 20.6 for evidence of how the residuals become more normally distributed once the heteroskedasticity is handled.

Empirical Example 20.14 (*Value at Risk*) The 95% value at risk (as fraction of the investment) is the (negative of the) 0.05 quantile of the return distribution. In particular,

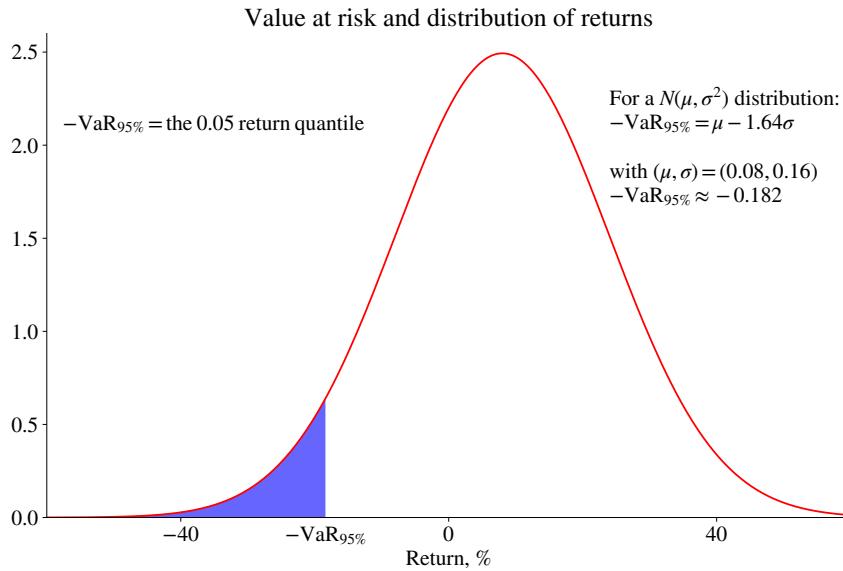


Figure 20.7: Value at risk

$VaR_{0.95} = 0.08$ says that there is only an 5% chance that the loss will be greater than 8% of the investment. See Figure 20.7 for an illustration. When the return has an $N(\mu, \sigma^2)$ distribution, then $VaR_{95\%} = -(\mu - 1.64\sigma)$. See Figure 20.8 for an example of time-varying VaR, based on a GARCH model.

20.4 Non-Linear Extensions

There is a wide array of extensions to the basic GARCH model. An *asymmetric GARCH* (Glosten, Jagannathan, and Runkle (1993), GJR) can be constructed as

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \delta(u_{t-1} < 0) u_{t-1}^2, \text{ where} \quad (20.14)$$

$\delta(z) = 1$ if z is true and 0 otherwise. This means that the effect of the shock u_{t-1}^2 is $\alpha + \gamma$ if the shock was negative and α if the shock was positive. With $\gamma > 0$, volatility increases more in response to a negative u_{t-1} (“bad news”) than to a positive.

The *EGARCH* (exponential GARCH, Nelson (1991)) sets

$$\ln \sigma_t^2 = \omega + \alpha |u_{t-1}|/\sigma_{t-1} + \beta \ln \sigma_{t-1}^2 + \gamma u_{t-1}/\sigma_{t-1}. \quad (20.15)$$

Apart from being written in terms of the log (which is a useful device to make $\sigma_t^2 > 0$ hold

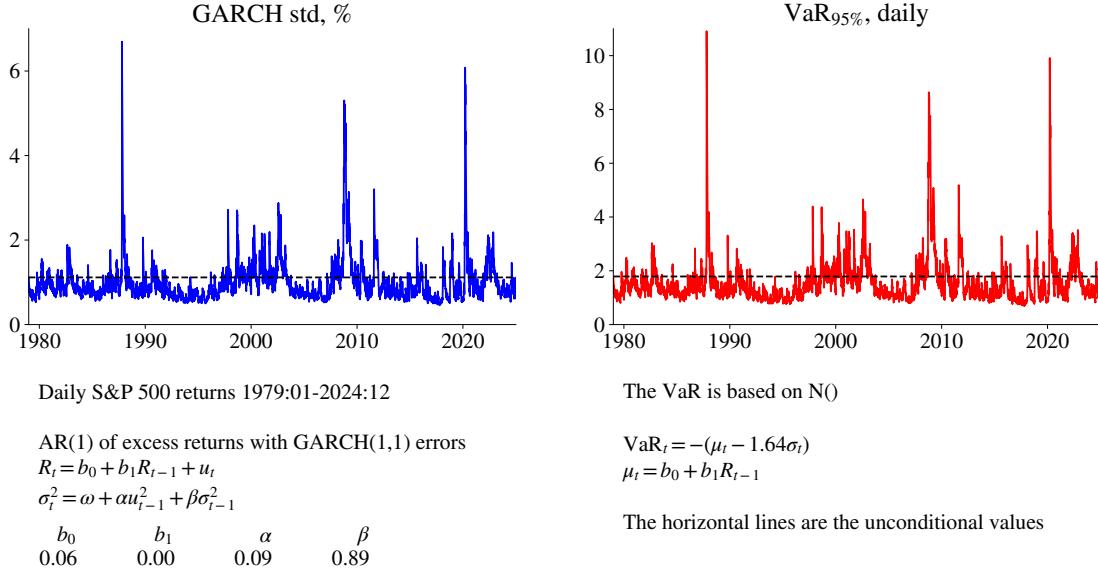


Figure 20.8: Conditional volatility and VaR

without any restrictions on the parameters), this is an asymmetric model. The $|u_{t-1}|$ term is symmetric: both negative and positive values of u_{t-1} affect the volatility in the same way. The linear term in u_{t-1} modifies this to make the effect asymmetric. In particular, if $\gamma < 0$, then the volatility increases more in response to a negative u_{t-1} than to a positive. This model is stationary if $|\beta| < 1$. Both (20.14) and (20.15) can be estimated by MLE using the log likelihood function (20.9).

20.5 Multivariate (G)ARCH

20.5.1 Descriptive Statistics

A first crude approach to estimate a time varying covariance of two series (u_{it} and u_{jt}) is the EWMA

$$\sigma_{ij,t} = (1 - \lambda)u_{i,t-1}u_{j,t-1} + \lambda\sigma_{ij,t-1}. \quad (20.16)$$

We could also use a moving data window as in Remark 20.4. Combining this with similar EWMA estimates (applying (20.3) or Remark 20.4) of the time varying variance of each series ($\sigma_{i,t}^2$ and $\sigma_{j,t}^2$), allows us to calculate a time varying correlation as

$$\rho_{ij,t} = \sigma_{ij,t} / \sqrt{\sigma_{ii,t}\sigma_{jj,t}}. \quad (20.17)$$

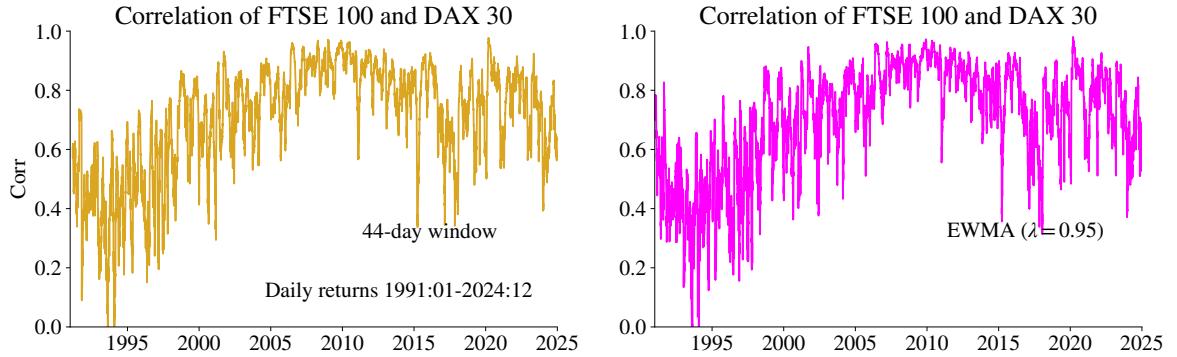


Figure 20.9: Time-varying correlations (different EWMA estimates)

This can be applied to several pairs, to create a correlation matrix of n series. The λ weights need not be the same as for the variances.

Empirical Example 20.15 (*Time-varying equity index correlations*) See Figure 20.9.

20.5.2 Different Multivariate Models

This section gives a brief summary of some multivariate models of heteroskedasticity. Suppose u_t is an $n \times 1$ vector. For instance, u_t could be the residuals from n different regressions or just n different demeaned return series.

We define the conditional (on the information set in $t - 1$) covariance matrix of u_t as

$$S_t = E_{t-1} u_t u_t'. \quad (20.18)$$

(The S_t notation is used to avoid the potentially confusing Σ_t . Still, we will refer to the elements as $\sigma_{ij,t}$.)

Remark 20.16 (*The vech operator*) *vech(A)* of a matrix A gives a vector with the elements on and below the principal diagonal A stacked on top of each other (column wise). For instance, $vech(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix})$ is a column vector with (a_{11}, a_{21}, a_{22}) .

It may seem as if a multivariate (matrix) version of the GARCH(1,1) model would be simple, but it is not. The reason is that it would contain far too many parameters. Although we only need to care about the unique elements of S_t , that is, $vech(S_t)$, this still gives very many parameters

$$vech(S_t) = C + A vech(u_{t-1} u_{t-1}') + B vech(S_{t-1}). \quad (20.19)$$

For instance, with $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = C + A \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1}u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + B \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \quad (20.20)$$

where C is 3×1 , A is 3×3 , and B is 3×3 . This gives 21 parameters, and with $n > 2$ series it quickly becomes very hard to manage, as there will be $n^2(n+1)^2/4$ elements in each of A and B . For instance, with $n = 5$, A and B are 15×15 , so 225 coefficients in each. Basically, the number of parameters grows with n^4 .

We thus have to limit the number of parameters. We also have to find a way to impose restrictions so S_t is positive definite to guarantee that every possible linear combination of the variables has a positive variance (compare the restrictions of positive coefficients in (20.11)).

The Diagonal Model

The *diagonal model* assumes that A and B are diagonal. This means that every element of S_t follows a univariate process. With $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1}u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \quad (20.21)$$

which gives $3 + 3 + 3 = 9$ parameters. More generally, $n(n+1)/2$. To make sure that S_t is positive definite we have to impose further restrictions. The obvious drawback of this model is that there is no spillover of volatility from one variable to another.

The Constant Correlation Model

The *constant correlation model* (CCC) assumes that every variance follows a univariate GARCH process and that the conditional correlations are constant. With $n = 2$ the covariance matrix is

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \quad (20.22)$$

and each of $\sigma_{11,t}$ and $\sigma_{22,t}$ follows a GARCH process. Assuming a GARCH(1,1) as in (20.11) gives just 7 parameters (2×3 GARCH parameters and one correlation, more

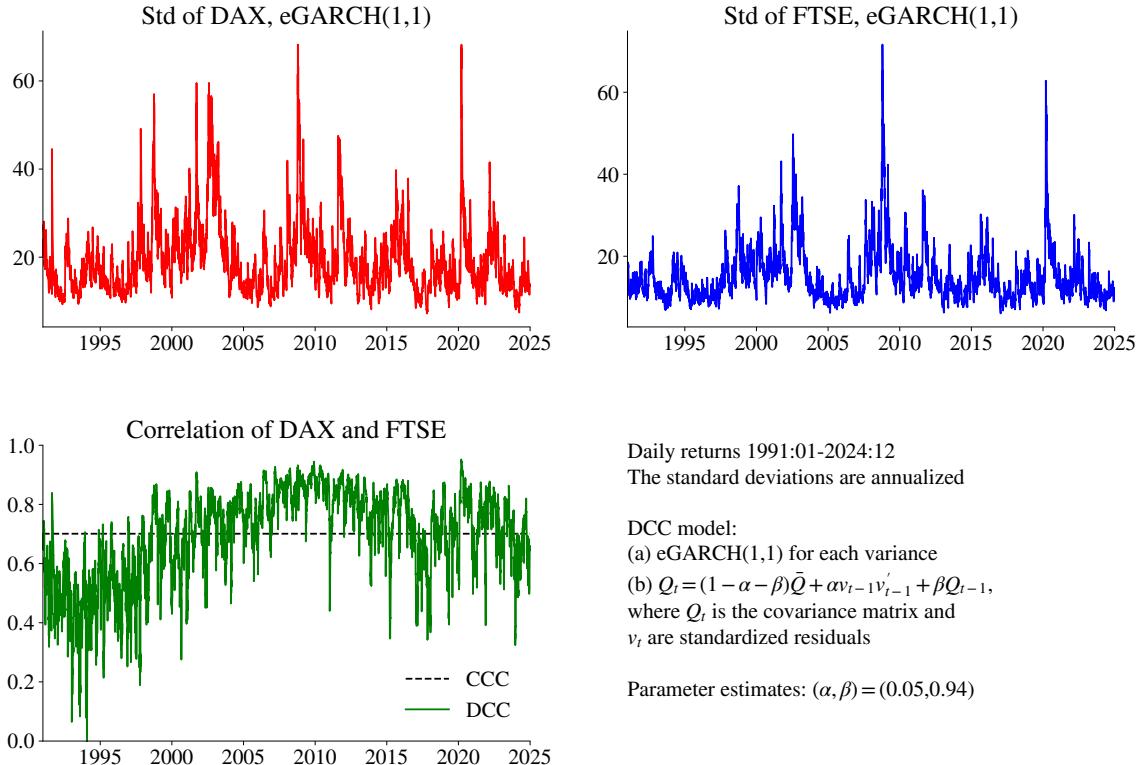


Figure 20.10: Results for multivariate eGARCH models

generally $3n + n(n - 1)/2$). The price is, of course, the assumption of no movements in the correlations. To get a positive definite S_t , each individual (e)GARCH model must generate a positive variance (same restrictions as before), and all the estimated (constant) correlations must be between -1 and 1 .

Remark 20.17 (*Estimating the constant correlation model*) A simple estimation approach is to first estimate the individual (e)GARCH processes and then estimate the correlation of the standardized residuals $u_{1t} / \sqrt{\sigma_{11,t}}$ and $u_{2t} / \sqrt{\sigma_{22,t}}$.

By also specifying how the correlation can change over time, we get a *dynamic correlation model* (DCC). It is slightly harder to estimate.

Empirical Example 20.18 (*Time-varying equity index correlations*) See Figure 20.10 for an illustration and Figure 20.9 for a comparison with the EWMA approach.

The Dynamic Correlation Model

The *dynamic correlation model* (DCC) discussed in Engle (2002)) allows the correlation to change over time. The model assumes that each conditional variance follows a univariate volatility model, for instance, an (e)GARCH, and the conditional correlation matrix is allowed to follow its own a univariate GARCH-like equation.

Remark 20.19 (*Covariance to correlation matrix*) If σ is a vector of standard deviations and R is a correlation matrix, then the variance-covariance matrix is $S = DRD$, where $D = \text{diag}(\sigma)$ create a diagonal matrix with σ along the main diagonal. Similarly, $R = D^{-1}SD^{-1}$. Clearly, D is a diagonal matrix with $1/\sigma_i$ along the main diagonal.

First, estimate conditional variance ($\sigma_{ii,t}$) for each series, for instance, by an (e)GARCH model. Second, let $v_{i,t}$ be the standardized residual for series i

$$v_{i,t} = u_{i,t} / \sqrt{\sigma_{ii,t}}, \quad (20.23)$$

and let \bar{Q} be the unconditional correlation matrix of the v_t vector.

Third, model a “GARCH(1,1)” structure for the covariance matrix of v_t

$$Q_t = (1 - \alpha - \beta)\bar{Q} + \alpha v_{t-1} v'_{t-1} + \beta Q_{t-1}, \quad (20.24)$$

where α and β are two *scalars*. We typically require that $\alpha, \beta \geq 0$ and that $\alpha + \beta < 1$, although the EWMA approach will use $\alpha + \beta = 1$. This model adds only 2 parameters (plus those in \bar{Q}) to the univariate models. Note that Q_t is almost a correlation matrix, since it refers to standardized residuals. However, we take an additional step to guarantee that we are working with a correlation matrix. Thus, fourth, calculate the correlation matrix R_t implied by Q_t , by letting element ij be

$$r_{ij,t} = q_{ij,t} / \sqrt{q_{ii,t} q_{jj,t}} \quad (20.25)$$

where $q_{ij,t}$ is element (i, j) of Q_t .

The fifth and final step is to combine this with the estimated variances ($\sigma_{ii,t}$) to get a full variance-covariance matrix of the original (not standardised) residuals (S) where element (i, j) is

$$\sigma_{ij,t} = \sqrt{\sigma_{ii,t}^2 \sigma_{jj,t}^2 r_{ij,t}}. \quad (20.26)$$

The basic idea of this model is to estimate a conditional correlation matrix as in (20.25) and then scale up with conditional variances (from univariate volatility models) to get a

conditional covariance matrix as in (20.26). To understand the dynamics of this model, consider element $q_{12,t}$ from (20.24)

$$q_{12,t} = (1 - \alpha - \beta)\bar{q}_{12} + \alpha v_{1,t-1}v_{2,t-1} + \beta q_{12,t-1}. \quad (20.27)$$

This “correlation” (almost) of series 1 and 2 depends positively in the lagged correlation ($q_{12,t-1}$), which generates persistence. Also, when the residuals have the same sign, so $v_{1,t-1}v_{2,t-1} > 0$, then the correlation increases (and vice versa).

This model can be implemented as an EWMA model or estimated with MLE. Both versions are discussed below.

Example 20.20 (*Dynamic correlation model, n = 2*) *To estimate the dynamic correlations, we first calculate (where α and β are two scalars)*

$$\begin{bmatrix} q_{11,t} & q_{12,t} \\ q_{12,t} & q_{22,t} \end{bmatrix} = (1 - \alpha - \beta) \begin{bmatrix} 1 & \bar{q}_{12} \\ \bar{q}_{12} & 1 \end{bmatrix} + \alpha \begin{bmatrix} v_{1,t-1} \\ v_{2,t-1} \end{bmatrix} \begin{bmatrix} v_{1,t-1} \\ v_{2,t-1} \end{bmatrix}' + \beta \begin{bmatrix} q_{11,t-1} & q_{12,t-1} \\ q_{12,t-1} & q_{22,t-1} \end{bmatrix},$$

where $v_{i,t} = u_{i,t}/\sqrt{\sigma_{ii,t}}$ and \bar{q}_{ij} is the unconditional correlation of $v_{i,t}$ and $v_{j,t}$. Each of σ_{11t} and σ_{22t} follows an (e)GARCH process. We get the conditional correlations by

$$\begin{bmatrix} 1 & r_{12,t} \\ r_{12,t} & 1 \end{bmatrix} = \begin{bmatrix} 1 & q_{12,t}/\sqrt{q_{11,t}q_{22,t}} \\ q_{12,t}/\sqrt{q_{11,t}q_{22,t}} & 1 \end{bmatrix}.$$

The covariance matrix S_t is then

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \begin{bmatrix} 1 & r_{12,t} \\ r_{12,t} & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix},$$

Empirical Example 20.21 (*DCC estimation, daily equity returns*) Figure 20.10 suggests that the DCC correlations look similar to what the crude EWMA method (20.17) delivers. Table 20.2 compares the ability of the CCC and DCC models to capture the covariance. It first constructs a portfolio return of two assets and then estimates an eGARCH model and calculates a time series of standard deviations. This is regressed on the results from a CCC model (using the variance-covariance matrix to calculate the implied standard deviation of the portfolio) and then also from a DCC model. The results suggest that a DCC model is considerably better than the CCC model: the slope is close to 1, the intercept close to 0 and the R^2 is very high.

	(1)	(2)
σ (CCC)	0.57 (23.66)	
σ (DCC)		0.95 (82.67)
constant	0.30 (15.43)	0.02 (2.88)
R^2	0.51	0.87
obs	8862	8862

Table 20.2: Regression of the daily return volatility (eGARCH(1,1)) of a long-short portfolio (long FTSE, short DAX) on the implied result from a CCC or DCC model. Daily data 1991:01-2024:12.

