

Lecture Notes in Financial Econometrics (MSc course)

Paul Söderlind¹

17 February 2019

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.

Contents

1 Review of Statistics	7
1.1 Random Variables and Distributions	7
1.2 Moments	15
1.3 Distributions Commonly Used in Tests	19
1.4 Normal Distribution of the Sample Mean	21
1.5 Appendix: Statistical Tables	23
1.6 Appendix: Data Sources	23
2 Least Squares Estimation	28
2.1 Least Squares: The Optimization Problem and Its Solution	28
2.2 The Distribution of $\hat{\beta}$	38
2.3 The Distribution of $\hat{\beta}$: More General Results	46
2.4 Appendix: A Primer in Matrix Algebra	52
3 Index Models	56
3.1 Single-Index Models	56
3.2 Estimating Beta	61
3.3 Multi-Index Models	63
3.4 Principal Component Analysis*	64
4 Least Squares: Testing	68
4.1 Hypothesis Testing	68
4.2 Heteroskedasticity	79
4.3 Autocorrelation	82
5 Testing CAPM and Multifactor Models	88
5.1 Market Model	88
5.2 Calendar Time Regressions	97

5.3	Several Factors	98
5.4	Fama-MacBeth*	99
6	Model Selection and Other Topics	103
6.1	Model Selection I	103
6.2	Model Selection II	104
6.3	Comparing Non-Nested Models	109
6.4	Non-Linear Models	109
6.5	Outliers	110
6.6	Estimation on Subsamples	111
6.7	Missing Data	114
6.8	Robust Estimation*	114
7	Asymptotic Results on OLS*	120
7.1	Motivation of Asymptotics	120
7.2	Asymptotics: Consistency	120
7.3	When LS Cannot be Saved	124
7.4	Instrumental Variables Method (IV)	128
7.5	Asymptotic Normality	130
8	Time Series Analysis	132
8.1	Descriptive Statistics	132
8.2	Stationarity	135
8.3	White Noise	136
8.4	AR(1)	137
8.5	AR(p)	144
8.6	Moving Average (MA)	150
8.7	ARMA(p,q)	151
8.8	VAR(p)	152
8.9	Non-stationary Processes	156
9	Predicting Asset Returns	166
9.1	Autocorrelations and Autoregressions	166
9.2	Other Predictors and Methods	172
9.3	Out-of-Sample Forecasting Performance	175
9.4	Forecast Averaging	183

9.5	Evaluating Forecasting Performance	184
9.6	Security Analysts	187
10	Maximum Likelihood Estimation	192
10.1	Maximum Likelihood	192
10.2	Key Properties of MLE	197
10.3	Three Test Principles	199
10.4	QMLE*	199
11	ARCH and GARCH	201
11.1	Heteroskedasticity	201
11.2	ARCH Models	205
11.3	GARCH Models	209
11.4	Non-Linear Extensions	212
11.5	(G)ARCH-M*	213
11.6	Multivariate (G)ARCH	214
12	Risk Measures	218
12.1	Value at Risk	218
12.2	Backtesting a VaR Model	221
12.3	Expected Shortfall	222
12.4	Semivariance and Max Drawdown	225
13	Return Distributions (Univariate)	229
13.1	Estimating and Testing Distributions	229
13.2	Tail Distribution*	238
14	Return Distributions (Multivariate)*	248
14.1	Recap of Univariate Distributions	248
14.2	Threshold Exceedance and Tail Distribution*	249
14.3	Exceedance Correlations*	256
14.4	Beyond (Linear) Correlations*	256
14.5	Copulas*	262
14.6	Simulating Joint Distributions*	268

15 Option Pricing and Estimation of Continuous Time Processes	273
15.1 The Black-Scholes Model	273
15.2 Estimation of the Volatility of a Random Walk Process	280
16 Event Studies	287
16.1 Basic Structure of Event Studies	287
16.2 Models of Normal Returns	289
16.3 Testing the Abnormal Return	293
16.4 Quantitative Events	296
17 Kernel Density Estimation and Regression	298
17.1 Non-Parametric Regression	298
17.2 Local Linear Regressions*	304
17.3 Examples of Non-Parametric Estimation	307
18 Simulating the Finite Sample Properties*	311
18.1 Introduction	311
18.2 Monte Carlo Simulations	311
18.3 Bootstrapping	317
19 Financial Panel Data	322
19.1 Introduction to Panel Data	322
19.2 Portfolio Sorts	323
19.3 Calendar Time Regressions and SURE	325
19.4 An Overview of Different Panel Data Models	326
19.5 Pooled OLS	328
19.6 The Within Estimator (“Fixed Effects Estimator”)	331
19.7 The First-Difference Estimator	335
19.8 Differences-in-Differences Estimator	336
19.9 Random Effects Model*	336
19.10 Fama-MacBeth	338
20 Binary Choice and Truncated Models*	341
20.1 Binary Choice Model	341
20.2 Truncated Regression Model	348
20.3 Censored Regression Model (Tobit Model)	352

20.4 Heckit: A Sample Selection Model	356
21 LAD and Quantile Regressions*	360
21.1 LAD	360
21.2 Quantile Regressions	363

Warning: a few of the tables and figures are reused in later chapters. This can mess up the references, so that the text refers to a table/figure in another chapter. No worries: it is really the same table/figure. I promise to fix this some day...

Chapter 1

Review of Statistics

Reference: Verbeek (2012) Appendix B

More advanced material is denoted by a star (*). It is not required reading.