20.5.3 Estimation of a Multivariate Model*

In principle, it is straightforward to specify the likelihood function of the model and then maximize it with respect to the model parameters. For instance, if the n -vector u_t is iid $N(\mathbf{0}, S_t)$, then the log likelihood function is

$$\ln L = \sum_{t=1}^T \ln L_t, \text{ where } \ln L_t = -n \ln(2\pi)/2 - \ln |S_t|/2 - u_t' S_t^{-1} u_t / 2. \quad (20.28)$$

In practice, the optimization problem can be difficult since there are typically many parameters. At least, good starting values are required. (The last expression in (20.28) is clearly the log pdf of the multivariate $N(\mathbf{0}, S_t)$ distribution.) In practice, a two-step approach is used in many cases.

Remark 20.22 (*Estimation of the DCC model with EWMA*) First, estimate $\sigma_{i,i,t}$ by EWMA. Use this in (20.23)–(20.26) with $\beta = \lambda$ and $\alpha = 1 - \lambda$, where the λ value may differ from that used in the univariate variance equations.

Remark 20.23 (*Estimation of the DCC model with MLE*) The DCC model can be estimated by two-step procedure. First, estimate the univariate (e)GARCH models. Second, use (20.23)–(20.26) in the likelihood function (20.28) to find the optimal (α, β) . The two-step approach is not the same as a full ML, but Engle (2009) argues that it is often very similar.

Empirical Example 20.24 It can be shown that applying an EWMA-DCC approach with $\lambda = \beta$ gives results very similar to those in Example 20.21 (which are based on MLE).

Further Reading

See Hull (2022) 23, Campbell, Lo, and MacKinlay (1997) 12, Franses and van Dijk (2000), Engle (2009) and Pesaran (2015) 18 and 25. For an application to asset pricing, see, for instance, Duffee (2005).

Chapter 21

Risk Measures

21.1 Value at Risk

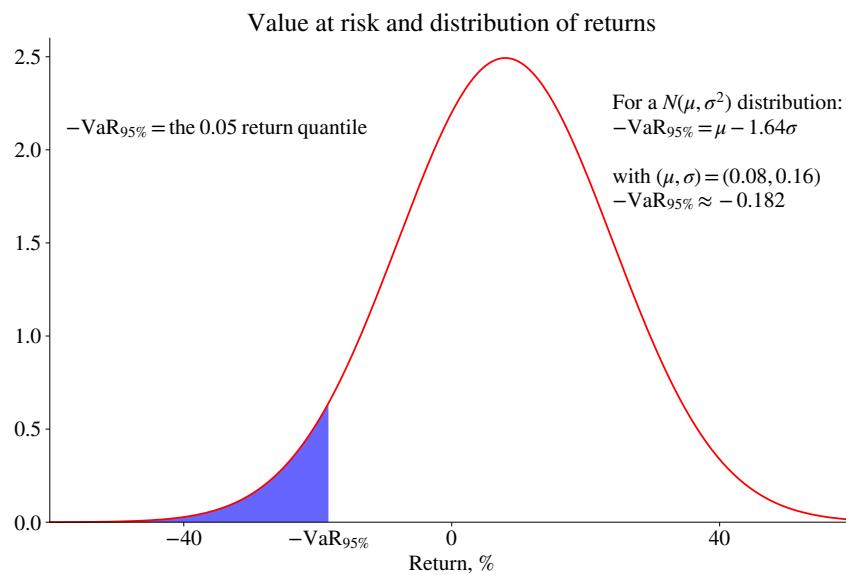


Figure 21.1: Value at risk

The mean-variance framework is often criticized for failing to distinguish between the downside of the return distribution (considered to be risk) and the upside (considered to be potential). The Value at Risk is one way of focusing on the downside.

Remark 21.1 (*Quantile of a distribution*) *The 0.05 quantile is the value such that there is only a 5% probability of a lower number, $\Pr(R \leq \text{quantile}_{0.05}) = 0.05$.*

The 95% Value at Risk ($\text{VaR}_{95\%}$) is a number such that there is only a 5% chance that the loss rate is larger than $\text{VaR}_{95\%}$

$$\Pr(\text{loss rate} \geq \text{VaR}_{95\%}) = 5\%. \quad (21.1)$$

Here, 95% is the confidence level of the VaR. For instance, if $\text{VaR}_{95\%} = 18\%$, then we are 95% sure that we will not lose more than 18% of our investment. To convert the VaR into monetary terms (in CHF, say), just multiply by the value of the investment (portfolio).

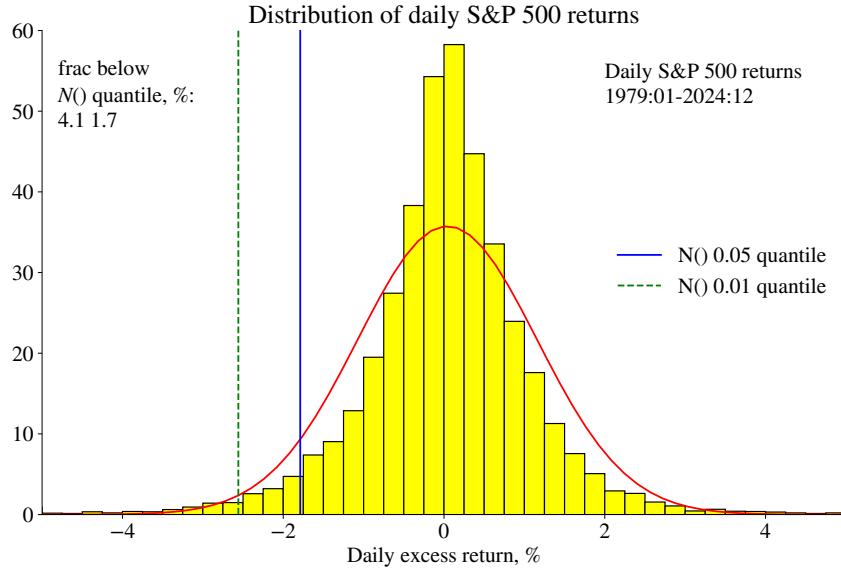


Figure 21.2: Return distribution and VaR for S&P 500

Clearly, the loss rate is the negative of the return, so $-R \geq \text{VaR}_{95\%}$ is true when (and only when) $R \leq -\text{VaR}_{95\%}$, so (21.1) can also be expressed as

$$\Pr(R \leq -\text{VaR}_{95\%}) = 5\%. \quad (21.2)$$

This says that $-\text{VaR}_{95\%}$ is a number such that there is only a 5% chance that the return is below it. That is, $-\text{VaR}_{95\%}$ is the 0.05 quantile (5th percentile) of the return distribution. Using (21.2) allows us to work directly with the return distribution (not the loss distribution), which is often convenient. See Figure 21.1 for an illustration. If the return is normally distributed, $R \sim N(\mu, \sigma^2)$ then

$$\text{VaR}_{95\%} = -(\mu - 1.64\sigma). \quad (21.3)$$

Replace 1.64 by 1.96 for a $\text{VaR}_{97.5\%}$ and 2.33 for a $\text{VaR}_{99\%}$. More generally, we can consider the confidence level α instead of just 0.95, so $\Pr(R \leq -\text{VaR}_\alpha) = 1 - \alpha$, which means that VaR_α is the $-(1 - \alpha)^{\text{th}}$ quantile of R . This is illustrated in Figure 21.3.

Example 21.2 (*VaR with $R \sim N(\mu, \sigma^2)$*) If daily returns have $\mu = 8\%$ and $\sigma = 16\%$, then the 1-day $\text{VaR}_{95\%} = -(0.08 - 1.64 \times 0.16) \approx 0.18$; we are 95% sure that we will not lose more than 18% of the investment over one day, that is, $\text{VaR}_{95\%} = 0.18$. In contrast, $\text{VaR}_{97.5\%} = -(0.08 - 1.96 \times 0.16) \approx 0.24$.

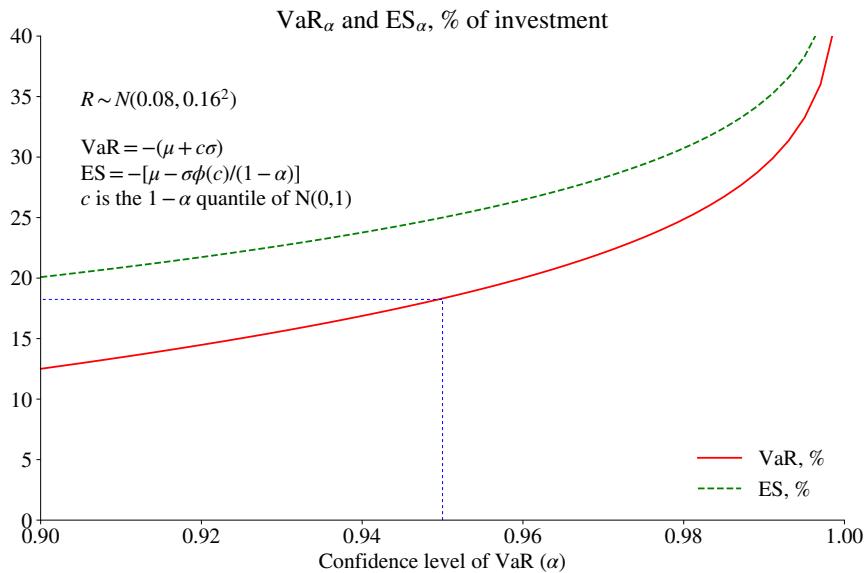


Figure 21.3: Value at risk, different probability levels

Empirical Example 21.3 (*Static VaR for S&P 500 returns*) Figure 21.2 shows the distribution and VaRs for different probability levels for the daily S&P 500 returns. The results indicate that the $N()$ -based model has a reasonable coverage for the 95%, but not for the 99% confidence level.

Example 21.4 (*VaR and regulation of bank capital*) Bank regulations have used 3 times the 99% VaR for 10-day returns as the required bank capital.

Notice that the return distribution depends on the investment horizon, so a VaR is typically calculated for a specific investment period, for instance, one day. Multi-period VaRs are calculated by either explicitly constructing the distribution of multi-period returns, or

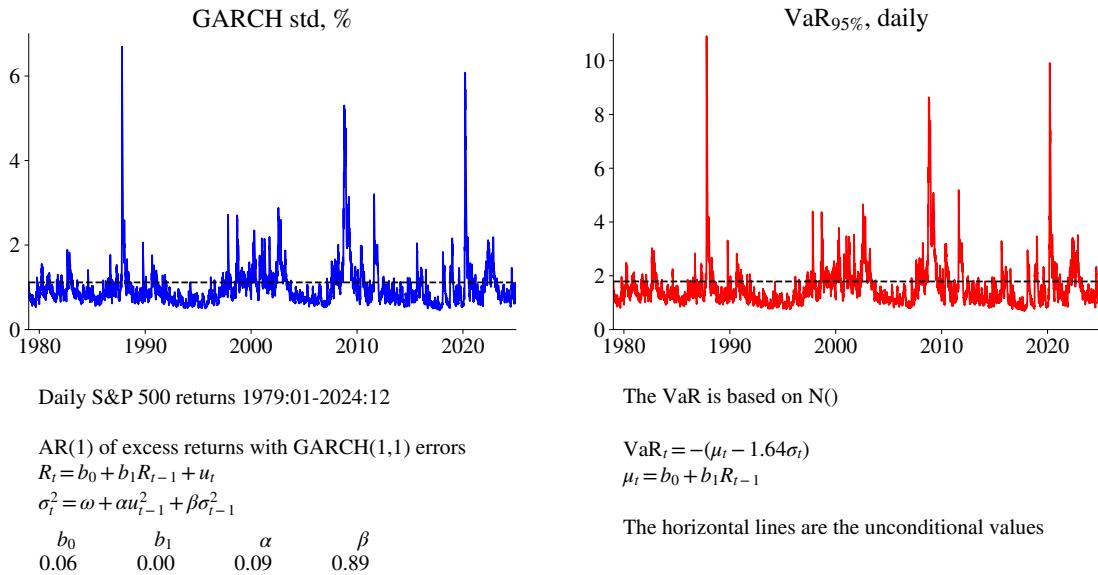


Figure 21.4: Conditional volatility and VaR

by making simplifying assumptions about the relation between returns in different periods, for instance, that they are iid.

Remark 21.5 (*Multi-period VaR*) *If the returns are iid $N(\mu, \sigma^2)$, then a q -period return has the mean $q\mu$ and variance $q\sigma^2$, where μ and σ^2 are the mean and variance of the one-period returns respectively. If the mean is zero, then the q -day VaR is \sqrt{q} times the one-day VaR.*

Empirical Example 21.6 (*Dynamic (time-varying) VaR for S&P 500 returns*) Figure 21.4 shows VaR calculated from $N(\mu_t, \sigma_t^2)$, but where the (μ_t, σ_t) values are allowed to change over time according to an AR(1)+GARCH(1,1) model.

21.2 Backtesting a VaR Model

Backtesting a VaR model amounts to checking if historical data fits with the VaR figures. For instance, we first find the $-\text{VaR}_{95\%}$ and then calculate what fraction of returns that is actually below this number. If the model is correct it should be 5%. We then repeat this for $-\text{VaR}_{96\%}$: only 4% of the returns should be below this number. This approach can be applied on the full sample, or on subsamples defined by time or market conditions.

Empirical Example 21.7 (*Backtesting a static and a dynamic VaR for S&P 500 returns*) See Figure 21.5 for results based on a static VaR model and Figure 21.6 for results from a VaR model where the volatility follows a GARCH process to capture the time varying volatility. The evidence suggests that the latter model, when combined with the assumption that the return is normally distributed with time-varying parameters, performs relatively well except at very high confidence levels.

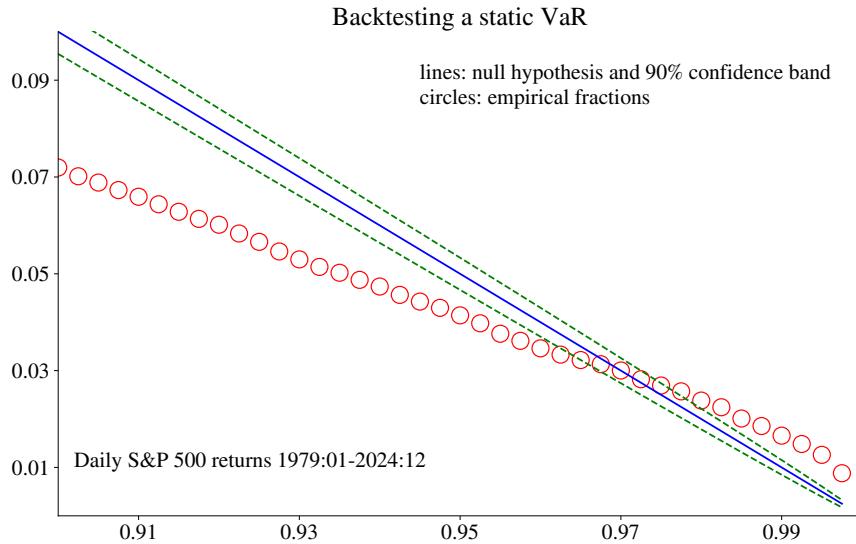


Figure 21.5: Backtesting a static VAR, assuming normally distributed shocks

Empirical Example 21.8 (*Evaluating the VaR model on moving subsamples*) See Figure 21.7 for results based on a static VaR model and Figure 21.8 for results from a dynamic model based on GARCH.

Remark 21.9 (*Bernoulli and binomial distributions*) In a Bernoulli distribution, the random variable X can only take two values: 1 or 0, with probability p and $1 - p$ respectively. This gives $E(X) = p$ and $\text{Var}(X) = p(1 - p)$. After n independent trials, the number of successes (y) has a binomial distribution with $E(y) = np$ and $\text{Var}(y) = np(1 - p)$. For the average outcome, y/n , we then get $E(y/n) = p$ and $\text{Var}(y/n) = p(1 - p)/n$.

To perform a *statistical test* of a VaR model, define a variable that is one if the return is less than the $-\text{VaR}$:

$$d_t = \begin{cases} 1 & \text{if } R_t < -\text{VaR}_\alpha \\ 0 & \text{otherwise} \end{cases} \quad (21.4)$$

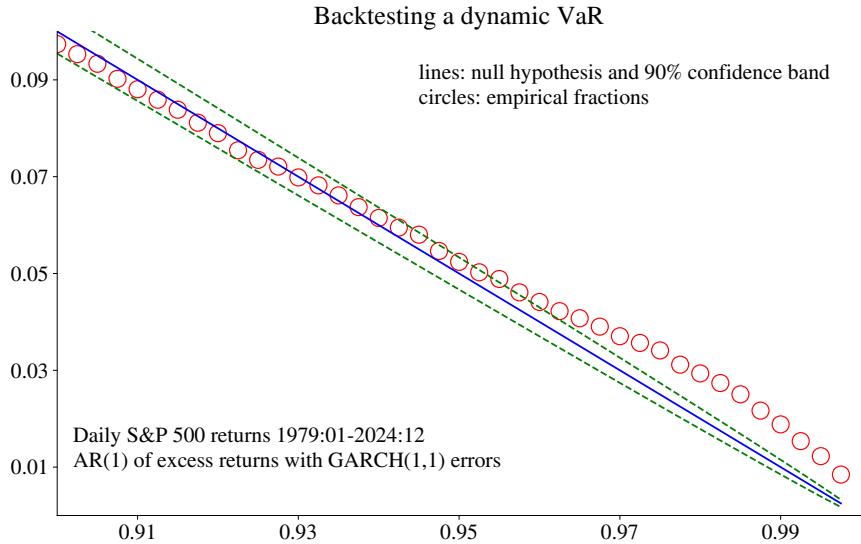


Figure 21.6: Backtesting a dynamic VaR, assuming normally distributed shocks

By using the properties of a binomial distribution, we notice that a sample average of d_t has the following distribution in a large sample

$$\sum_{t=1}^T d_t / T \xrightarrow{d} N(1 - \alpha, \alpha(1 - \alpha) / T). \quad (21.5)$$

This can be used to create a t-stat or a confidence interval.

21.3 Expected Shortfall

The VaR concept has been criticized for having poor aggregation properties. In particular, the VaR for a portfolio is not necessarily (weakly) lower than the portfolio of the VaRs, which contradicts the concept of diversification benefits. (To get this unfortunate property, the return distributions must be heavily skewed.) The *expected shortfall* has better aggregation properties. It also provides a measure of *how large* the loss is, not just whether it will be large or not.

The expected shortfall (also called conditional VaR, average value at risk and expected tail loss) is the expected loss when the return actually is below the VaR_α , that is,

$$\text{ES}_\alpha = -E(R | R \leq -\text{VaR}_\alpha). \quad (21.6)$$

Compare this with the VaR_α , which is the *minimum loss* that will happen with a $1 - \alpha$

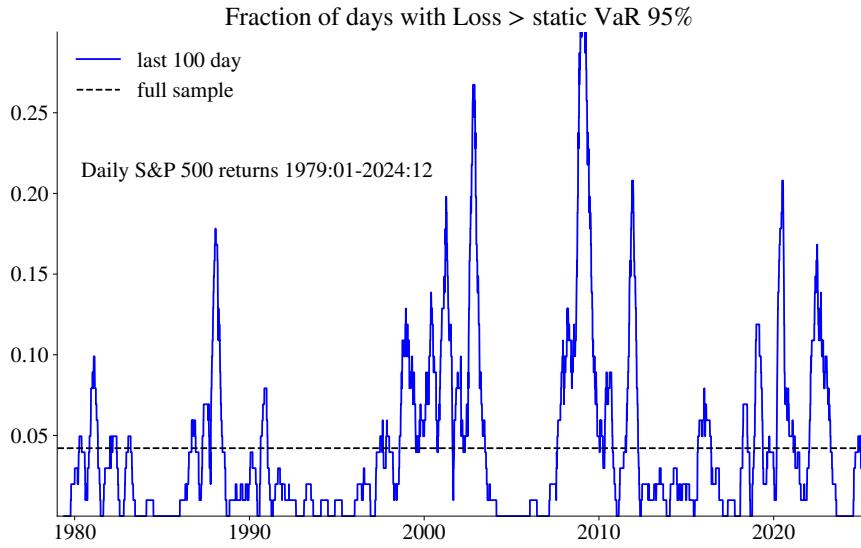


Figure 21.7: Backtesting a static VaR model on a moving data window

probability. See Figure 21.9 for an illustration.

For a normally distributed return $R \sim N(\mu, \sigma^2)$ we have

$$ES_\alpha = -[\mu - \phi(c_{1-\alpha})\sigma/(1-\alpha)], \quad (21.7)$$

where $\phi()$ is the pdf of a $N(0, 1)$ variable and where $c_{1-\alpha}$ is the $1-\alpha$ quantile of a $N(0, 1)$ distribution, for instance, -1.64 for $1-\alpha = 0.05$.

Example 21.10 (ES) If $\mu = 8\%$ and $\sigma = 16\%$, the 95% expected shortfall is $ES_{95\%} = -(0.08 - 2.08 \times 0.16) \approx 0.25$ (since $\phi(-1.64)/0.05 \approx 2.08$) and the 97.5% expected shortfall is $ES_{97.5\%} = -(0.08 - 2.34 \times 0.16) \approx 0.29$ (since $\phi(-1.96)/0.025 \approx 2.34$)

To estimate the average shortfall from a sample (for back-testing, say), calculate the average $-R_t$ for observations where $R_t \leq -\text{VaR}_\alpha$

$$ES_\alpha = -\sum_{t=1}^T \delta_t R_t / (\sum_{t=1}^T \delta_t), \quad (21.8)$$

where $\delta_t = 1$ if $R_t \leq -\text{VaR}_\alpha$ and 0 otherwise.

Empirical Example 21.11 See Figure 21.10 for an empirical example based on S&P 500 returns.

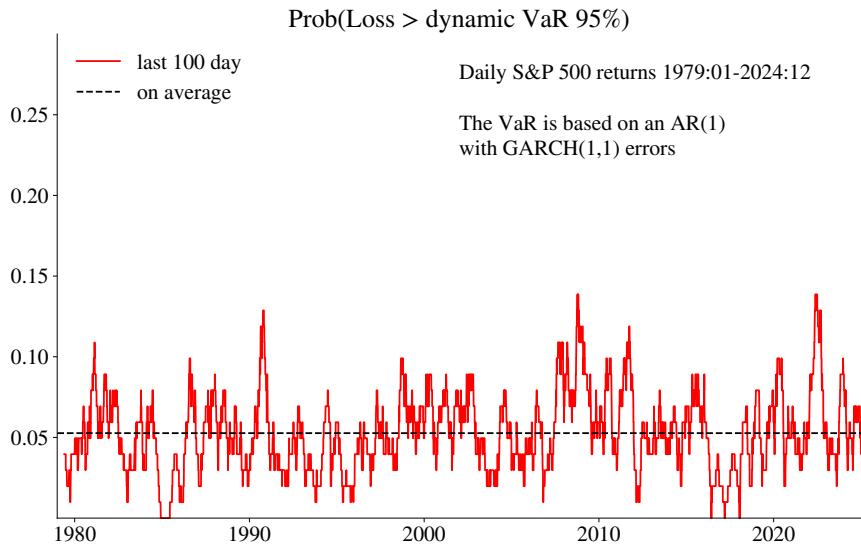


Figure 21.8: Backtesting a dynamic VaR on a moving data window

Empirical Example 21.12 (*Comparison of different downside risk measures for two equity portfolios*) See Table 21.1 for an empirical comparison of the VaR, ES and some more downside risk measures (discussed below). In this case, the ranking of the two portfolios is the same irrespective of measure.

	Small growth	Large value
Std	8.1	5.8
VaR (95%)	12.9	8.8
ES (95%)	17.9	13.1
SemiStd	5.7	3.9
Drawdown	78.4	63.2

Table 21.1: Risk measures of monthly returns of two stock indices (%), US data 1970:01-2024:12.

Proof (of (21.7)) If $x \sim N(\mu, \sigma^2)$, then $E(x|x \leq b) = \mu - \sigma\phi(b_0)/\Phi(b_0)$ where $b_0 = (b - \mu)/\sigma$ and where $\phi()$ and $\Phi()$ are the pdf and cdf of a $N(0, 1)$ variable respectively. To apply this, use $b = -\text{VaR}_\alpha$ so $b_0 = c_{1-\alpha}$. Clearly, $\Phi(c_{1-\alpha}) = 1 - \alpha$ (by definition of the $1 - \alpha$ quantile). Multiply by -1 . \square

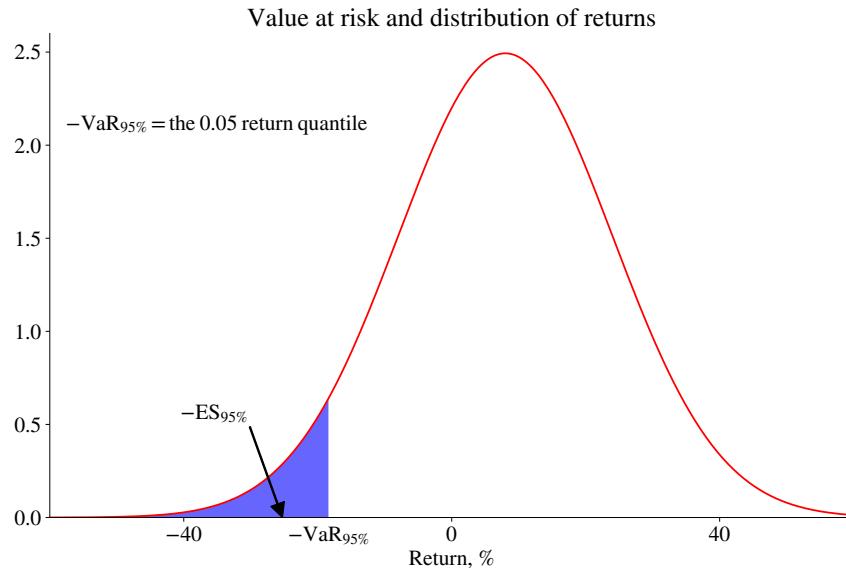


Figure 21.9: Value at risk and expected shortfall

21.4 Semi-Variance and Maximum Drawdown

A semi-variance is defined as

$$\sigma_-^2 = E[\min(R - E R, 0)^2]. \quad (21.9)$$

The square root of σ_-^2 is called the semi-standard deviation. In comparison with a variance, only negative deviations from the mean are given any weight. In some cases, we replace $E R$ by a target level h (this is called a target semi-variance).

To estimate the semi-variance from data (for backtesting) use

$$\sigma_-^2 = \sum_{t=1}^T \delta_t (R_t - \bar{R})^2 / T, \quad (21.10)$$

where $\delta_t = 1$ if $R_t \leq \bar{R}$ and 0 otherwise.

Remark 21.13 (*Alternative scaling of $\sigma_-^2(h)$) Some analysts define $\sigma_-^2(h)$ by just including those observations for which $R_t \leq \bar{R}$. This means multiplying (21.10) by $T / \sum_{t=1}^T \delta_t$, which is actually estimating $E[(R - \bar{R})^2 | R_t \leq \bar{R}]$.

Remark 21.14 (Sortino ratio) The Sortino ratio is an alternative to the Sharpe ratio as a measure of performance. It is $(E R - h) / \sqrt{\sigma_-^2(h)}$.

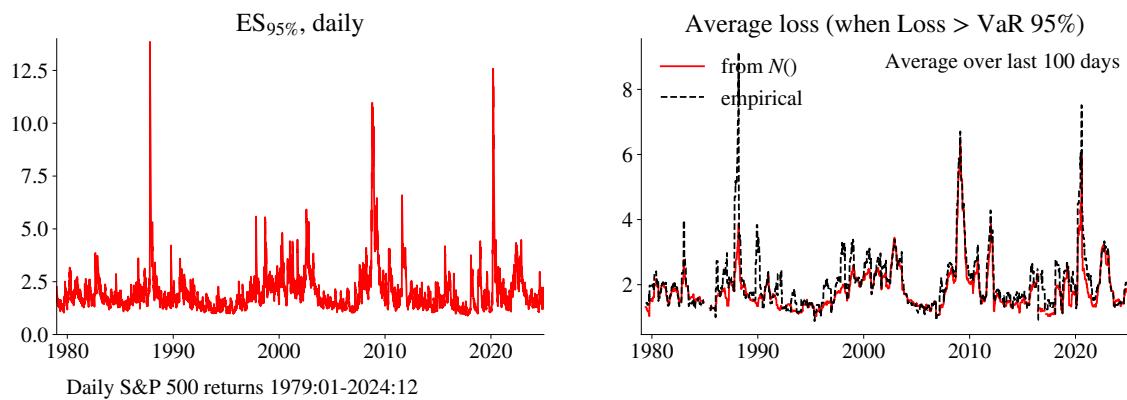


Figure 21.10: Expected shortfall from GARCH model (with backtesting)

An alternative measure is the (percentage) *maximum drawdown* in a given period, for instance, 5 years, say. This is the largest loss from peak to bottom within the period. This is a useful measure when the investor does not know exactly when they have to exit the investment—since it indicates the worst (peak-to-bottom) outcome over the sample. See [21.12](#) for an illustration of the maximum drawdown, in this case over the full sample.

Empirical Example 21.15 (*Max drawdown for different assets*) See [Figure 21.13](#).

Further Reading

See Hull (2022) 22, McDonald (2014) 31, Fabozzi, Focardi, and Kolm (2006) 4–5, McNeil, Frey, and Embrechts (2005), and Alexander (2008).

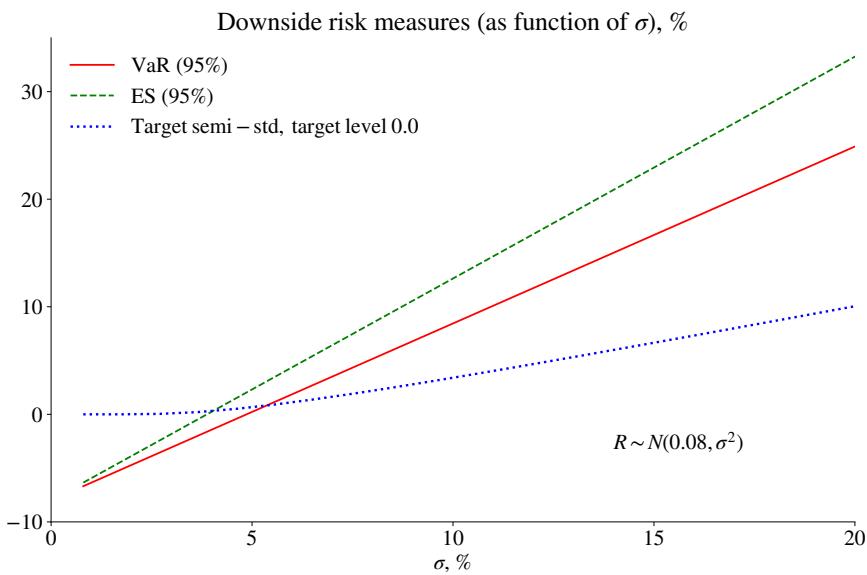


Figure 21.11: Downside risk measures as functions of the standard deviation for a $N(\mu, \sigma^2)$ variable

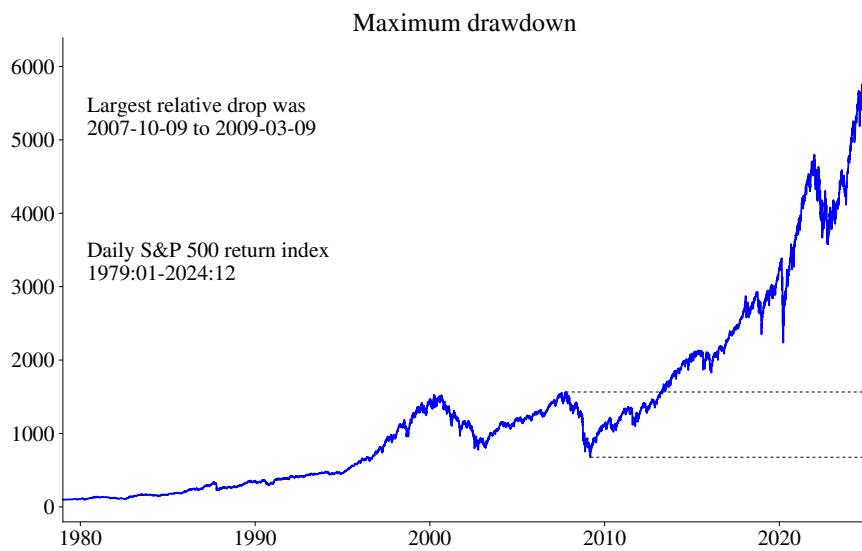


Figure 21.12: Maximum drawdown over the full sample

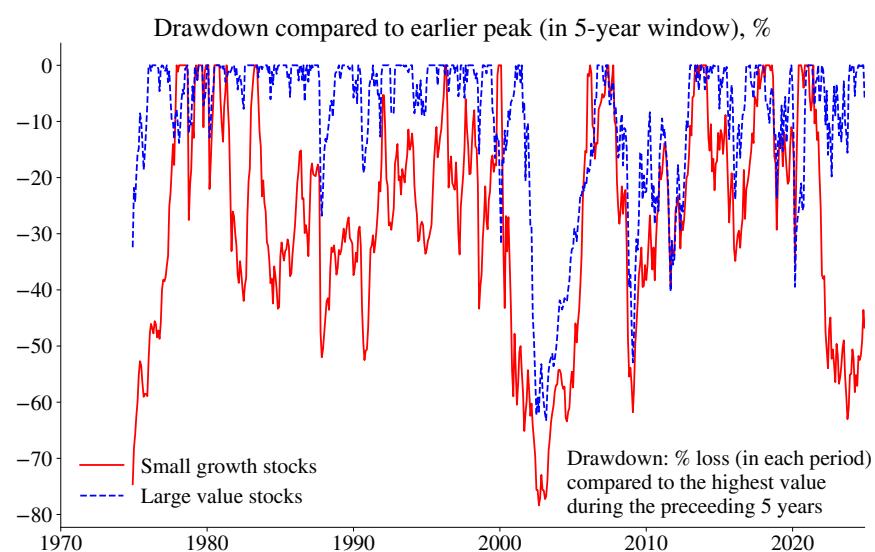


Figure 21.13: Drawdown in moving (rolling) data window

Chapter 22

Non-Parametric Regressions

22.1 Kernel Regressions

22.1.1 Simple Kernel Regression

Non-parametric regressions are used when we are unwilling to impose a parametric form on the regression equation and when a large amount of data is available.

Let the scalars y_t and x_t be related as

$$y_t = b(x_t) + \varepsilon_t, \quad (22.1)$$

where ε_t is iid and $E \varepsilon_t = 0$, $Cov [b(x_t), \varepsilon_t] = 0$ and here $b()$ is an unknown, possibly non-linear, function. In comparison, in a linear regression we have $b(x_t) = \beta x_t$.

One possibility of estimating such a function is to approximate $b(x_t)$ by a polynomial (or some other basis). This approach provides quick estimates; however, the results are “global” in the sense that the value of $b(x)$ at a particular x value ($x = 1.9$, say) will depend on all the data points, and potentially very strongly so.

Non-parametric regressions are more local, allowing only observations close to x (again, $x = 1.9$) to influence the estimate there. For instance, a naive (and bad) approach would be to locate those observations in the sample (say, $t = 3, 27$, and 99) where $x_t = 1.9$ and then let $\hat{b}(1.9) = (y_3 + y_{27} + y_{99})/3$.

Unfortunately, such repeated x_t observations are seldom available and there might be a point in using also nearby x_t values (say, 1.91), provided $b(x)$ is assumed to be smooth. Therefore, we try to approximate the value of $b(x)$ by averaging over observations where

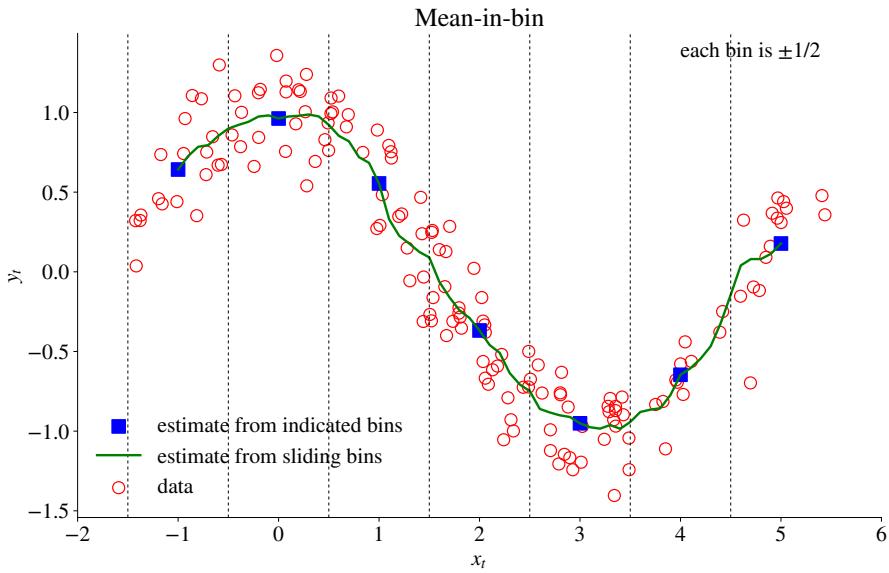


Figure 22.1: Example of a mean-in-bin estimation

x_t is close to x . The general form of this type of estimator is

$$\hat{b}(x) = \frac{\sum_{t=1}^T w(x_t - x)y_t}{\sum_{t=1}^T w(x_t - x)}, \quad (22.2)$$

where $w(x_t - x)/\sum_{t=1}^T w(x_t - x)$ is the weight given to observation t . The function $w(x_t - x)$ is positive and (weakly) decreasing in the distance between x_t and x . Note that the denominator makes the weights sum to unity. The basic assumption behind (22.2) is that the $b(x)$ function is smooth so local averaging (around x) makes sense.

As an example of a $w(\cdot)$ function, consider equal (positive) weights to all values of x_t that are in a certain bin around x and zero weight to all other observations. This is the “mean-in-bin” approach. See Figure 22.1 for an example. Alternatively, use equal weights to the k values of x_t which are closest to x and zero to all other observations (this is the “ k -nearest neighbor” estimator, see Härdle (1990) 3.2). As another example, the weight function could be defined so that it trades off the expected squared errors, $E[y_t - \hat{b}(x)]^2$, and the expected squared acceleration, $E[d^2\hat{b}(x)/dx^2]^2$, which gives a cubic spline.