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

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

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

The probability that $x \leq B$ (that is, $-\infty < x \leq B$) is measured by the *cumulative distribution function* (cdf). For instance, if x has a $N(0, 1)$ distribution, then $\Pr(x \leq -1.645) = 0.05$ and $\Pr(x \leq 0) = 0.5$. See Figure 15.2 for an illustration.

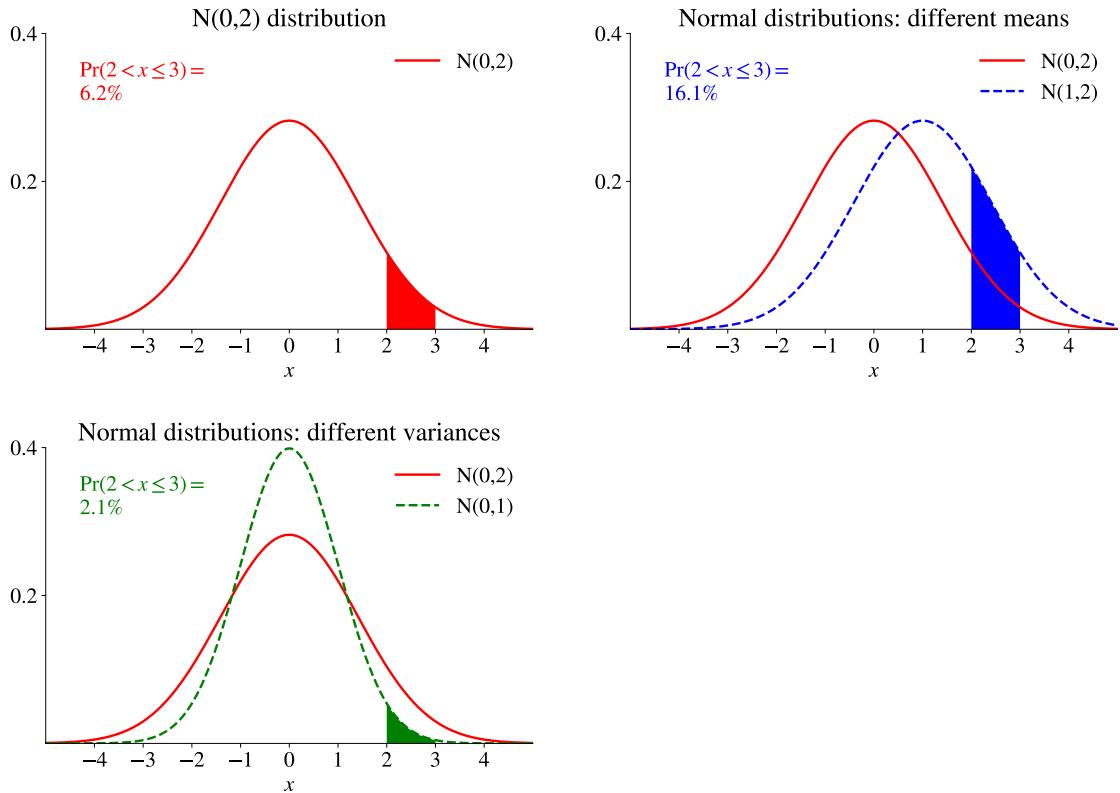


Figure 1.1: A few different normal distributions

If we invert the cdf, then we get the *quantiles* of the random variable. For instance, the 0.05th quantile 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 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.