A *Kernel regression* uses a probability density function (pdf) as the weight function

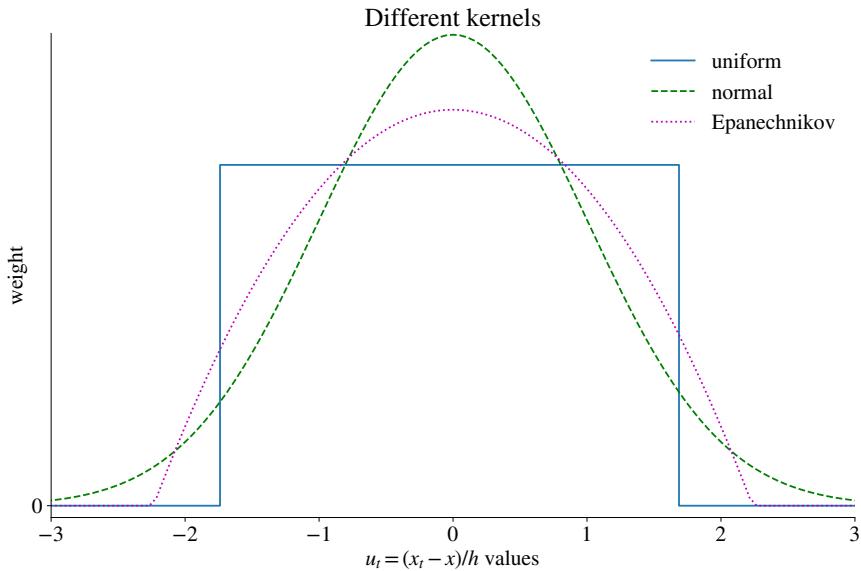


Figure 22.2: Different weighting functions for non-parametric regression

$w(\cdot)$. The perhaps simplest choice is a uniform density function

$$w(u_t) = \begin{cases} 1/(2\sqrt{3}) & \text{for } |u| \leq \sqrt{3} \\ 0 & \text{otherwise,} \end{cases} \quad (22.3)$$

$$\text{with } u_t = (x_t - x)/h. \quad (22.4)$$

See Figure 22.2 for an illustration of the weights. This gives a uniform distribution over $x \pm \sqrt{3}h$. For instance, with $h = 1/(2\sqrt{3})$ we get a flat function over $x \pm 1/2$. (Note that some authors use the convention of a uniform distribution over $x \pm h$ or $x \pm h/2$.)

The reason for the $\sqrt{3}$ terms is that it makes area under the function equal to 1 ($\int w(u)du = 1$) and the variance also equal to one ($\int w(u)u^2du = 1$). This standardisation makes it easy to compare with a $N(0, 1)$ distribution. In any case, we can adjust h to get the intervals we want. The mean-in-bin approach in Figure 22.1 is implemented by using (22.3).

Remark 22.1 (*Interpretation of the pdf in (22.3)*) If $w(u_t)$ is a pdf of the u_t variable, then $w(u_t)/h$ is the pdf of $x_t - x$. Notice that both give the same result in (22.2).

However, we can gain efficiency and get a smoother estimate by using another density function than the uniform. In particular, using a density function that tapers off continuously instead of suddenly dropping to zero improves the properties, for instance, from a

normal distribution. With this kernel, we get the following weights (for estimation of $b(x)$ at point x , for instance, $x = 1.9$)

$$w(u_t) = \phi(u), \text{ with } u_t = (x_t - x)/h \quad (22.5)$$

and where $\phi(u) = \exp(-u^2/2)/\sqrt{2\pi}$, that is, the pdf of an $N(0, 1)$ variable. This weighting function is positive, so all observations receive a positive weight, but the weights are highest for observations close to x and then taper off in a bell-shaped way. See Figure 22.2 for an illustration.

In practice we have to estimate $\hat{b}(x)$ at a finite number of points x . This could, for instance, be 100 evenly spread points in the interval between the minimum and maximum values observed in the sample. See Figure 22.3 for an illustration

Empirical Example 22.2 (*Kernel regression of an AR(1) for equity returns*) See Figure 22.5. The results indicate mean reversion but only after extremely low or high returns.

Example 22.3 Suppose the sample has three data points $[x_1, x_2, x_3] = [1.5, 2, 2.5]$ and $[y_1, y_2, y_3] = [5, 4, 3.5]$. Consider the estimation of $b(x)$ at $x = 1.9$ and suppose we use the gaussian kernel (22.5). With $h = 1$, $u = [-0.4, 0.1, 0.6]$, so the numerator in (22.2) is (skipping the $\sqrt{2\pi}$ term)

$$\begin{aligned} \sum_{t=1}^T w(x_t - x)y_t &= e^{-(0.4)^2/2} \times 5 + e^{-0.1^2/2} \times 4 + e^{-0.6^2/2} \times 3.5 \\ &\approx 0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5 \\ &= 11.52. \end{aligned}$$

The denominator is (again skipping the $\sqrt{2\pi}$ term, since they cancel)

$$\begin{aligned} \sum_{t=1}^T w(x_t - x) &= e^{-(0.4)^2/2} + e^{-0.1^2/2} + e^{-0.6^2/2} \\ &\approx 2.75. \end{aligned}$$

The estimate at $x = 1.9$ is therefore $\hat{b}(1.9) \approx 11.52/2.75 \approx 4.19$.

22.1.2 Choice of Bandwidth (h)

A low value of h in (22.5) means that the weights taper off fast, so the weight function is a normal pdf with a low variance. When $h \rightarrow 0$, then $\hat{b}(x) = y_t$ for $x = x_t$ and zero otherwise, that is, no averaging is performed. In contrast, as $h \rightarrow \infty$, $\hat{b}(x)$ becomes the

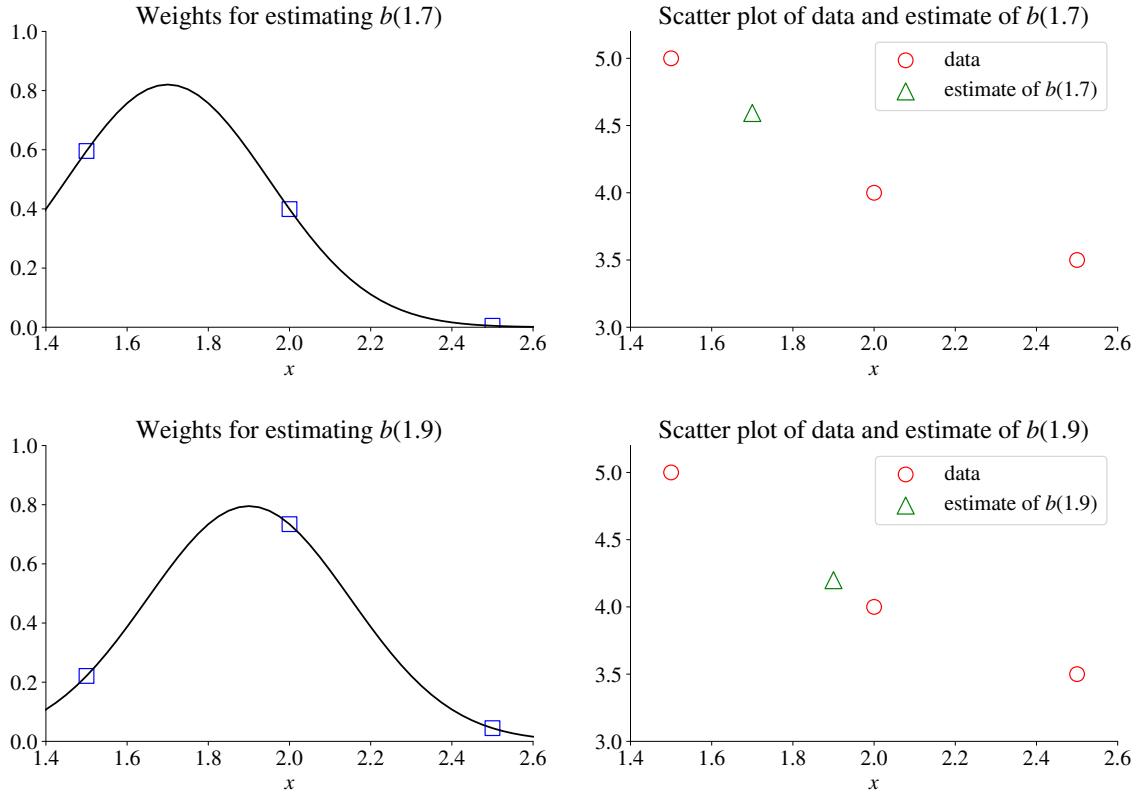


Figure 22.3: Example of kernel regression with three data points

sample average of y_t , so we have global averaging. Clearly, some value of h in between is needed. See Figures 22.4 and 22.5 for illustrations.

Empirical Example 22.4 (*Kernel regression of an AR(1) for equity returns*) Figure 22.5 clearly illustrates the importance of the bandwidth.

A rule of thumb value of h is

$$h = 0.6 [\sigma_\varepsilon^2 (x_{\max} - x_{\min}) / (T \gamma^2)]^{1/5}, \quad (22.6)$$

where γ is a from the regression $y = \alpha + \beta x + \gamma x^2 + \varepsilon$ and σ_ε^2 is an estimate of $\text{Var}(\varepsilon_t)$. In practice, we often replace $x_{\max} - x_{\min}$ by the difference between the 90th and 10th percentiles of x .

A useful, though computationally intensive, approach to choose h is by the leave-one-out *cross-validation* technique. This approach would, for instance, choose h to minimize

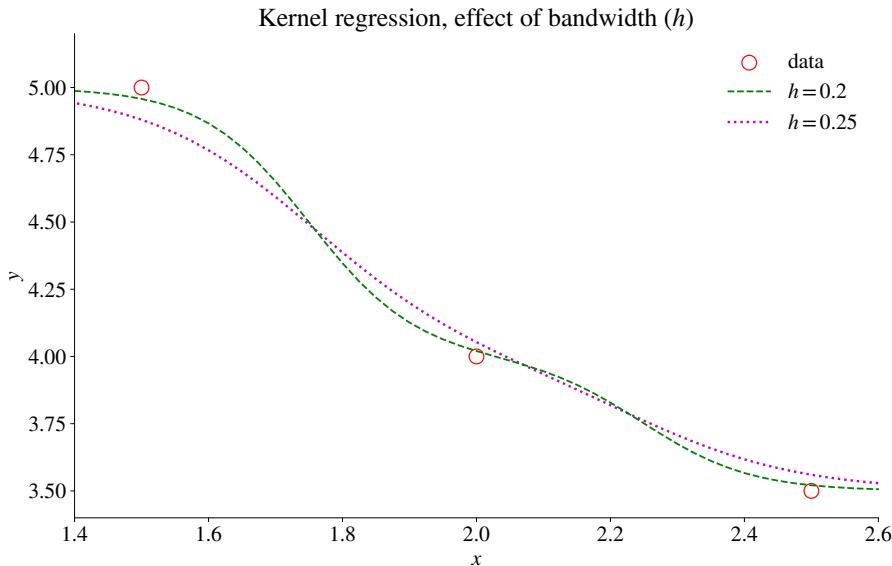


Figure 22.4: Example of kernel regression with three data points

the expected (or average) prediction error

$$\text{EPE}(h) = \sum_{t=1}^T [y_t - \hat{b}_{-t}(x_t, h)]^2 / T, \quad (22.7)$$

where $\hat{b}_{-t}(x_t, h)$ is the fitted value at x_t when we use the bandwidth h , but exclude observation t . This means that each prediction is out-of-sample. To calculate (22.7) we clearly need to make T estimations (for each x_t), and then repeat this for different values of h to find the minimum.

Empirical Example 22.5 (*Kernel regression of an AR(1) for equity returns*) Figure 22.6 shows the EPE for different values of the bandwidth for the kernel regressions previously illustrated in Figure 22.5.

Remark 22.6 (*EPE calculations*) Step 1: pick a value for h

Step 2: estimate the $b(x)$ function on all data, but exclude $t = 1$, then calculate $\hat{b}_{-1}(x_1)$ and the error $y_1 - \hat{b}_{-1}(x_1)$

Step 3: redo step 2, but now exclude $t = 2$, include $t = 1$, and calculate the error $y_2 - \hat{b}_{-2}(x_2)$. Repeat this for $t = 3, 4, \dots, T$. Calculate the EPE as in (22.7).

Step 4: redo steps 1–3, but for another value of h . Keep doing this until you find the best h (the one that gives the lowest EPE)

22.1.3 Implementing the Kernel Regression as Weighted Least Squares

The kernel regression is about finding a weighted mean of y_t . This makes it similar to a weighted least squares (WLS) estimation where the only regressor is a constant. Recall that a WLS estimation minimizes the weighted sum of squared residuals

$$\sum_{t=1}^T w_t (y_t - z'_t b)^2, \quad (22.8)$$

where w_t is the weight for observation t , and the equation allows for z_t to be a K -vector and b a corresponding vector of regression coefficients.

It is straightforward to show that this can be implemented by running a regression on transformed variables

$$\tilde{y}_t = \tilde{z}'_t b + \varepsilon_t, \text{ where } \tilde{z}_t = \sqrt{w_t} z_t \text{ and } \tilde{y}_t = \sqrt{w_t} y_t, \quad (22.9)$$

and the robust (heteroskedasticity consistent) variance-covariance matrix from this regression is a valid approach to estimate $\text{Var}(\hat{b})$.

In the case of a *kernel regression* $z_t = 1$ and $w_t = w(x_t - x)$, and it can be implemented by performing one such regression for each choice of x (typically for a grid of values, as discussed before).

Kernel regressions are typically consistent, provided longer samples are accompanied by smaller values of h , so the weighting function becomes more and more local as the sample size increases.

22.2 Local Linear Regressions

Notice that (22.2) solves the problem $\min_{\alpha_x} \sum_{t=1}^T w(x_t - x)(y_t - \alpha_x)^2$ for each value of x . For a given value of x , α_x is a constant, but it can vary across x values. The first order condition (at a given x value) is $\sum_{t=1}^T w(x_t - x)(y_t - \alpha_x) = 0$, so the solution is as in (22.2), that is, $\hat{\alpha}_x = \hat{b}(x)$. This can be interpreted as a “local constant” regression model: for each x , it is just a constant.

This can be extended to solving a problem like

$$\min_{\alpha_x, \gamma_x} \sum_{t=1}^T w(x_t - x)[y_t - \alpha_x - \beta_x(x_t - x)]^2, \quad (22.10)$$

which defines the *local linear estimator*. (The convention is to use $x_t - x$ as the regressor, but this could easily be changed.) The first order conditions are similar to the usual

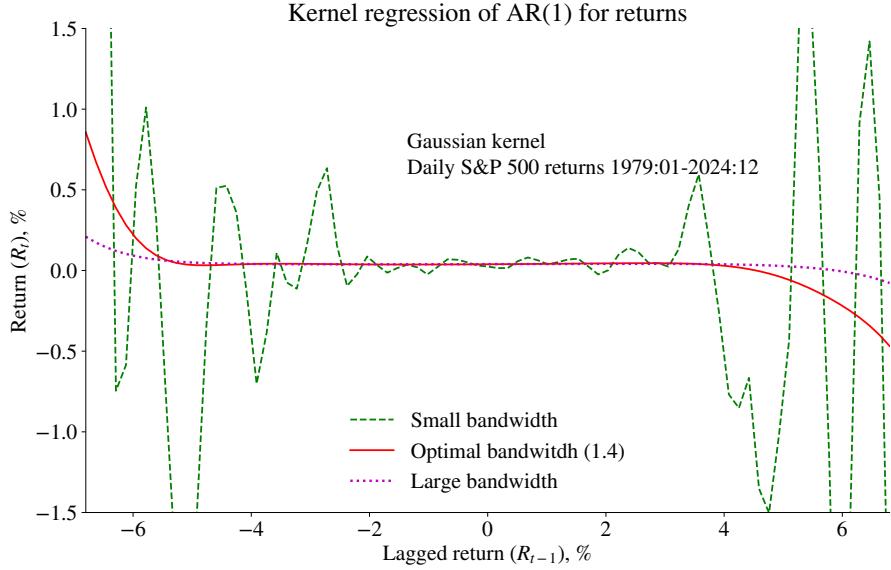


Figure 22.5: Non-parametric regression, importance of bandwidth

normal equations for LS, except that data point t has the weight $w(x_t - x)$ and that we use $x_t - x$ as the regressor. In fact, if we let $z_t = [1, x_t - x]'$ and collect the coefficients in $\theta_x = [\alpha_x, \beta_x]'$, then the first order conditions can be written

$$\sum_{t=1}^T w(x_t - x) z_t y_t = \sum_{t=1}^T w(x_t - x) z_t z_t' \hat{\theta}_x. \quad (22.11)$$

It is straightforward to solve these for $\hat{\theta}_x$.

However, it is easier if we create to run a regression on transformed variables,

$$\tilde{y}_t = \tilde{z}_t' \theta_x + \varepsilon_t, \text{ where } \tilde{z}_t = \sqrt{w(x_t - x)} z_t \text{ and } \tilde{y}_t = \tilde{y}_t = \sqrt{w(x_t - x)} y_t, \quad (22.12)$$

because that gives the same as solving (22.11). That is, the local linear estimator can be implemented by a traditional OLS routine as a regression of \tilde{y}_t on \tilde{z}_t , and this will also give a valid variance-covariance matrix of $\hat{\theta}_x$. (A White's type of covariance matrix might be a good alternative.)

Clearly, solving (22.11) gives one $\hat{\theta}_x$ vector for each x value that we consider. Once we have the estimates, the fitted value at the value x is just $\hat{\alpha}_x$ (since the regression function is $y_t = \alpha_x + \beta_x(x_t - x) + \varepsilon_t$ and we evaluate it at $x_t = x$.) The bandwidth parameter, which appears only in the calculations of the weights, $w(x_t - x)$, can be chosen by the same rule of thumb (22.6) as before, or by a leave-one-out cross validation approach.

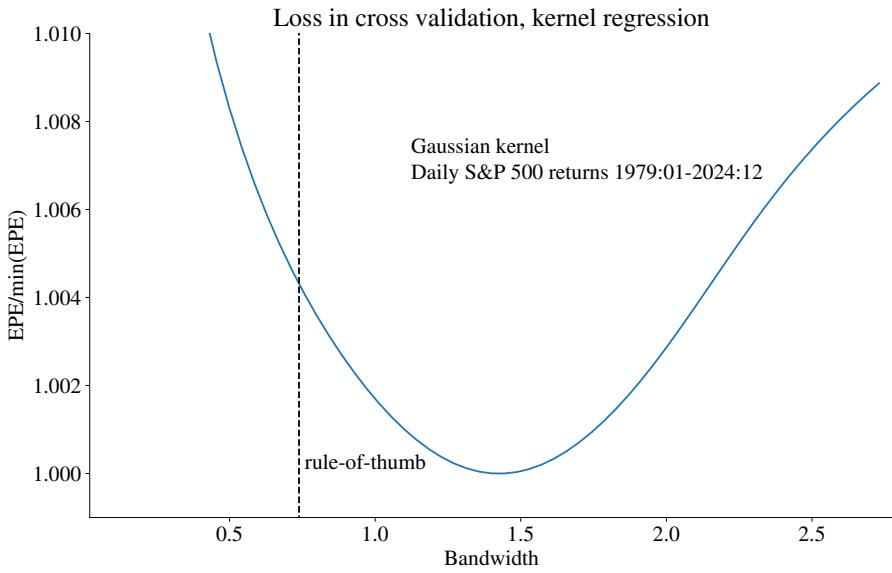


Figure 22.6: Cross-validation

Empirical Example 22.7 (*Local linear regression of an AR(1) for equity returns*) See Figures 22.9 – 22.10.

22.2.1 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \quad \varepsilon_t \text{ is iid and } E \varepsilon_t = 0. \quad (22.13)$$

This makes the estimation problem more data demanding. To illustrate this, suppose a uniform density function is used as weighting function (see in (22.3)). However, with two regressors, the interval becomes a rectangle. With as little as a 20 intervals of each of x and z , we get 400 bins, so we need a large sample to have a reasonable number of observations in every bin.

The most common way to implement the kernel regressor is to let

$$\hat{b}(x, z) = \frac{\sum_{t=1}^T w(x_t - x)v(z_t - z)y_t}{\sum_{t=1}^T w(x_t - x)v(z_t - z)}, \quad (22.14)$$

where $w(x_t - x)$ and $v(z_t - z)$ are two kernels like in (22.5) and where we may allow the bandwidth (h) to be different for x_t and z_t (and depend on the variance of x_t and y_t).

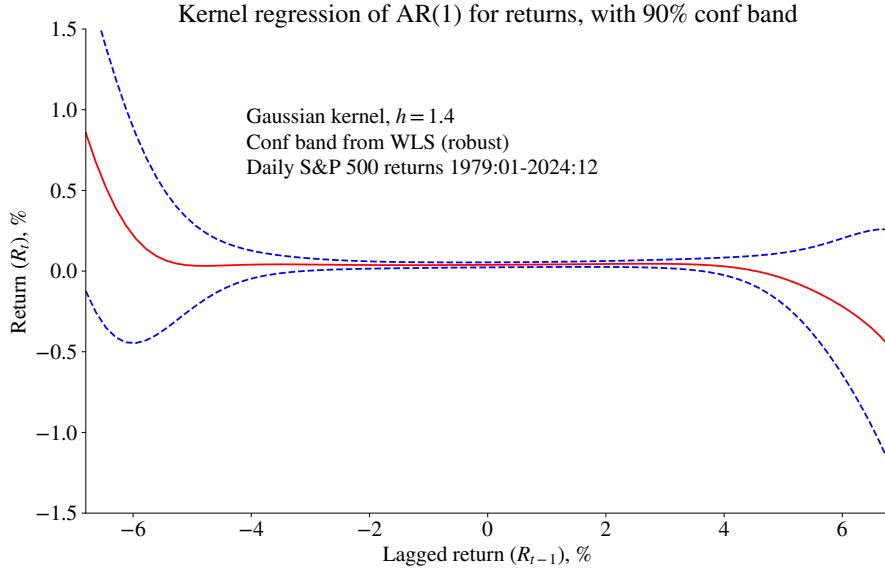


Figure 22.7: Non-parametric regression with confidence bands

In this case, the weight of the observation (x_t, z_t) is proportional to $w(x_t - x)v(z_t - z)$, which is high if both x_t and z_t are close to x and z respectively.

Empirical Example 22.8 (*Kernel regression of an AR(2) for equity returns*) See Figure 22.11.

Empirical Example 22.9 (*Interest rate models*) Interest rate models are typically designed to describe the movements of the entire yield curve in terms of a small number of factors. For instance, the Vasicek model assumes that the (demeaned) short interest rate, r_t , is a mean-reverting AR(1) process

$$r_t = \rho r_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim N(0, \sigma^2), \text{ so}$$

$$r_t - r_{t-1} = (\rho - 1)r_{t-1} + \varepsilon_t,$$

and that all term premia are constant. This means that the drift is decreasing in the interest rate, but that the volatility is constant. For instance, if $\rho = 0.95$ (a very persistent interest rate), then

$$r_t - r_{t-1} = -0.05r_{t-1} + \varepsilon_t,$$

so the reversion to the mean (here zero) is very slow. (The usual assumption is that the short interest rate follows an Ornstein-Uhlenbeck diffusion process, which implies the

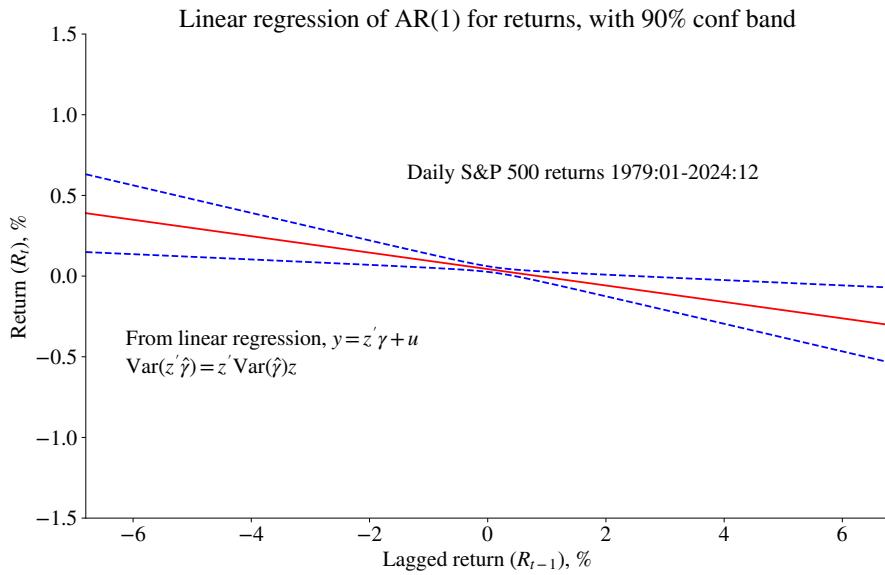


Figure 22.8: Linear regression with confidence bands

discrete time model above.) It can then be shown that all interest rates (for different maturities) are linear functions of short interest rates. To capture more movements in the yield curve, models with richer dynamics are used. For instance, Cox, Ingersoll, and Ross (1985) construct a model which implies that the short interest rate follows an AR(1), except that the variance is proportional to the interest rate level, so $\varepsilon_t \sim N(0, r_{t-1}\sigma^2)$. Non-parametric methods have been used to estimate how the drift and volatility are related to the interest rate level (see, for instance, Ait-Sahalia (1996)). (Kernel regression of a short interest rate process) Figure 22.12 gives an example. Note that the volatility is defined as the square of the drift minus expected drift (from the same estimation method).

Further Reading

See Campbell, Lo, and MacKinlay (1997) 12.3, Härdle (1990), Pagan and Ullah (1999), Mittelhammer, Judge, and Miller (2000) 21, and Hansen (2022a) 19.

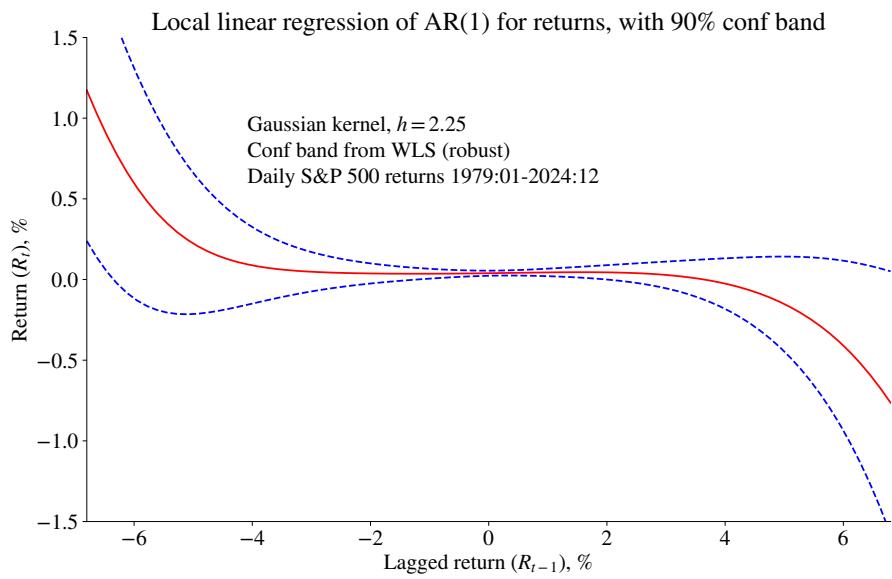


Figure 22.9: Non-parametric local linear regression with confidence bands

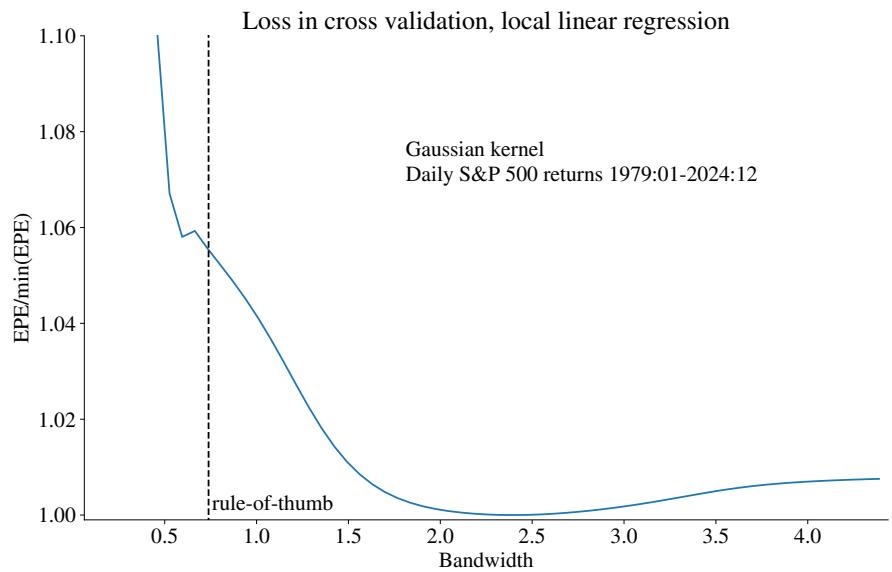


Figure 22.10: Cross-validation

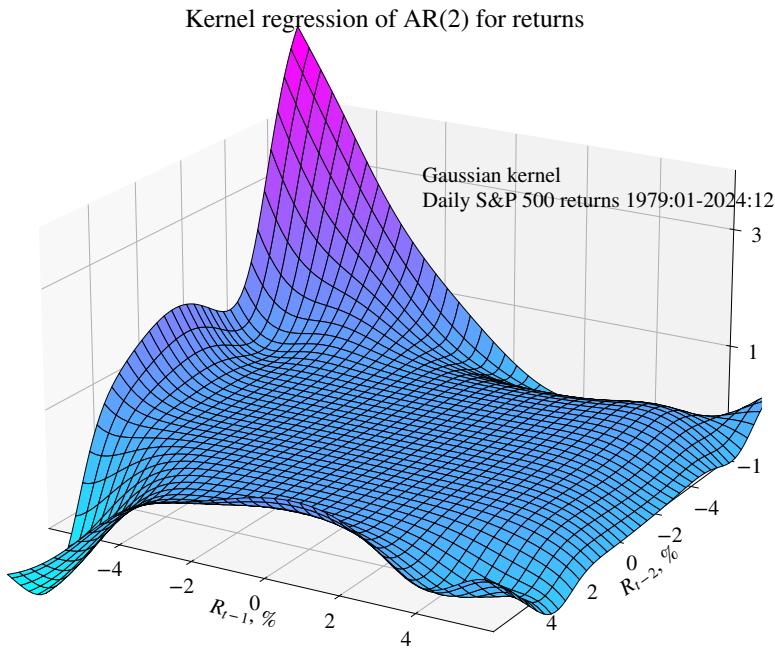


Figure 22.11: Non-parametric regression with two regressors

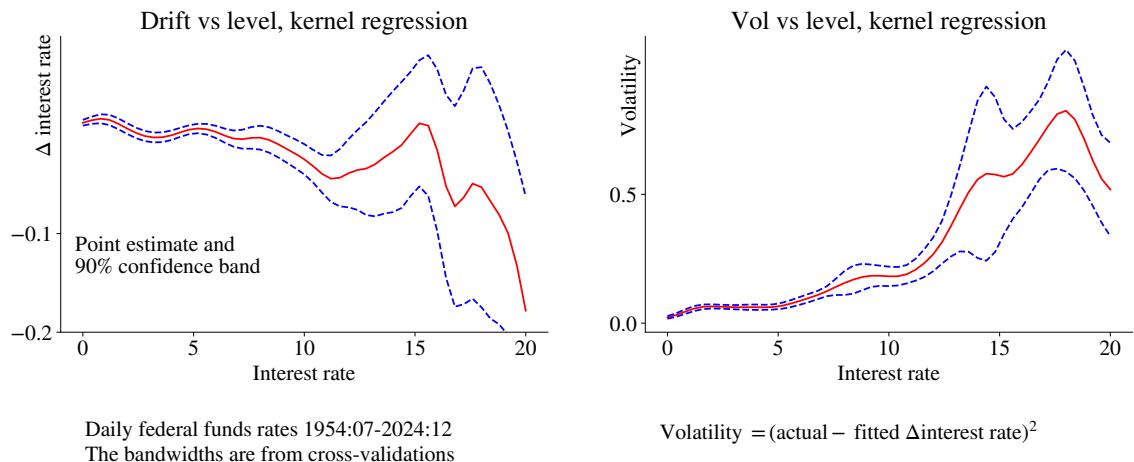


Figure 22.12: Kernel regression, confidence band

Chapter 23

LAD and Quantile Regressions

23.1 LAD

The least absolute deviations (LAD) estimator ($\hat{\beta}_{LAD}$) minimizes the sum of absolute residuals (rather than the squared residuals)

$$\min_b \sum_{t=1}^T |y_t - x'_t b| \quad (23.1)$$

The optimization is a non-linear problem, but a simple iterative method works well (see below). The estimator is typically less sensitive to outliers than OLS. (There are also other ways to estimate robust regression coefficients.) This point is illustrated in Figure 23.1. We can interpret the LAD estimator as an alternative way of getting good estimates of β , especially when the error distribution has fat tails. In fact, when the errors have a Laplace distribution, $f(u) = \exp(-|u|/\sigma)/(2\sigma)$, then LAD is the MLE.

Empirical Example 23.1 (LAD betas for industry portfolios) See Figure 23.2.

If we assume that the median of the true residual (u_t) is zero, then (under strict assumptions, discussed below) we have

$$\begin{aligned} \sqrt{T}(\hat{\beta}_{LAD} - \beta) &\rightarrow^d N[0, f(0)^{-2} \Sigma_{xx}^{-1}/4], \text{ where} \\ \Sigma_{xx} &= \text{plim } \Sigma_{t=1}^T x_t x'_t / T, \end{aligned} \quad (23.2)$$

where $f(0)$ is the value of the pdf of the residual at zero. Unless we know this density function, we need to estimate it, for instance, with a kernel density method. However, to arrive at the result in (23.2) we must assume that the residual is independent of the regressors. (This is discussed in some detail below, see quantile regressions).

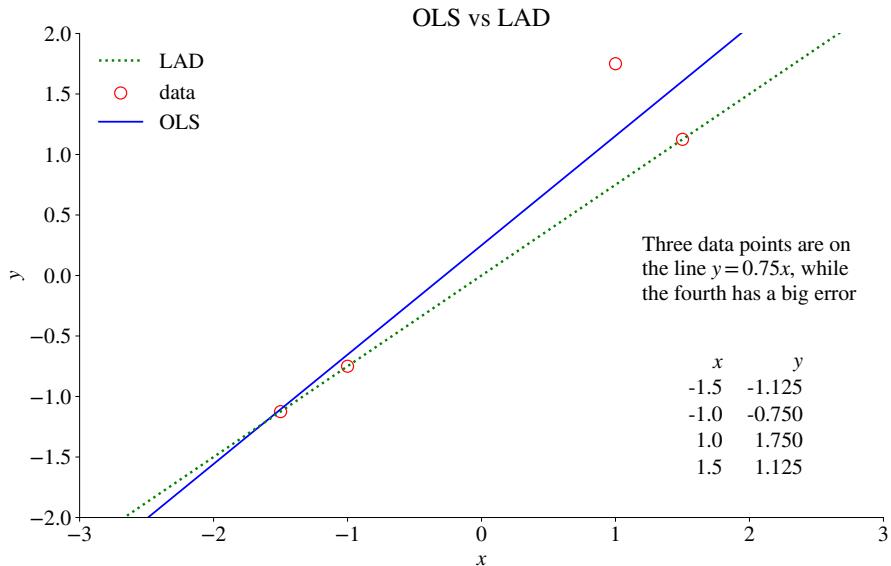


Figure 23.1: Data and regression line from OLS and LAD

Example 23.2 ((23.2) when $u_t \sim N(0, \sigma^2)$) When $u_t \sim N(0, \sigma^2)$, then $f(0) = 1/\sqrt{2\pi\sigma^2}$, so the covariance matrix in (23.2) becomes $\pi\sigma^2\Sigma_{xx}^{-1}/2$. This is $\pi/2$ times larger than when using OLS.

Remark 23.3 (Estimation of $f(0)$) The $f(0)$ value can be estimated as $\hat{f}(0) = \sum_{t=1}^T \phi(u_t/h)/(Th)$, where $\phi()$ is the pdf of an $N(0, 1)$ variable.

Remark 23.4 (Algorithm for LAD*) The LAD estimator solves

$$\min_b \sum_{t=1}^T w_t \hat{u}_t^2, \text{ where } w_t = 1/|\hat{u}_t| \text{ and } \hat{u}_t = y_t - x_t' b,$$

so it is a weighted least squares where both y_t and x_t are multiplied by $1/|\hat{u}_t|^{0.5}$. It can be shown that iterating on LS with the weights given by $1/|\hat{u}_t|^{0.5}$, where the residuals are from the previous iteration, converges very quickly to the LAD estimator.

23.1.1 Reinterpreting the LAD

Consider a linear regression

$$y_t = x_t' \beta + u_t. \quad (23.3)$$

Recall that if u is a random variable, then (typically) the mean, μ , is the solution to $\min_\mu E(u - \mu)^2$, while the median, m , is the solution to $\min_m E|u - m|$. This suggests

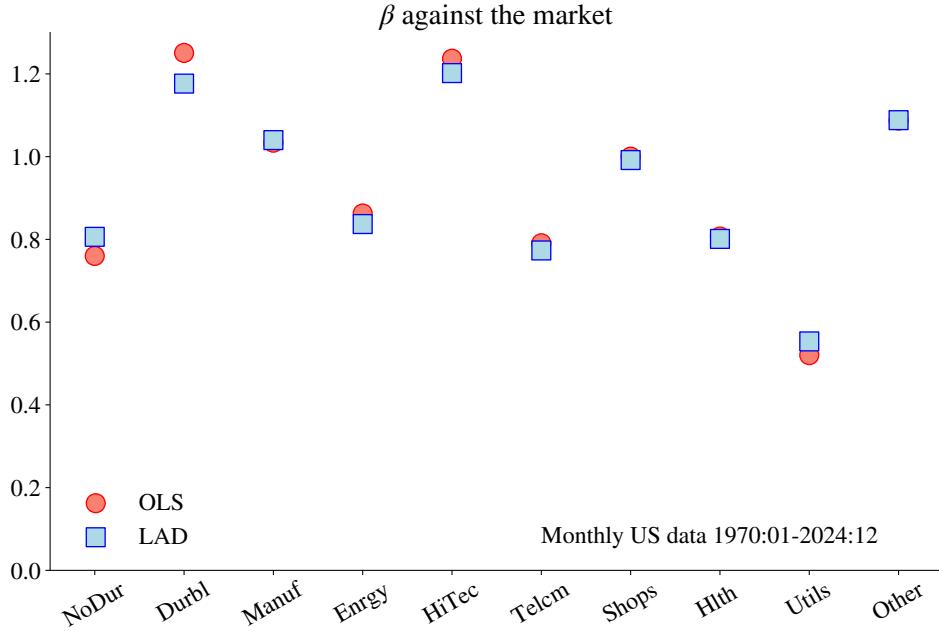


Figure 23.2: Betas of US industry portfolios

that the LAD estimator (23.1) amounts to finding $\hat{\beta}$ coefficients (in a linear model) such that

$$\text{median}(\hat{u}_t | x_t) = 0, \text{ which implies } \text{median}(y_t | x_t) = x_t' \hat{\beta}. \quad (23.4)$$

This is an alternative interpretation of the LAD: it tries to set the median of the fitted residuals, at a given x_t vector, equal to zero. In contrast, OLS tries to set the mean of the fitted residuals, at a given x_t vector, to zero.

23.2 Quantile Regressions

A quantile regression is a generalization of the LAD. Instead of focusing on the 0.5th quantile (the median), as is done in (23.4), it rather states that the q th quantile (conditional on x_t) of the residual is zero

$$Q(\hat{u}_t | x_t; q) = 0, \text{ which implies } Q(y_t | x_t; q) = x_t' \hat{\beta}^{(q)}. \quad (23.5)$$

Here $Q(\hat{u}_t | x_t; q)$ denotes the q th quantile of \hat{u}_t at a particular value of x_t and we also index the estimated coefficients $\hat{\beta}^{(q)}$ to remember that this refers to the q th quantile. We use $\beta^{(q)}$ to denote the true value. Clearly, the LAD is the special case when $q = 0.5$.