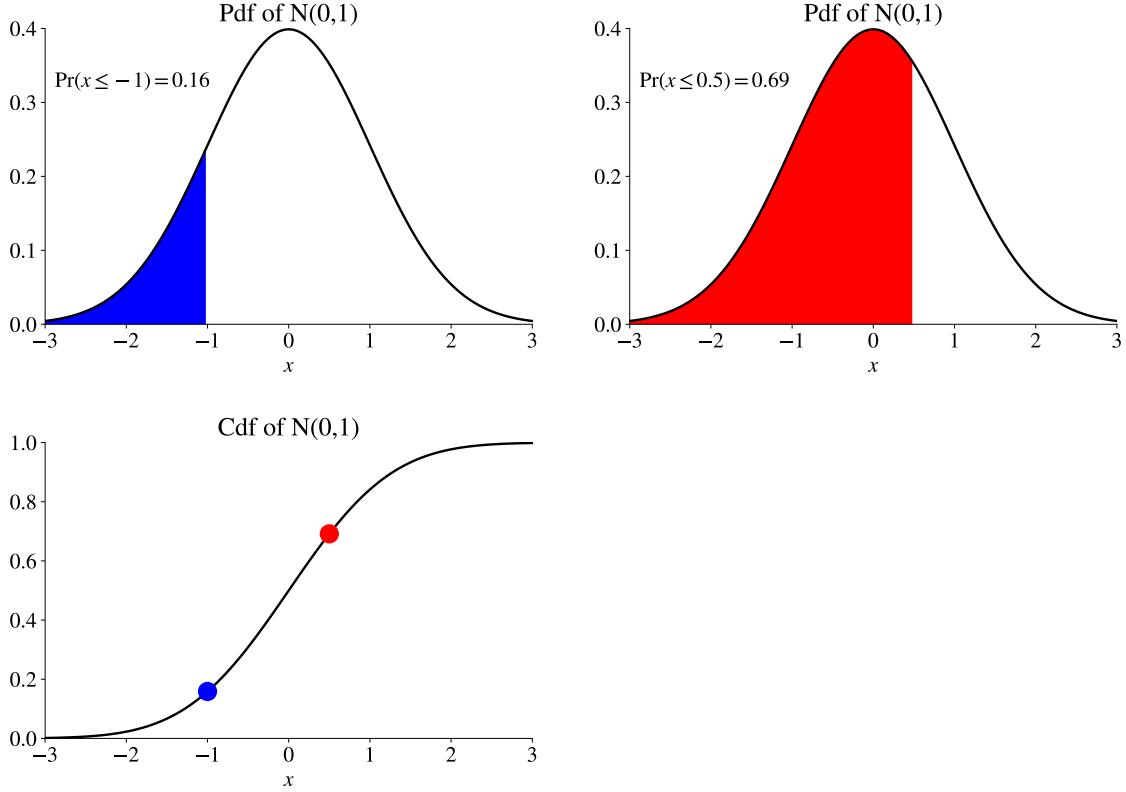


Figure 1.2: Pdf and cdf of $N(0,1)$

A joint normal distributions is completely described by the means and the 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 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 1.4 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 guess of y . The correlation of x and y is defined as

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

See Figure 1.4 for an example.

If two random variables happen to be 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.

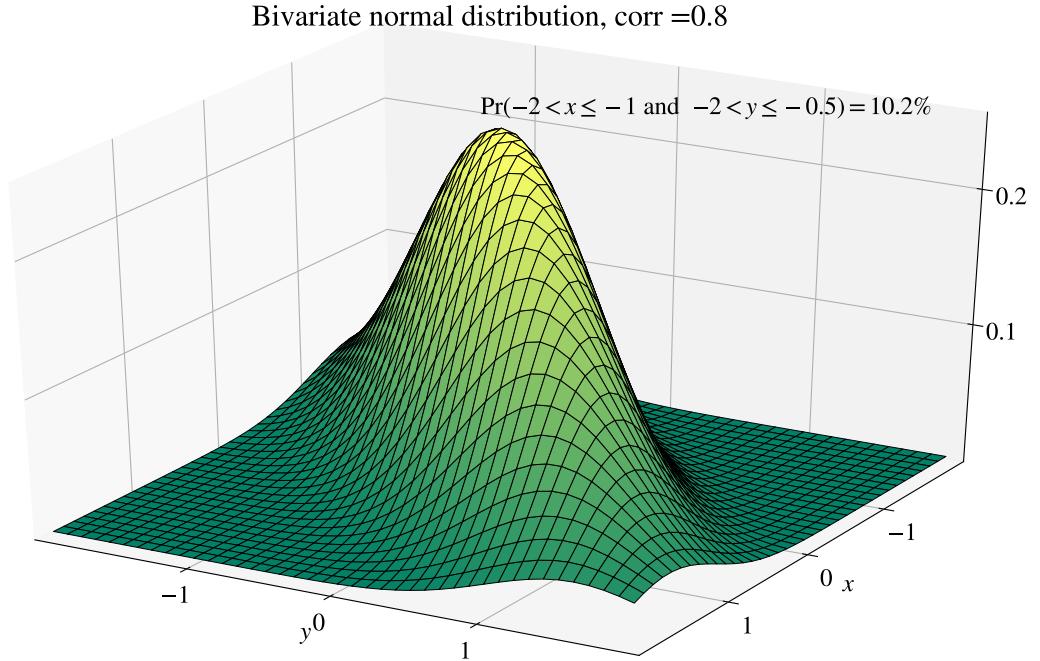


Figure 1.3: Density function bivariate normal distribution

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 guess of y given that we know x . Similarly, the conditional variance can be interpreted as the variance of the

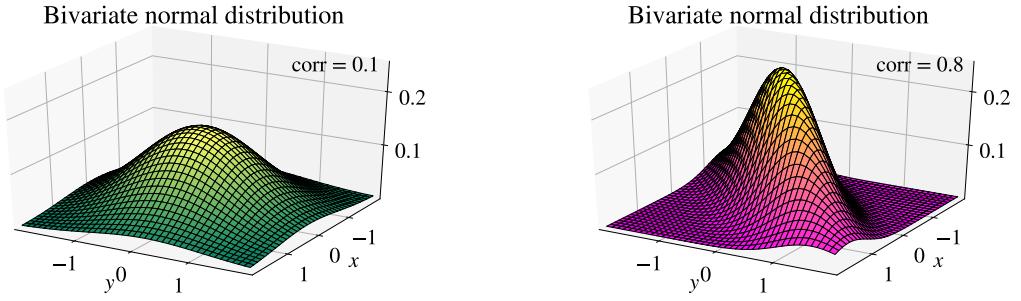


Figure 1.4: Density function bivariate normal distributions

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

$$y|x \sim N\left(\mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x), \sigma_y^2 - \frac{\sigma_{xy}\sigma_{yx}}{\sigma_x^2}\right). \quad (1.7)$$

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 1.5 for an illustration: notice how the location and the width of the conditional distribution of y changes as a function of the correlation and the value of x .

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_t 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 (as a probability density function).

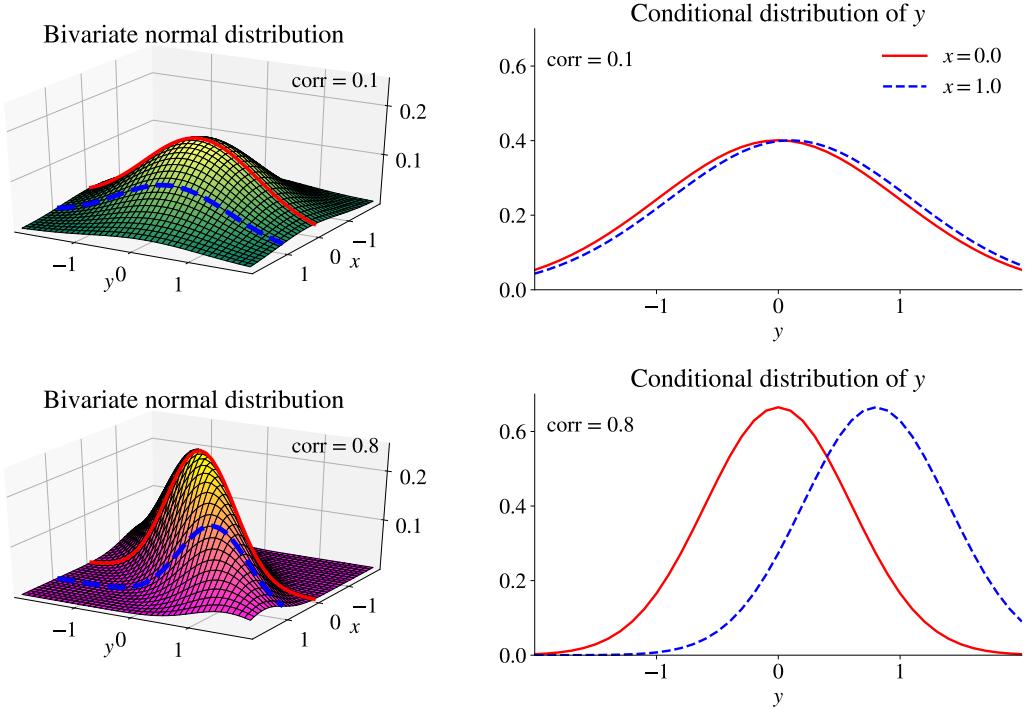


Figure 1.5: Density functions of normal distributions

See Figure 1.6 for an illustration.

1.1.5 Confidence Bands and t-tests

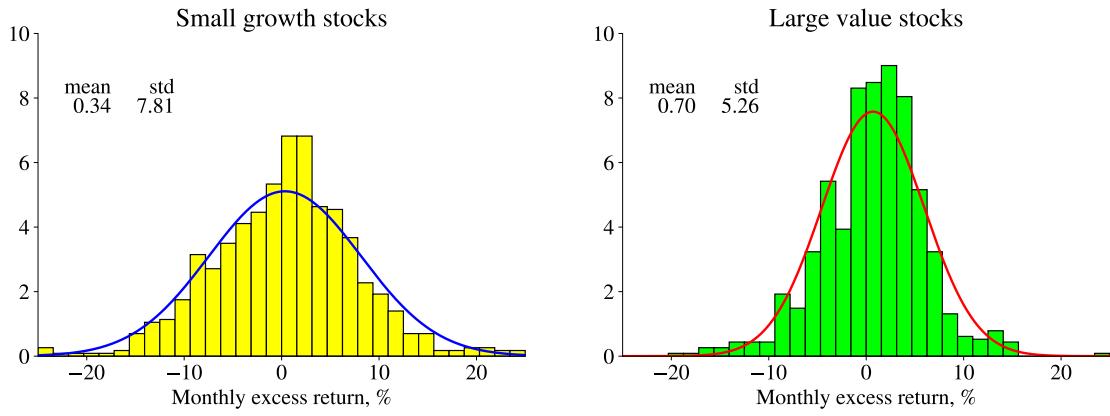
Confidence bands are typically only used for symmetric distributions. For instance, a 90% confidence band is constructed by finding a critical value c such that

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

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

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

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



Monthly data on two U.S. indices, 1957:01-2017:12
Sample size: 732
Solid line: estimated normal distribution

Figure 1.6: Histogram of returns, the curve is a normal distribution with the same mean and standard deviation as the return series

a 90% confidence band around the point estimate ($\hat{\beta}$) as

$$[\hat{\beta} - 1.64\sigma, \hat{\beta} + 1.64\sigma]. \quad (1.10)$$

In case this band does not include your null hypothesis $\beta = q$ ($q = 0$ is a commonly used special case), then we would be 90% that the (true) regression coefficient is different from q .

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

$$[q - 1.64\sigma, q + 1.64\sigma]. \quad (1.11)$$

If this band does not include the point estimate ($\hat{\beta}$), then we are also 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

$$t = \frac{\hat{\beta} - q}{\sigma}, \quad (1.12)$$

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

$$[-1.64, 1.64]. \quad (1.13)$$

(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 , say). The parameter could, for instance, be the mean or a regression coefficient. This estimate is actually a random variable so it makes sense to construct a confidence band as in (1.10). The reason for why it is a random variable is that another sample is most likely to produce a different estimate—and that if we could try all possible samples then the different estimates would have some sort of distribution. *If* we are willing to assume that data for those other samples would be similar (scattered around the same mean, showing the same degree of dispersion, etc) to the sample we actually study (observations 1 to T), then we can use our sample to guess how much other samples would differ. For instance, we can estimate the variance of the data (σ^2) and draw the conclusions about how different the sample averages would be in different samples (it would have a variance of σ^2/T as discussed in (1.16)).

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 $H_0: \beta = 0$.)