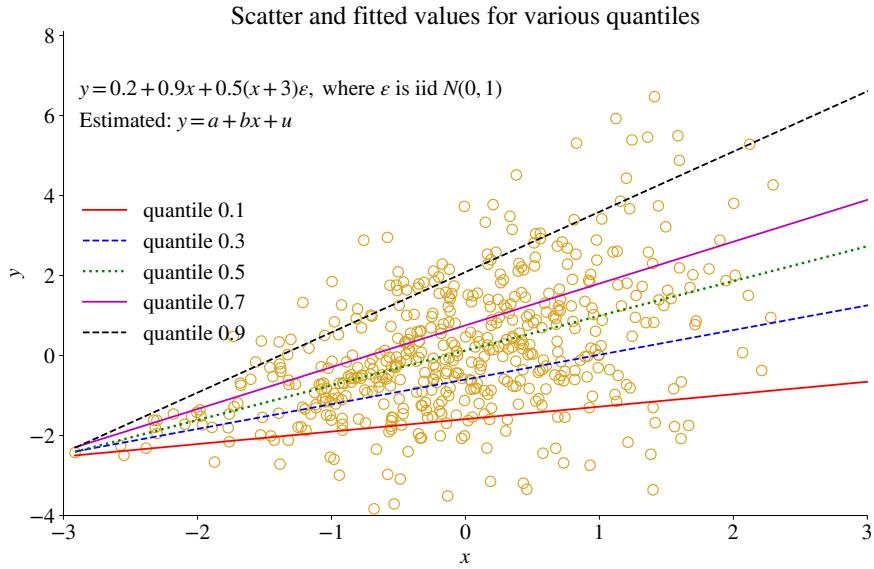


Figure 23.3: Example of quantile regression

We could estimate (see below for how) such coefficients for various quantiles. When x_t just contains a constant and one more regressor, then it is easy to illustrate. See Figure 23.3 for an example in which the slopes differ across quantiles, and Figure 23.4 for one in which they do not. In particular, these figures are based on a *location and scale model*

$$y_t = x_t' \theta + u_t \text{ where } u_t = x_t' \gamma \varepsilon_t, \text{ and } \varepsilon_t \text{ is iid.} \quad (23.6)$$

This is basically a linear model ($y_t = x_t' \theta$), but where the residuals (u_t) are heteroskedastic. In particular, the volatility of u_t is increasing in $|x_t' \gamma|$. This highlights the key feature of quantile regressions: they are well suited for showing how both the typical (median) and tails (for instance, the 0.1th and 0.9th quantiles) are related to the regressors. Notice, however, that we are always referring to *conditional quantiles*, that is, to quantiles of y_t at a particular value of x_t . We are *not* referring to unconditional quantiles of y_t . This means that the slopes for a high quantile, for instance, 0.9, do not necessarily describe the relation between y_t and x_t at generally (unconditionally) high y_t (or x_t) values, see Figure 23.3. Rather, the slopes describe the relationship between y_t and x_t at high ε_t values.

This is perhaps most clearly illustrated using the location and scale model (23.6).

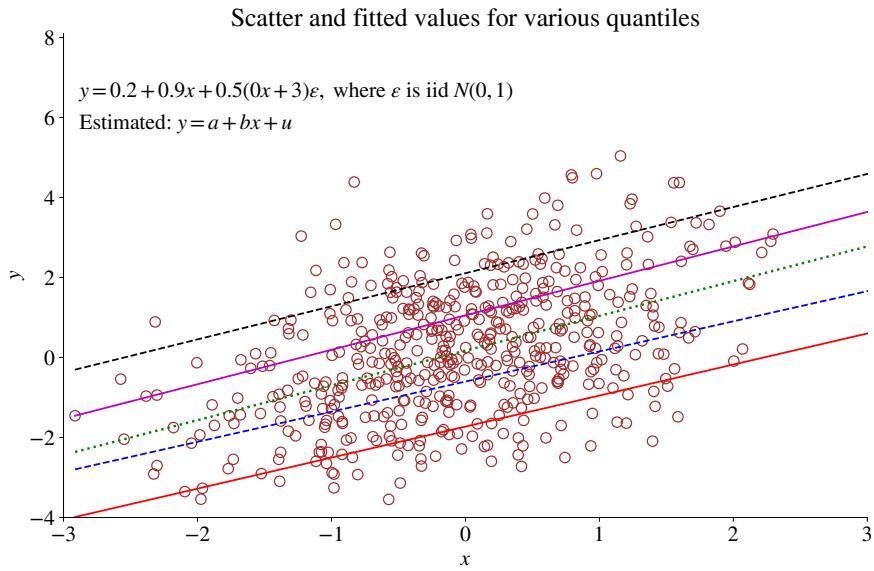


Figure 23.4: Example of quantile regression

Using the properties of the true model (not estimates), we have

$$Q(y_t|x_t; q) = x_t'\theta + Q(u_t|x_t; q) \quad (23.7)$$

$$= x_t'[\theta + \gamma Q(\varepsilon_t; q)]. \quad (23.8)$$

(In the second line, $Q(\varepsilon_t; q)$ need not be conditioned on x_t , if we assume that ε_t is independent of x_t .) Comparing with (23.5) shows that

$$\beta^{(q)} = \theta + \gamma Q(\varepsilon_t; q). \quad (23.9)$$

For instance, if $\gamma > 0$, then $\beta^{(q)}$ is increasing with q , since $Q(\varepsilon_t; q)$ is. This is the case illustrated in Figure 23.3, where the higher slopes at elevated quantiles effectively capture heteroskedasticity. In contrast, Figure 23.4 shows the case where the γ coefficient on the non-constant regressors are all zero: the $\beta^{(q)}$ coefficients (except for the constants) are the same across quantiles.

Empirical Example 23.5 (*Quantile regressions of an AR(1) for S&P 500 returns*) Figure 23.5 illustrates these points by showing the predicted quantiles of a return as a function of the lagged return. The empirical evidence suggests that the typical (median) effect of a lagged return on today's return is almost zero (there is a weak pattern of negative return to be followed by positive returns and vice versa). More pronounced is the smaller

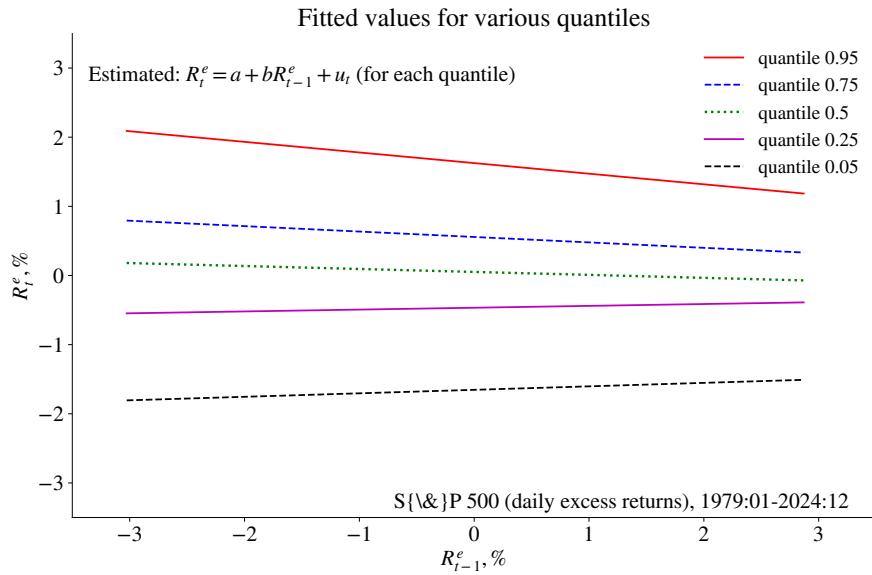


Figure 23.5: Quantile regression of AR(1) of daily returns

dispersion of returns after positive returns—and this is where the real payoff of the quantile regressions is.

The estimated coefficients for the q th quantile, $\hat{\beta}^{(q)}$, solves

$$\min_b \sum_{t=1}^T \delta(\hat{u}_t \geq 0) q |\hat{u}_t| + [1 - \delta(\hat{u}_t \geq 0)] (1 - q) |\hat{u}_t|, \text{ where} \quad (23.10)$$

$$\hat{u}_t = y_t - x'_t b,$$

and where $\delta(z) = 1$ if z is true and 0 otherwise. This is a highly non-linear problem, and the objective function does not have continuous derivatives. However, it can be solved using a linear programming method, a derivative-free minimization algorithm, or an iterative (LS) approach. As a special case, $q = 0.5$ gives the LAD where (23.10) becomes

$$\min_b 0.5 \sum_{t=1}^T |u_t|, \text{ where } u_t = y_t - x'_t b, \quad (23.11)$$

which is clearly the same problem as (23.1).

Remark 23.6 (*An iterative algorithm for quantile regressions**) Applying the same idea

as for LAD, we can write (23.10) as

$$\min_b \sum_{t=1}^T w_t \hat{u}_t^2, \text{ where}$$

$$w_t = \delta(\hat{u}_t \geq 0)q / |\hat{u}_t| + [1 - \delta(\hat{u}_t \geq 0)](1 - q) / |\hat{u}_t|,$$

$$\hat{u}_t = y_t - x'_t b.$$

We iterate by using the weights (w_t) from the previous iteration. (By dividing the loss function by $q(1 - q)$, we could also write the weights as $1/[(1 - q)|\hat{u}_t|]$ and $1/[q|\hat{u}_t|]$, which is how some commonly used code sets it up.)

The asymptotic distribution of the estimates gives (see Wooldridge (2010) for details)

$$\sqrt{T}(\hat{\beta}^{(q)} - \beta^{(q)}) \xrightarrow{d} N[0, q(1 - q)C^{-1}\Sigma_{xx}C^{-1}], \text{ where} \quad (23.12)$$

$$\Sigma_{xx} = \text{plim } \sum_{t=1}^T x_t x'_t / T \text{ and}$$

$$C = \text{plim } \sum_{t=1}^T f(0|x_t) x_t x'_t / T,$$

where $f(0|x_t)$ is the value of the pdf of the residual, conditional on the regressor value, at a zero residual.

Several ways of estimating the variance-covariance matrix have been proposed. First, if the residual is independent of the regressor, then $f(0|x_t) = f(0)$ where the latter is the unconditional density of the residual. In this case, the covariance matrix can be written $q(1 - q)f(0)^{-2}\Sigma_{xx}^{-1}$, which gives the result in (23.2) once we set $q = 0.5$. To estimate $f(0)$, apply the usual kernel density estimator (see the chapter on distributions). Second, to allow for a dependence between the residual and regressor (for instance, heteroskedasticity), then one way of obtaining a consistent estimate of C is

$$C = \sum_{t=1}^T w_t x_t x'_t / T, \text{ with} \quad (23.13)$$

$$w_t = \phi(\hat{u}_t/h)/h,$$

where $\phi()$ is the pdf of an $N(0, 1)$ variable and where \hat{u}_t is the fitted residual. The “bandwidth” h could be chosen as $h = 1.06 \text{Std}(\hat{u}_t)T^{-1/5}$, see Silverman (1986). Notice that the product of w_t and $x_t x'_t$ in (23.13) is aimed at capturing the fact that $f(0|x_t)$ and $x_t x'_t$ are related, something that was ruled out in (23.2). Alternatively, w_t could be calculated using the uniform kernel instead, $w_t = \delta(-h/2 \leq \hat{u}_t \leq h/2)/h$, where $\delta(z) = 1$ when z is true.

Further Reading

See Amemiya (1985) 4.6, Greene (2018) 7, Wooldridge (2010) 12.10, and Hansen (2022a) 24.

Chapter 24

Binary Choice and Truncated Models

Reference: Verbeek (2017) 7; Greene (2018) 17.

24.1 Binary Choice Model

24.1.1 Basic Model Setup

A binary choice model is used when the dependent variable y_i has only two values (here denoted 0 and 1) and we want to estimate how the probability of $y_i = 1$ depends on some variables x_i according to

$$\Pr(y_i = 1|x_i) = F(x'_i \beta), \quad (24.1)$$

where $F()$ is some predefined function. The aim is to estimate the β coefficients. As an example, y_i could indicate whether a firm pays dividends or not and x_i could be a vector of firm characteristics.

The choice of the $F()$ function is mostly a matter of convenience. A *probit model* assumes that $F()$ is a standard normal cumulative distribution function. Other choices of $F()$ give the *logit model* ($F()$ is a logistic function), or the *linear probability model* ($F(x'_i \beta) = x'_i \beta$). See Figure 24.1 for an illustration.

Remark 24.1 (*Logistic function*) The logistic function is $F(v) = 1/[1 + \exp(-v)]$. The derivative of $F(v)$ is $(1 - F(v))F(v)$.

The results are often interpreted by looking at the marginal effects. For instance, the marginal effect of changing regressor k is

$$\frac{\partial F(x'_i \beta)}{\partial x_{k,i}} = f(x'_i \beta)\beta_k, \quad (24.2)$$

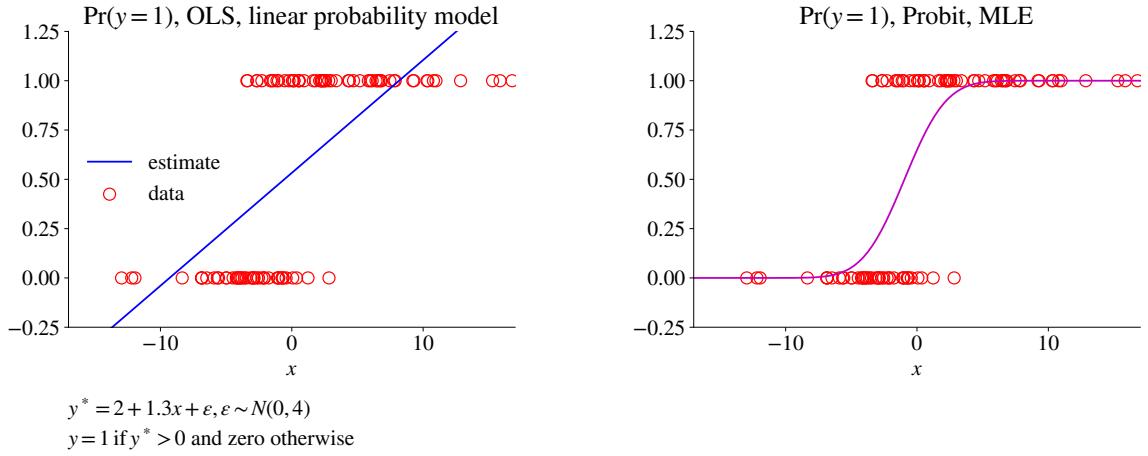


Figure 24.1: Example of probit model

where $f()$ is the derivative of $F()$. (The notation $x_{k,i}$ indicate regressor k for firm i and β_k is the coefficient on regressor k) This is calculated at some typical value x_i (for instance, at the sample average of the regressors). As an example, this could answer the question: how does the probability paying dividends change when profits change? Notice that if the $F()$ function is increasing (all three alternatives mentioned above are), then the derivative (24.2) has the same sign as β_k , since $f() > 0$. In a linear probability model, $f() = 1$ so the marginal effect equals β_k .

Example 24.2 Constant plus two more regressors (w and z): $x'_i\beta = \beta_0 + \beta_1 w_i + \beta_2 z_i$, then $\partial F(x'_i\beta)/\partial w = f(\beta_0 + \beta_1 w_i + \beta_2 z_i)\beta_1$, where $f()$ is the derivative of $F()$. This is often calculated at some typical values of (w_i, z_i) .

Example 24.3 If a regressor is a dummy variable, then we use a simple difference instead of a derivative. For instance, if w_i is either 0 or 1, then we can use $F(\beta_0 + \beta_1 + \beta_2 z_i) - F(\beta_0 + \beta_2 z_i)$. This is calculated at some typical value of z_i . In a linear probability model the result is β_1 .

Notice from (24.2) that the ratio of two coefficients equals the ratio of their marginal effect on the probability

$$\beta_k/\beta_m = \frac{\partial F(x'_i\beta)}{\partial x_{k,i}} / \frac{\partial F(x'_i\beta)}{\partial x_{m,i}}.$$

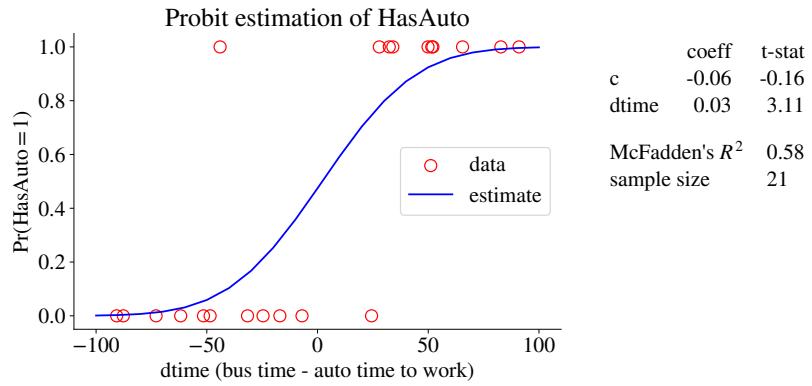


Figure 24.2: Example of probit model, Hill et al (2008), Table 16.1

24.1.2 Estimation

A linear probability model can be estimated by OLS: regress y_i (which is 0 or 1) on a vector of regressors x_i .

In contrast, with a non-linear $F()$, the model is typically estimated with MLE. To do that we need to construct the likelihood function

$$\ln L = \sum_{i=1}^N \ln L_i, \text{ where} \quad (24.3)$$

$$\ln L_i = y_i \ln F(x'_i \beta) + (1 - y_i) \ln[1 - F(x'_i \beta)].$$

This works since y_i is similar to a dummy variable (0 or 1). A similar approach will be used for the models in subsequent sections.

We find the ML estimate by maximizing this log likelihood function with respect to the parameters β . Standard errors are estimates as usual for MLE, based on the information matrix or the more robust sandwich approach. See Figure 24.2 for an empirical example.

Proof (of (24.3)*) Recall that a Bernoulli distribution is specified as $\Pr(y_i = 1) = p_i$, $\Pr(y_i = 0) = 1 - p_i$. Assume independent observations. The probabilities (likelihoods) for the different outcomes are, for instance, $\Pr(y_i = 1 \text{ and } y_j = 1) = p_i p_j$ and $\Pr(y_i = 1 \text{ and } y_j = 0) = p_i(1 - p_j)$. Notice that $p_i^{y_i}(1 - p_i)^{1-y_i} = 1$ if $y_i = 1$, and $1 - p_i$ if $y_i = 0$. For the sample with two data points, the probability can thus be written $L = p_i^{y_i}(1 - p_i)^{1-y_i} \times p_j^{y_j}(1 - p_j)^{1-y_j}$. Take logs and rearrange to get $L = y_i \ln p_i + y_j \ln p_j + (1 - y_i) \ln(1 - p_i) + (1 - y_j) \ln(1 - p_j)$. Substitute $F(x'_i \beta)$ for p_i and extend to N observations for get (24.3). \square

24.1.3 Goodness of Fit

To measure the fit, several different approaches are used, since a traditional R^2 is not appropriate for a non-linear model. *First*, McFadden's R^2

$$\text{McFadden's } R^2 = 1 - \frac{\log \text{likelihood value (at max)}}{\log \text{likelihood value (slope coeffs = 0)}}. \quad (24.4)$$

Notice: $\ln L < 0$ since it is a log of a probability (the likelihood function value), but gets closer to zero as the model improves. McFadden's R^2 (24.4) is therefore between 0 (as bad as a model with only a constant) and 1 (a perfect model).

Example 24.4 If $\ln L = \ln 0.9$ (at max) and the model with only a constant has $\ln L = \ln 0.5$, McFadden's $R^2 = 1 - \ln 0.9 / \ln 0.5 \approx 0.84$. If, instead, the model has $\ln L = \ln 0.8$ (at max), then McFadden's $R^2 = 1 - \ln 0.8 / \ln 0.5 \approx 0.68$.