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

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

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 Mean and Standard Deviation

The mean and variance of a series are estimated as

$$\bar{x} = \sum_{t=1}^T x_t / T \text{ and } \hat{\sigma}^2 = \sum_{t=1}^T (x_t - \bar{x})^2 / T. \quad (1.15)$$

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 T .) See Figure 1.6 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

$$\text{Var}(\bar{x}) = \sigma^2 / T, \text{ if } x_t \text{ is iid.} \quad (1.16)$$

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

$$E \bar{x} = E x_t = \mu. \quad (1.17)$$

Since sample averages are typically normally distributed in large samples, we thus have

$$\bar{x} \sim N(\mu, \sigma^2 / T), \quad (1.18)$$

so we can construct a *t-stat* as

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{T}}, \quad (1.19)$$

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

Proof. (of (1.16)–(1.17)) To prove (1.16), notice that

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

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 (1.17)

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

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. ■

1.2.2 Skewness and Kurtosis

The skewness, kurtosis and Bera-Jarque test for normality are useful diagnostic tools. They are

	<u>Test statistic</u>	<u>Distribution</u>	
skewness	$= \frac{1}{T} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)^3$	$N(0, 6/T)$	(1.20)
kurtosis	$= \frac{1}{T} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)^4$	$N(3, 24/T)$	
Bera-Jarque	$= \frac{T}{6} \text{skewness}^2 + \frac{T}{24} (\text{kurtosis} - 3)^2$	χ^2_2	

This is implemented by using the estimated mean and standard deviation. See Figure 1.6 for an illustration.

The distributions stated on the right hand side of (1.20) 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 (Bera-Jarque) 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 and kurtosis by traditional t-stats as in

$$t = \frac{\text{skewness}}{\sqrt{6/T}} \text{ and } t = \frac{\text{kurtosis} - 3}{\sqrt{24/T}}, \quad (1.21)$$

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

1.2.3 Covariance and Correlation

The covariance of two variables (here x and y) is typically estimated as

$$\hat{\sigma}_{xy} = \sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y}) / T. \quad (1.22)$$

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

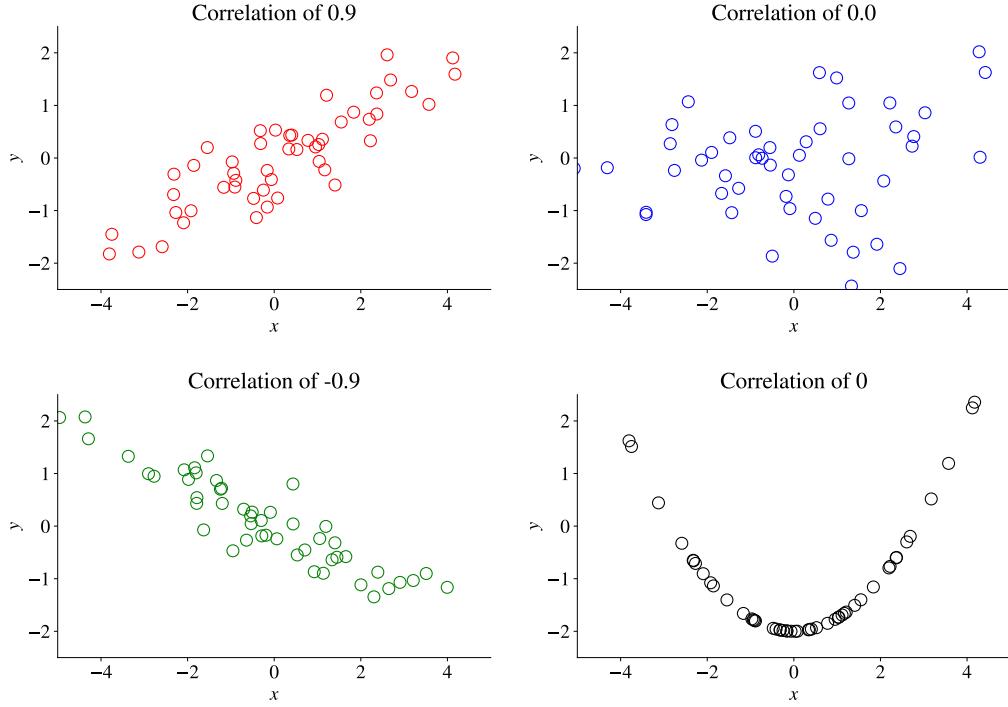


Figure 1.7: Example of correlations.

The correlation of two variables is then estimated as

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}, \quad (1.23)$$

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 1.7. See Figure 1.8 for an empirical illustration.

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

$$\frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}} \sim N(0, 1/T), \quad (1.24)$$

so we can form a t-stat as

$$t = \sqrt{T} \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}, \quad (1.25)$$

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

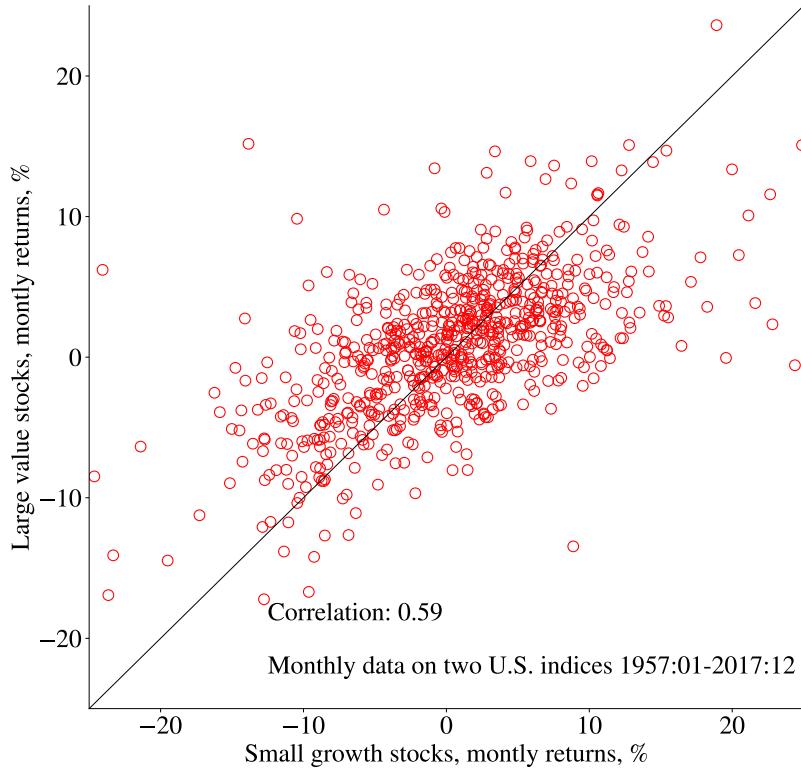


Figure 1.8: Scatter plot of two different portfolio returns

1.2.4 Correlations vs. Causality

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

$$(x, \varepsilon) \Rightarrow y \quad (1.26)$$

$$(y, u) \Rightarrow x$$

$$(z, u) \Rightarrow x \text{ and } (z, \varepsilon) \Rightarrow y$$

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 . This reasoning carries over to regression analysis too. In many regressions we would like to capture the causality, although forecasting models are more focused on the correlation per se.

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

$$t = \frac{x - \mu}{\sigma} \sim N(0, 1). \quad (1.27)$$

To see this, notice that $x - \mu$ has a mean of zero and that x/σ has a standard deviation of unity. (This result is the motivation for why the confidence band (1.13) gives the same result as (1.11).)

1.3.2 t -distribution

If we instead need to estimate σ to use in (1.27), then the test statistic has t_n -distribution

$$t = \frac{x - \mu}{\hat{\sigma}} \sim t_n, \quad (1.28)$$

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) increases. See Figure 1.9 for a comparison and Table 1.1 for critical values.

Example 1.2 (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.

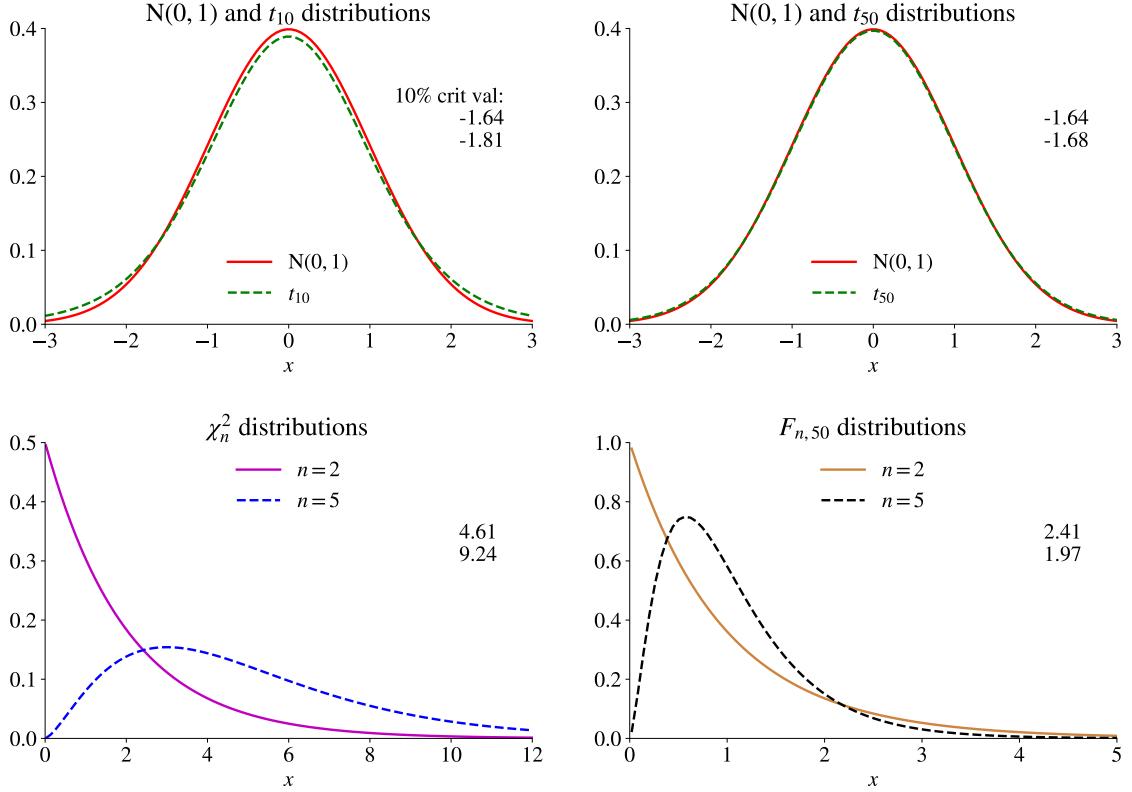


Figure 1.9: Probability density functions

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

$$v' \Sigma^{-1} v \sim \chi_n^2, \text{ if the } n \times 1 \text{ vector } v \sim N(0, \Sigma). \quad (1.29)$$

See Figure 1.9 for an illustration and Table 1.2 for critical values.