Second, an alternative measure is an “ R^2 ” for the predicted probabilities

$$\text{“}R_{\text{pred}}^2\text{”} = 1 - \frac{\text{number of incorrect predictions}}{\text{number of incorrect predictions, constant probabilities}}. \quad (24.5)$$

This is somewhat reminiscent of a traditional R^2 since it measures the errors as the number of incorrect predictions, and compare the model with a very static benchmark (here, constant probability). To compare predictions to data, let the predictions be

$$\hat{y}_i = \begin{cases} 1 & \text{if } F(x_i' \hat{\beta}) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (24.6)$$

This says that if the fitted probability $F(x_i' \hat{\beta})$ is higher than 50%, then we define the fitted binary variable to be one, otherwise zero. Then, cross-tabulate (as in a contingency table) the actual (y_i) and predicted (\hat{y}_i) values:

	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Total
$y_i = 0$:	n_{00}	n_{01}	N_0
$y_i = 1$:	n_{10}	n_{11}	N_1
Total:	\hat{N}_0	\hat{N}_1	N

(24.7)

There are $n_{01} + n_{10}$ incorrect predictions to be used in the numerator of (24.5).

For the constant probability in the denominator, we first notice the fraction of data where $y_i = 1$ is

$$\hat{p} = N_1/N. \quad (24.8)$$

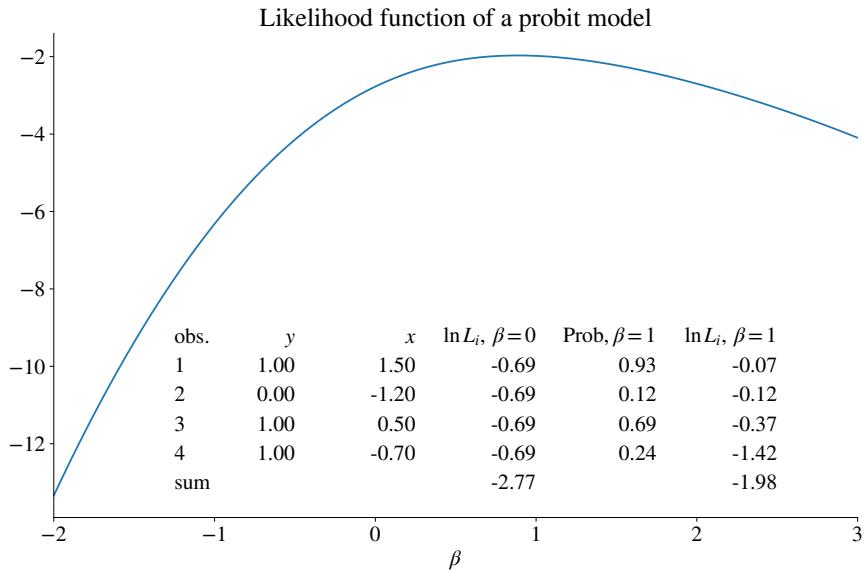


Figure 24.3: Example of ML estimation of probit model

When $\hat{p} \leq 0.5$, the constant probability model always predicts $y_i = 0$, so the number of incorrect predictions is N_1 . Otherwise it is N_0 . This gives the “ R^2_{pred} ” in (24.5) as

$$\text{“}R^2_{pred}\text{”} = \begin{cases} 1 - \frac{n_{01} + n_{10}}{N_1} & \text{if } \hat{p} \leq 0.5 \\ 1 - \frac{n_{01} + n_{10}}{N_0} & \text{if } \hat{p} > 0.5. \end{cases} \quad (24.9)$$

Example 24.5 Consider the data in Figure 24.3. If $\beta = 1$ happened to maximize the likelihood function (it almost does...), then McFadden’s $R^2 = 1 - (-1.98)/(-2.77) \approx 0.29$ and the predicted values would be $\hat{y}_i \approx [1 \ 0 \ 1 \ 0]$. Cross-tabulation of actual (y_i) and predicted (\hat{y}_i) values gives

	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Total
$y_i = 0:$	1	0	1
$y_i = 1:$	1	2	3
<i>Total:</i>	2	2	4

Since the constant probability is $\hat{p} = 3/4$, the constant probability model always predicts $y_i = 1$. We therefore get “ R^2_{pred} ” = $1 - 1/(1+0) = 0$.

24.1.4 Related Models

There are several related models (not treated in detail here), for instance, multi-response models answers questions like “a little, more, most?” (ordered logit or probit) or “Red, blue or yellow car?” (unordered models: multinomial logit or probit). Also, models for count data are useful for answering questions like: “how many visits to the supermarket this week?” They are like the probit model, but y_i can take on a finite number of integer values (0, 1, 2, 3, ...).

24.2 Truncated Regression Model

24.2.1 Basic Model Setup

The truncated regression model deals with the case when data is not observed when the dependent variable is below some threshold. To be precise, suppose the correct model is linear and that the residuals are normally distributed

$$y_i^* = x_i' \beta + \varepsilon_i, \varepsilon_i \sim \text{iidN}(0, \sigma^2). \quad (24.10)$$

The assumption of normally distributed residuals is not necessary. However, changing it requires redefining the likelihood function below.

However, data on both the dependent variable (y_i) and the regressors (x_i) is completely missing if $y_i^* \leq 0$, so

$$\begin{aligned} (y_i, x_i) &= (y_i^*, x_i) && \text{if } y_i^* > 0 \\ (y_i, x_i) &\text{ not observed} && \text{otherwise.} \end{aligned} \quad (24.11)$$

The cutoff at zero is merely a normalization and can be changed without affecting the model’s validity.

The problem with this truncation is that the sample is no longer random. For instance, suppose y_i^* is dividends and x_i is profits, and it so happens that firms with low dividends are not in the sample. This is likely to bias the results. In fact, running OLS to estimate

$$y_i = x_i' \beta + u_i \quad (24.12)$$

on the available data gives biased and inconsistent estimates. See Figure 24.4 for an illustration.

The reason for the bias is that we only use those data points where y_i^* is unusually high

(for a given value of x_i). That is, for observation i to be observed, it must be the case that

$$\varepsilon_i > -x'_i \beta, \quad (24.13)$$

since otherwise (24.12) becomes negative, and those observations do not enter the sample. The ε_i realizations that enter the sample are not random, and they depend on the x_i value: when $x'_i \beta < 0$, then only positive ε_i realizations enter the sample and vice versa. This correlation is a classical reason for why OLS estimates of the regression coefficients are inconsistent.

Remark 24.6 (*Details on the bias**) To see the bias, notice that the expected value of y_i , conditional on x_i and that we observe the data ($y_i > 0$), is $x'_i \beta + E(\varepsilon_i | y_i^* > 0)$ which equals $x'_i \beta + E(\varepsilon_i | \varepsilon_i > -x'_i \beta)$. The second result follows from the fact that $y_i^* > 0$ happens when $x'_i \beta + \varepsilon_i > 0$, that is, when $\varepsilon_i > -x'_i \beta$. Notice that $E(\varepsilon_i | \varepsilon_i > -x'_i \beta)$ is positive (compare with $E(\varepsilon_i | \varepsilon_i > -\infty) = 0$ and notice that $-x'_i \beta > -\infty$) and correlated with x_i . From basic properties of normal distributions, we know that if $\varepsilon \sim N(\mu, \sigma^2)$, then $E(\varepsilon | \varepsilon > a) = \mu + \sigma \phi(a_0) / [1 - \Phi(a_0)]$, where $a_0 = (a - \mu) / \sigma$. Here $a = -x'_i \beta$, so $a_0 = -x'_i \beta / \sigma$.

24.2.2 Estimation

Remark 24.7 (*Pdf of truncated variable*) Let $\text{pdf}(\varepsilon)$ be the density function of ε (without any truncation). The density function, conditional on $a < \varepsilon \leq b$ is $\text{pdf}(\varepsilon | a < \varepsilon \leq b) = \text{pdf}(\varepsilon) / \Pr(a < \varepsilon \leq b)$.

If ε_i is normally distributed, then the log likelihood function is

$$\begin{aligned} \ln L &= \sum_{i=1}^N \ln L_i, \text{ where} \\ \ln L_i &= -\ln \sigma + \ln \phi[(y_i - x'_i \beta) / \sigma] - \ln \Phi(x'_i \beta / \sigma), \end{aligned} \quad (24.14)$$

where $\phi()$ and $\Phi()$ are the pdf and the cdf of an $N(0, 1)$ variable. (Recall that $\phi(z/\sigma)/\sigma$ is the pdf of an $N(0, \sigma^2)$ variable z .) MLE maximizes this with respect to β and σ^2 (or σ). Notice that $\Phi(x'_i \beta / \sigma)$ is the new part compared with OLS. As before, standard errors for MLE can be based in the information matrix or the sandwich approach. See Figure 24.4 for an illustration.

Proof (of (24.14)*) The density function of ε_i , conditional on $y_i^* = x'_i \beta + \varepsilon_i > 0$ (so

$\varepsilon_i > -x'_i \beta$) is

$$\text{pdf}(\varepsilon_i | \varepsilon_i > -x'_i \beta) = \frac{\text{pdf}(\varepsilon_i)}{\Pr(\varepsilon_i > -x'_i \beta)}.$$

The numerator is the pdf of an $N(0, \sigma^2)$ variable, while the denominator equals $\Pr(\varepsilon_i / \sigma > -x'_i \beta / \sigma) = \Phi(x'_i \beta / \sigma)$, where the last equality follows from $\Phi(z) = 1 - \Phi(-z)$. Combining and replacing ε_i by $y_i - x'_i \beta$, noting that $\text{pdf}(\varepsilon_i) = \phi(\varepsilon/\sigma)/\sigma$, and taking logs gives (24.14). \square

24.3 Censored Regression Model

The censored regression model is similar to truncated model, but with the difference that we always observe the regressors x_i . We thus have *more information* than in the truncated case. In short, the model and data are

$$y_i^* = x'_i \beta + \varepsilon_i, \varepsilon_i \sim \text{iidN}(0, \sigma^2) \quad (24.15)$$

$$\text{Data: } (y_i, x_i) = \begin{cases} (y_i^*, x_i) & \text{if } y_i^* > 0 \\ (0, x_i) & \text{otherwise.} \end{cases}$$

Values $y_i^* \leq 0$ are said to be *censored*, and assigned the value 0, which is just a normalization. This is the classical *Tobit model*. Again, the assumption of normally distributed residuals can be relaxed, but requires changing the likelihood function below.

As an example of where this model would make sense, let y_i^* represent investment into stocks by a household, and x_i household income, where households with low income are assigned a common value (normalized to $y_i = 0$) in the survey.

If we estimate

$$y_i = x'_i \beta + u_i \quad (24.16)$$

with OLS, using all data with $y_i > 0$, then we are in same situation as in truncated model: OLS is not consistent. See Figure 24.4.

24.3.1 Estimation of Censored Regression Model

Remark 24.8 (*Likelihood function with different “states”*) The likelihood contribution of observation i is $\text{pdf}(y_i) = \text{pdf}(y_i | \text{state } K) \times \Pr(\text{state } K)$. The total likelihood function is the sum over all i (observations).

Here there are two states: $y_i^* \leq 0$ (no data on y_i , but on x_i) and $y_i^* > 0$ (data on both

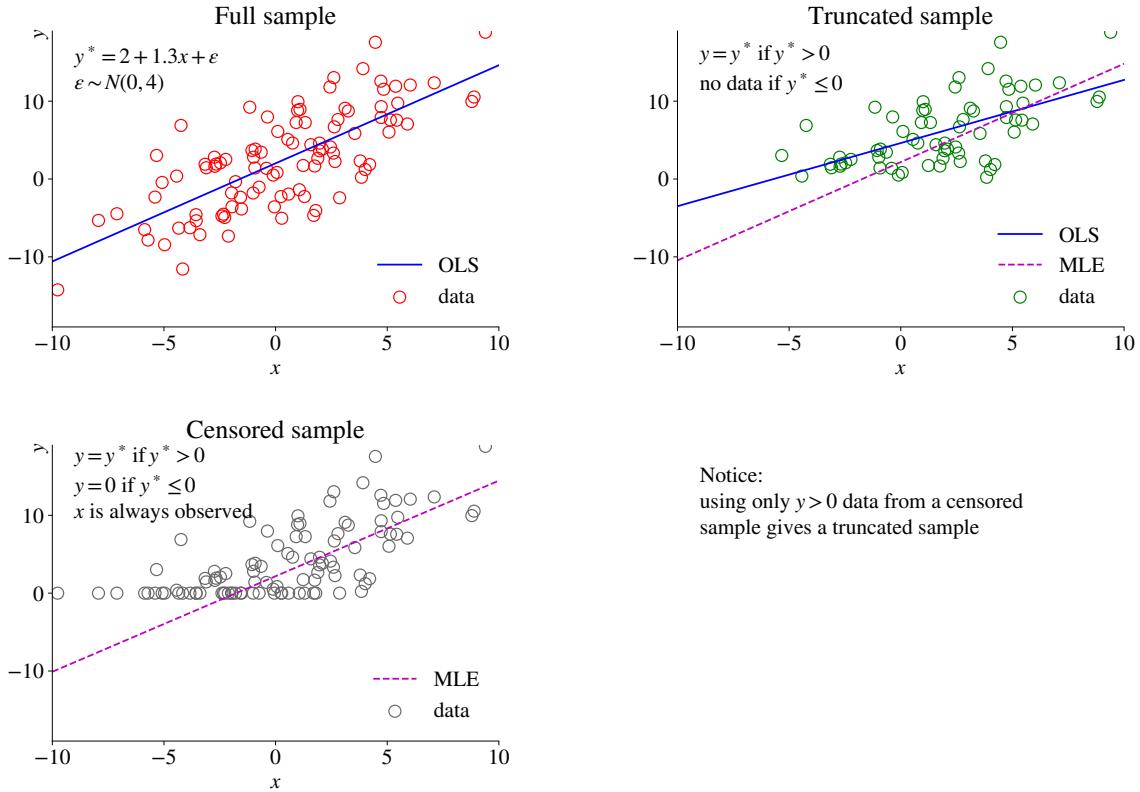


Figure 24.4: Estimation on full, truncated and censored sample

y_i and x_i) This gives the log likelihood function

$$\begin{aligned} \ln L &= \sum_{i=1}^N \ln L_i \text{ where} \\ \ln L_i &= \delta_i \ln \Phi(-x_i' \beta / \sigma) + (1 - \delta_i)(\ln \phi[(y_i - x_i' \beta) / \sigma] - \ln \sigma), \end{aligned} \quad (24.17)$$

where $\delta_i = 1$ if $y_i = 0$ and 0 otherwise. As before, $\phi()$ is the pdf of an $N(0, 1)$ variable, so $\phi(z/\sigma)/\sigma$ is the pdf of an $N(0, \sigma^2)$ variable z . MLE maximizes this with respect to β and σ^2 (or σ). Compared to OLS, the new part is that we have a way of calculating the probability of censored data (first term), since we know all x_i values.

Proof (of (24.17)*) State $y_i^* \leq 0$ happens when $y_i^* = x_i' \beta + \varepsilon_i \leq 0$, that is, when $\varepsilon_i \leq -x_i' \beta$. The probability of this is $\Phi(-x_i' \beta / \sigma)$. The conditional density function in this state has the constant value of one, so the likelihood contribution (see Remark 24.8) is

$$L_i(\text{if } y_i^* \leq 0) = \text{pdf}(y_i | y_i^* \leq 0) \times \Pr(y_i^* \leq 0) = 1 \times \Phi(-x_i' \beta / \sigma).$$

State $y_i^* > 0$ happens in the same way as in the truncated model (24.16), but the difference

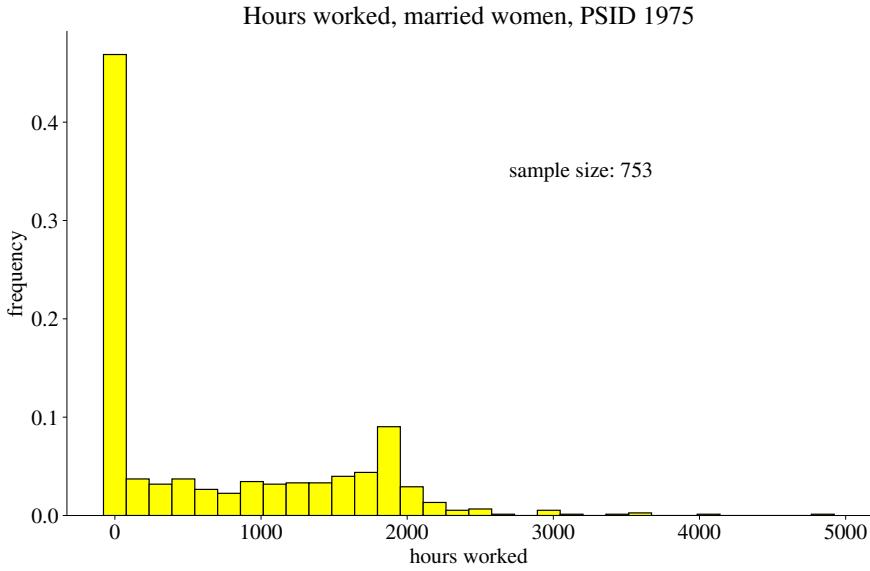


Figure 24.5: Example of probit model, Hill et al (2008), Table 16.1

here is that the contribution to the likelihood function (again, see Remark 24.8) is

$$L_i(\text{if } y_i^* > 0) = \text{pdf}(\varepsilon_i | \varepsilon_i > -x_i' \beta) \times \Pr(\varepsilon_i > -x_i' \beta) = \text{pdf}(\varepsilon_i),$$

where $\text{pdf}(\varepsilon_i)$ is the pdf of $N(0, \sigma^2)$, which is the same as $\phi(\varepsilon/\sigma)/\sigma$. Take logs and introduce the δ_i to pick out observations where $y_i = 0$. \square

24.3.2 Interpretation of the Tobit Model

We could be interested in several things. *First*, how is the probability of $y_i = 0$ affected by a change in regressor k ? The probability is $\Phi(-x_i' \beta/\sigma)$ as in (24.17) and the derivative is

$$\frac{\partial \Pr(y_i = 0)}{\partial x_{k,i}} = -\phi(-x_i' \beta/\sigma) \beta_k/\sigma, \quad (24.18)$$

where $\phi()$ is the pdf of a $N(0, 1)$ variable.

(Clearly, $\phi()$ is symmetric, so $-\phi(x_i' \beta/\sigma) \beta_k/\sigma$ is the same.) This is similar to (24.2), but here refers to $y_i = 0$ (not $y_i = 1$). The derivative (24.18) is high (in absolute value) when $x_i' \beta \approx 0$, since a small change in x_k can then tip the balance towards $y_i = 0$. In contrast, when $x_i' \beta$ is very small or very large, then a small change in x_k does not matter much for the probability. *Second*, how is the expected value of y_i affected by a change in

	OLS	MLE
c	1335.3 (5.7)	1349.9 (3.4)
educ	27.1 (2.2)	73.3 (3.6)
exper	48.0 (13.2)	80.5 (13.1)
age	-31.3 (-7.9)	-60.8 (-9.1)
kids16	-447.9 (-7.7)	-918.9 (-8.0)
N	753.0	753.0

Table 24.1: Tobit estimation of hours worked. Example of a Tobit model, Hill et al (2008), Table 16.8. Numbers in parentheses are t-stats ('sandwich' approach for MLE).

regressor k ? Once again, we can calculate a derivative

$$\frac{\partial \mathbb{E} y_i}{\partial x_{k,i}} = \Phi(x'_i \beta / \sigma) \beta_k. \quad (24.19)$$

For low values of $x'_i \beta$, the derivative in (24.19) is close to zero (since $\Phi(x'_i \beta / \sigma) \approx 0$). In contrast, for high values of $x'_i \beta$, the derivative is close to β_k .