Example 1.3 (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.

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 1.9 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 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 (1.17)) and that the variance goes to zero as the sample size goes to infinity (as in (1.16)). (This is also called convergence in mean square.) To see the LLN in action, see Figure 1.10.

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 1.10 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 be applied.

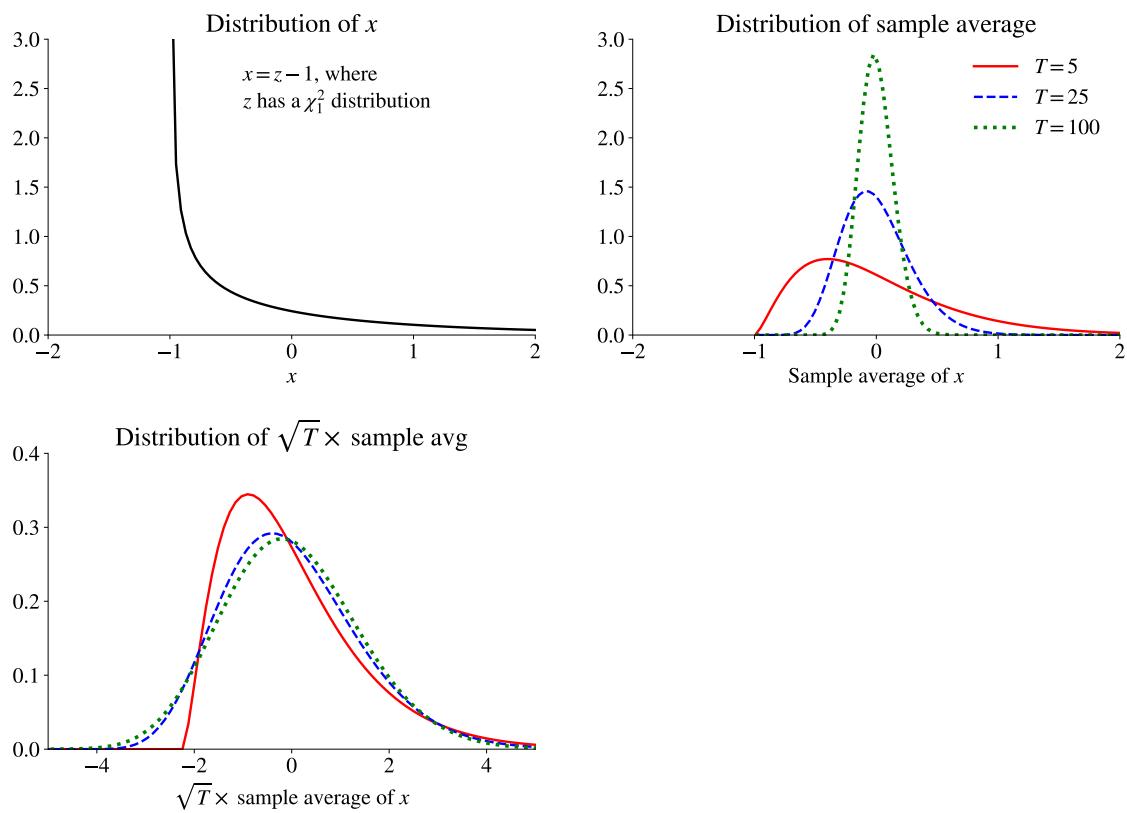


Figure 1.10: Sampling distributions

1.5 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.

1.6 Appendix: Data Sources

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

1. website of Kenneth French,
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
2. Datastream
3. Federal Reserve Bank of St. Louis (FRED), <http://research.stlouisfed.org/fred2/>
4. website of Robert Shiller, <http://www.econ.yale.edu/~shiller/data.htm>
5. yahoo! finance, <http://finance.yahoo.com/>
6. OlsenData, <http://www.olsendata.com>

<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 chi-square distribution (different degrees of freedom, n).

<u>n_1</u>	<u>$n_2, 10-300$</u>					<u>$\chi^2_{n_1}/n_1$</u>
	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).

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

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	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
-2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
-2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
-2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
-2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
-2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
-2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
-2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
-2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
-2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
-2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
-1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
-1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
-1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537
-1.6	0.0548	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
-1.5	0.0668	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
-1.4	0.0808	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
-1.3	0.0968	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
-1.2	0.1151	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
-1.1	0.1357	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
-1.0	0.1587	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
-0.9	0.1841	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
-0.8	0.2119	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
-0.7	0.2420	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709
-0.6	0.2743	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
-0.5	0.3085	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
-0.4	0.3446	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783
-0.3	0.3821	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
-0.2	0.4207	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
-0.1	0.4602	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960

Table 1.5: Values of the standard normal 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	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table 1.6: Values of the standard normal 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

Reference: Verbeek (2012) 2 and 4

More advanced material is denoted by a star (*). It is not required reading.

2.1 Least Squares: The Optimization Problem and Its Solution

2.1.1 Simple Regression

The simplest 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 . In principle, the residual should account for all the movements in y_t that we cannot explain 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 $y_i = a + b x_i + u_i$.

Notice the two very important assumptions: (i) the mean of the residual is zero; and (ii) the residual is not correlated with the regressor, x_t . This basically says that the residual is 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 $\beta_0 + \beta_1 x_t$.