Proof (*of (24.19)) We have $\mathbb{E} y_i = \Phi(-x'_i \beta / \sigma) \times 0 + \Phi(x'_i \beta / \sigma) \mathbb{E}(y_i | y_i^* > 0)$, since $1 - \Phi(-x'_i \beta / \sigma) = \Phi(x'_i \beta / \sigma)$. From Remark 24.6, expectations of (24.15) gives $\mathbb{E}(y_i | y_i^* > 0)$ as $x'_i \beta + \sigma \phi(-x'_i \beta / \sigma) / \Phi(x'_i \beta / \sigma)$. Combing gives $\mathbb{E}(y_i | y_i^* > 0) = \Phi(x'_i \beta / \sigma) x'_i \beta + \sigma \phi(x'_i \beta / \sigma)$. Differentiating gives $\Phi(x'_i \beta / \sigma) \beta + \phi(x'_i \beta / \sigma) x'_i \beta \beta / \sigma + \phi'(x'_i \beta / \sigma) \beta$, but it is well known that $\phi'(z) = -z \phi(z)$, so the last two terms cancel. \square

24.4 A Sample Selection Model

Recall that in a Tobit model, $x'_i \beta + \varepsilon_i$ determines both the probability of observing y_i^* and its value. The Heckit model (sample selection model) relaxes that. It is a two equation model

$$w_i^* = x'_{1i} \beta_1 + \varepsilon_{1i} \quad (24.20)$$

$$h_i^* = x'_{2i} \beta_2 + \varepsilon_{2i}. \quad (24.21)$$

In early applications, w_i^* could be individual productivity (measured by the wage) and h_i^* could be labour supply, and x_{1i} and x_{2i} could contain information about education, age, etc. The data on w_i^* (hourly wage) is only observed for people who work, and h_i^* is only observed as 0 or 1 (does not work/works). In short,

$$\text{Data: } (w_i, h_i) = \begin{cases} w_i = w_i^*, h_i = 1 & \text{if } h_i^* > 0 \\ w_i \text{ not observed, } h_i = 0 & \text{otherwise.} \end{cases} \quad (24.22)$$

The regressors x_{1i} and x_{2i} are observed for all i . In the special case where $\text{Corr}(h_i^*, w_i^*) = 1$, then we are back in standard Tobit model, where we used the notation y_i^* for what is here called w_i^* .

It is typical to assume that the residuals in the two equations could be jointly normally distributed, but correlated

$$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix} \right). \quad (24.23)$$

Notice that $\text{Var}(\varepsilon_{2i}) = 1$ is a normalization. A correlation, $\sigma_{12} \neq 0$, means that some unobserved characteristics (part of the residuals) are affecting both equations. For instance, “ability” may be hard to measure (so it is not in the x_{1i} or x_{2i} vectors, but in the residuals), but is likely to affect both productivity and the labour supply choice.

To understand the properties of this model, notice that the expected value of w_i , conditional on $h_i = 1$, is

$$E(w_i | h_i = 1) = x'_{1i} \beta_1 + \sigma_{12} \lambda_i, \text{ where } \lambda_i = \phi(x'_{2i} \beta_2) \Phi(x'_{2i} \beta_2), \quad (24.24)$$

where λ_i is called the inverse Mill’s ratio or Heckman’s lambda. The point of (24.24) is that the covariance of the residuals in the two equations (24.20)–(24.21) is crucial. In fact, when $\sigma_{12} = 0$, then we can estimate (24.20) with OLS. Otherwise, it is biased (and inconsistent).

Proof (of (24.24))

$$\begin{aligned} E(w_i | h_i = 1) &= x'_{1i} \beta_1 + E(\varepsilon_{1i} | h_i = 1) \\ &= x'_{1i} \beta_1 + E(\varepsilon_{1i} | \varepsilon_{2i} > -x'_{2i} \beta_2), \end{aligned}$$

since $h_i = 1$ when $h_i^* = x'_{2i} \beta_2 + \varepsilon_{2i} > 0$. If $(\varepsilon_{1i}, \varepsilon_{2i})$ have a joint normal distribution as in (24.23), then it is a standard result for bivariate normally distributed variables that

$$E(\varepsilon_{1i} | \varepsilon_{2i} > -q) = \sigma_{12} \lambda, \text{ where } \lambda = \phi(q) / \Phi(q).$$

□

Another way to see the problem highlighted by (24.24) is the following. Consider the observable data (when $h_i = 1$)

$$w_i = x'_{1i}\beta_1 + \varepsilon_{1i} \quad (24.25)$$

and ask if $E(x_{1i}\varepsilon_{1i}) = 0$ for this data? To keep it simple, suppose $x_{2i} = 0$: w_i is observed only when $\varepsilon_{2i} > 0$. If $\text{Corr}(\varepsilon_{1i}, \varepsilon_{2i}) > 0$, our sample of w_i actually contains mostly observations when $\varepsilon_{1i} > 0$, so ε_{1i} isn't zero on average in the sample. This gives a *sample selection bias*.

Is $\sigma_{12} \neq 0$? To assess that, we must think about the economics of the problem. In wage and labour supply equations: ε_{1t} and ε_{2t} may capture some unobservable factor that makes a person more productive at the same time as more prone to supply more labour.

What if $\text{Cov}(x_{1i}, \lambda_i) = 0$ (although $\sigma_{12} \neq 0$)? In that case, OLS estimation of equation (24.25) is consistent (recall the case of uncorrelated regressors: we can then estimate one slope coefficient at a time). The conclusion is that the bias of OLS comes from $\sigma_{12} \neq 0$ and $\text{Cov}(x_{1i}, x_{2i}) \neq 0$ since then $\text{Cov}(x_{1i}, \lambda_i) \neq 0$ (although λ is a non-linear function of x_{2i}).

24.4.1 Estimation

Use MLE or Heckman's 2-step approach, which is as follows: First, estimate (24.21) with the probit method (recall $h_i = 0$ or 1) using the likelihood function in (24.17). Extract $x'_{2i}\hat{\beta}_2$ and create $\hat{\lambda}_i$ as in (24.24). See Table 24.2 for an example. Second, estimate $(\beta_1$ and $\sigma_{12})$ with LS

$$w_i = x'_{1i}\beta_1 + \sigma_{12}\hat{\lambda}_i + \eta_i \quad (24.26)$$

on the data where w_i is observed (and not artificial set to zero or some other value). This approach gives consistent estimates, but we may need to adjust standard errors (unless you test under the null hypothesis that $\sigma_{12} = 0$).

Further Reading

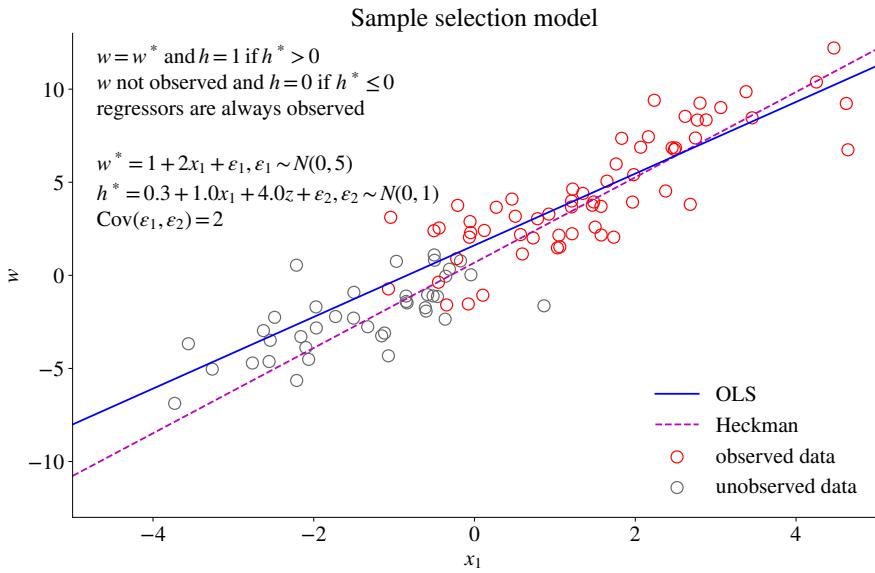


Figure 24.6: Sample selection model

Heckman, step 1	
c	-4.16 (-2.97)
age	0.19 (2.79)
age2/100	-0.24 (-3.12)
faminc/1000	0.05 (1.00)
kids	-0.45 (-3.54)
educ	0.10 (4.35)

Table 24.2: Probit estimation of labour market participation. 1st step of Heckman estimation. Example of a Heckman model, Greene (2003), Table 22.7 (corrected). Numbers in parentheses are t-stats.

	LS	Heckman
c	-2.56 (-2.77)	-0.97 (-0.48)
exper	0.03 (0.53)	0.02 (0.34)
exper2/100	-0.00 (-0.14)	0.00 (0.07)
educ	0.48 (7.24)	0.42 (4.24)
cit	0.45 (1.42)	0.44 (1.41)
lambda		-1.10 (-0.88)

Table 24.3: OLS and Heckman estimation of log wages, married women, PSID 1975. Example of a Heckman model, Greene (2003), Table 22.7 (corrected). Numbers in parentheses are t-stats.

Bibliography

- Ait-Sahalia, Y., 1996, “Testing continuous-time models of the spot interest rate,” *Review of Financial Studies*, 9, 385–426.
- Alexander, C., 2008, *Market Risk Analysis: Value at Risk Models*, Wiley.
- Amemiya, T., 1985, *Advanced econometrics*, Harvard University Press, Cambridge, Massachusetts.
- Andrews, D. W. K., and J. C. Monahan, 1992, “An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator,” *Econometrica*, 60, 953–966.
- Ang, J. S., and S. J. Ciccone, 2001, “International differences in analyst forecast properties,” mimeo, Florida State University.
- Arellano, M., 2003, *Panel Data Econometrics*, Oxford University Press, Oxford.
- Bali, T. G., R. F. Engle, and S. Murray, 2016, *Empirical Asset Pricing*, Wiley, Hoboken, New Jersey.
- Baltagi, D. H., 2008, *Econometric Analysis of Panel Data*, Wiley, 4th edn.
- Barber, B., R. Lehavy, M. McNichols, and B. Trueman, 2001, “Can investors profit from the prophets? Security analyst recommendations and stock returns,” *Journal of Finance*, 56, 531–563.
- Berkowitz, J., and L. Kilian, 2000, “Recent developments in bootstrapping time series,” *Econometric-Reviews*, 19, 1–48.
- Blume, M. E., 1971, “On the Assessment of Risk,” *Journal of Finance*, 26, 1–10.
- Bodie, Z., A. Kane, and A. J. Marcus, 2002, *Investments*, McGraw-Hill/Irwin, Boston, 5th edn.

- Bondt, W. F. M. D., 1991, “What do economists know about the stock market?,” *Journal of Portfolio Management*, 17, 84–91.
- Bondt, W. F. M. D., and R. H. Thaler, 1990, “Do security analysts overreact?,” *American Economic Review*, 80, 52–57.
- Boni, L., and K. L. Womack, 2006, “Analysts, industries, and price momentum,” *Journal of Financial and Quantitative Analysis*, 41, 85–109.
- Brock, W., J. Lakonishok, and B. LeBaron, 1992, “Simple technical trading rules and the stochastic properties of stock returns,” *Journal of Finance*, 47, 1731–1764.
- Brockwell, P. J., and R. A. Davis, 1991, *Time series: theory and methods*, Springer Verlag, New York, second edn.
- Campbell, J. Y., 2018, *Financial decisions and markets*, Princeton University Press, Princeton, New Jersey.
- Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The econometrics of financial markets*, Princeton University Press, Princeton, New Jersey.
- Campbell, J. Y., and S. B. Thompson, 2008, “Predicting the equity premium out of sample: can anything beat the historical average,” *Review of Financial Studies*, 21, 1509–1531.
- Chance, D. M., and M. L. Hemler, 2001, “The performance of professional market timers: daily evidence from executed strategies,” *Journal of Financial Economics*, 62, 377–411.
- Chen, N.-F., R. Roll, and S. A. Ross, 1986, “Economic forces and the stock market,” *Journal of Business*, 59, 383–403.
- Clark, T. E., and M. W. McCracken, 2001, “Tests of equal forecast accuracy and encompassing for nested models,” *Journal of Econometrics*, 105, 85–110.
- Clark, T. E., and K. D. West, 2007, “Approximately normal tests for equal predictive accuracy in nested models,” *Journal of Ec*, 138, 291–311.
- Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.
- Cox, J. C., J. E. Ingersoll, and S. A. Ross, 1985, “A theory of the term structure of interest rates,” *Econometrica*, 53, 385–407.

- Dahlquist, M., J. V. Martinez, and P. Söderlind, 2017, “Individual Investor Activity and Performance,” *The Review of Financial Studies*, 30, 866–899.
- Davidson, J., 2000, *Econometric theory*, Blackwell Publishers, Oxford.
- Davidson, R., and J. G. MacKinnon, 1993, *Estimation and inference in econometrics*, Oxford University Press, Oxford.
- Davison, A. C., and D. V. Hinkley, 1997, *Bootstrap methods and their applications*, Cambridge University Press.
- DeGroot, M. H., 1986, *Probability and statistics*, Addison-Wesley, Reading, Massachusetts.
- DeMiguel, V., L. Garlappi, and R. Uppal, 2009, “Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?,” *Review of Financial Studies*, 22, 1915–1953.
- Diebold, F. X., 2001, *Elements of forecasting*, South-Western, 2nd edn.
- Diebold, F. X., and R. S. Mariano, 1995, “Comparing predictive accuracy,” *Journal of Business and Economic Statistics*, 13, 253–265.
- Duan, J., 1995, “The GARCH option pricing model,” *Mathematical Finance*, 5, 13–32.
- Duffee, G. R., 2005, “Time variation in the covariance between stock returns and consumption growth,” *Journal of Finance*, 60, 1673–1712.
- Ederington, L. H., and J. C. Goh, 1998, “Bond rating agencies and stock analysts: who knows what when?,” *Journal of Financial and Quantitative Analysis*, 33, 569–585.
- Efron, B., T. Hasti, I. Johnstone, and R. Tibshirani, 2004, “Least angle regression,” *The Annals of Statistics*, 32, 407–499.
- Efron, B., and R. J. Tibshirani, 1993, *An introduction to the bootstrap*, Chapman and Hall, New York.
- Elliot, G., and A. Timmermann, 2016, *Economic forecasting*, Princeton University Press, Princeton, New Jersey.
- Elton, E. J., M. J. Gruber, S. J. Brown, and W. N. Goetzmann, 2014, *Modern portfolio theory and investment analysis*, John Wiley and Sons, 9th edn.

- Enders, W., 2004, *Applied econometric time series*, John Wiley and Sons, New York, 2nd edn.
- Engle, R., 2009, *Anticipating Correlations*, Princeton University Press.
- Engle, R. F., 2002, “Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models,” *Journal of Business and Economic Statistics*, 20, 339–351.
- Fabozzi, F. J., S. M. Focardi, and P. N. Kolm, 2006, *Financial modeling of the equity market*, Wiley Finance.
- Fama, E., and J. MacBeth, 1973, “Risk, return, and equilibrium: empirical tests,” *Journal of Political Economy*, 71, 607–636.
- Fama, E. F., and K. R. French, 1993, “Common risk factors in the returns on stocks and bonds,” *Journal of Financial Economics*, 33, 3–56.
- Fama, E. F., and K. R. French, 1996, “Multifactor explanations of asset pricing anomalies,” *Journal of Finance*, 51, 55–84.
- Franses, P. H., and D. van Dijk, 2000, *Non-linear time series models in empirical finance*, Cambridge University Press.
- Gibbons, M., S. Ross, and J. Shanken, 1989, “A test of the efficiency of a given portfolio,” *Econometrica*, 57, 1121–1152.
- Glosten, L. R., R. Jagannathan, and D. Runkle, 1993, “On the relation between the expected value and the volatility of the nominal excess return on stocks,” *Journal of Finance*, 48, 1779–1801.
- Gourieroux, C., and J. Jasiak, 2001, *Financial econometrics: problems, models, and methods*, Princeton University Press.
- Goyal, A., and I. Welch, 2008, “A comprehensive look at the empirical performance of equity premium prediction,” *Review of Financial Studies* 2008, 21, 1455–1508.
- Greene, W. H., 2018, *Econometric analysis*, Pearson Education Ltd, 8th edn.
- Hamilton, J. D., 1994, *Time series analysis*, Princeton University Press, Princeton.

- Hansen, B. E., 2022a, *Econometrics*, Princeton University Press, Princeton.
- Hansen, B. E., 2022b, *Probability and Statistics for Economists*, Princeton University Press, Princeton.
- Härdle, W., 1990, *Applied nonparametric regression*, Cambridge University Press, Cambridge.
- Harvey, A. C., 1989, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.
- Hastie, T., R. Tibshirani, and J. Friedman, 2001, *The elements of statistical learning: data mining, inference and prediction*, Springer Verlag.
- Heston, S. L., and S. Nandi, 2000, “A closed-form GARCH option valuation model,” *Review of Financial Studies*, 13, 585–625.
- Hill, R. C., W. E. Griffiths, and G. C. Lim, 2008, *Principles of Econometrics*, John Wiley and Sons, 3rd edn.
- Horowitz, J. L., 2001, “The Bootstrap,” in J.J. Heckman, and E. Leamer (ed.), *Handbook of Econometrics* . , vol. 5, Elsevier.
- Hull, J. C., 2022, *Options, futures, and other derivatives*, Pearson, 11th edn.
- JP Morgan, 1996, “RiskMetrics Technical Document,” Discussion paper, JP Morgan, New York, NY.
- Lo, A. W., H. Mamaysky, and J. Wang, 2000, “Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation,” *Journal of Finance*, 55, 1705–1765.
- MacKinlay, C., 1995, “Multifactor models do not explain deviations from the CAPM,” *Journal of Financial Economics*, 38, 3–28.
- Makridakis, S., S. C. Wheelwright, and R. J. Hyndman, 1998, *Forecasting: methods and applications*, Wiley, New York, 3rd edn.
- McDonald, R. L., 2014, *Derivatives markets*, Pearson, 3rd edn.
- McNeil, A. J., R. Frey, and P. Embrechts, 2005, *Quantitative risk management*, Princeton University Press.

- Mittelhammer, R. C., 1996, *Mathematical statistics for economics and business*, Springer-Verlag, New York.
- Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric foundations*, Cambridge University Press, Cambridge.
- Nelson, D. B., 1991, “Conditional heteroskedasticity in asset returns,” *Econometrica*, 59, 347–370.
- Pagan, A., and A. Ullah, 1999, *Nonparametric econometrics*, Cambridge University Press.
- Pesaran, M. H., 2015, *Time series and panel data econometrics*, Oxford University Press.
- Petersen, M. A., 2009, “Estimating standard errors in finance panel data sets: comparing approaches,” *The Review of Financial Studies*, 22, 435–480.
- Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric models and economic forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.
- Priestley, M. B., 1981, *Spectral analysis and time series*, Academic Press.
- Reilly, F. K., and K. C. Brown, 2012, *Analysis of investments & management of portfolios*, South-Western, 10th edn.
- Scott, D. W., 1985, “Averaged shifted histograms: effective non-parametric density estimators in several dimensions,” *The Annals of Statistics*, 13, 1024–1040.
- Silverman, B. W., 1986, *Density estimation for statistics and data analysis*, Chapman and Hall, London.
- Singleton, K. J., 2006, *Empirical dynamic asset pricing*, Princeton University Press.
- Söderlind, P., 2010, “Predicting stock price movements: regressions versus economists,” *Applied Economics Letters*, 17, 869–874.
- Stekler, H. O., 1991, “Macroeconomic forecast evaluation techniques,” *International Journal of Forecasting*, 7, 375–384.
- Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.
- Verbeek, M., 2017, *A guide to modern econometrics*, Wiley, 5th edn.

Wooldridge, J. M., 2010, *Econometric analysis of cross section and panel data*, MIT Press, 2nd edn.