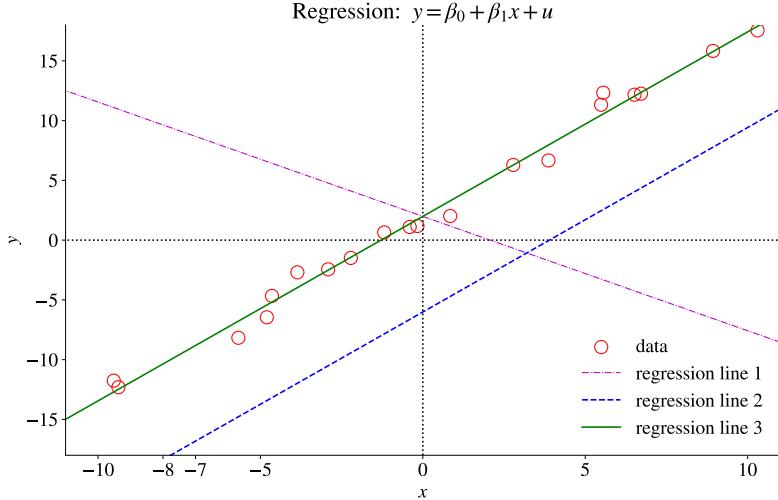


Figure 2.1: Example of OLS

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

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

by choosing b_0 and b_1 to make the loss function value as small as possible. The objective is thus to pick values of b_0 and b_1 in order to make the model fit the data as closely as possible—where close is taken to be a small variance of the unexplained part (the residual). See Figures 2.1–2.2 for illustrations. (Least squares is only one of many possible ways to estimate regression coefficients. We will discuss other methods later on.)

Remark 2.2 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 clearly $\hat{\beta}_i$.

Remark 2.3 (First order condition for minimizing a differentiable function). We want to find the value of b in the interval $b_{low} \leq b \leq b_{high}$, which makes the value of the differentiable function $f(b)$ as small as possible. The answer is b_{low} , b_{high} , or the value of b where $df(b)/db = 0$. See Figure 2.3.

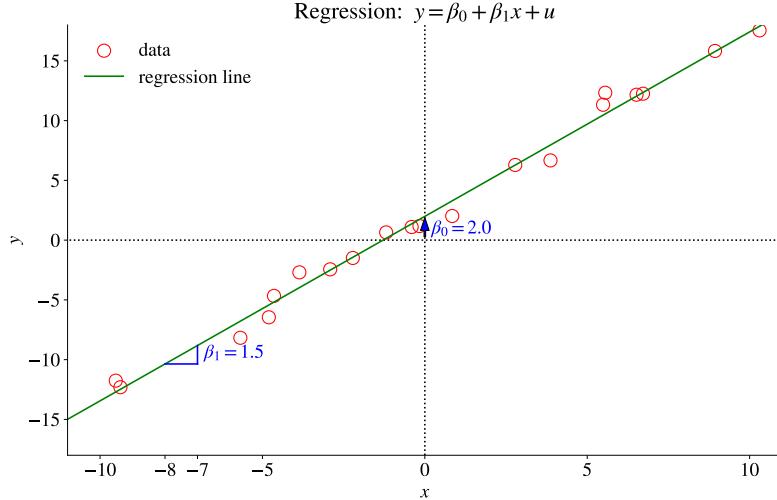


Figure 2.2: Example of OLS

The first order conditions for a minimum are that the derivatives of this loss function with respect to b_0 and b_1 should be zero. Notice that

$$\frac{\partial}{\partial b_0} (y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)1 \quad (2.3)$$

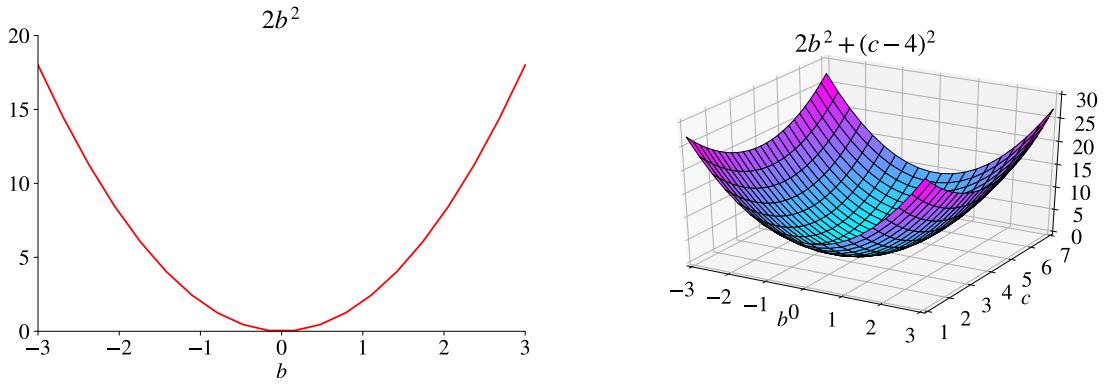
$$\frac{\partial}{\partial b_1} (y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)x_t. \quad (2.4)$$

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

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

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

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 is indeed a defining feature of LS and can be seen as the sample analogues of the assumptions in (2.1) that $E u_t = 0$ and $Cov(x_t, u_t) = 0$. To see this, note that (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



The minimum is where $\frac{d^2b^2}{db} = 4b = 0$

The minimum is where
 $\frac{\partial 2b^2}{\partial b} = 4b = 0$ and $\frac{\partial(c-4)^2}{\partial c} = 2(c-4) = 0$

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

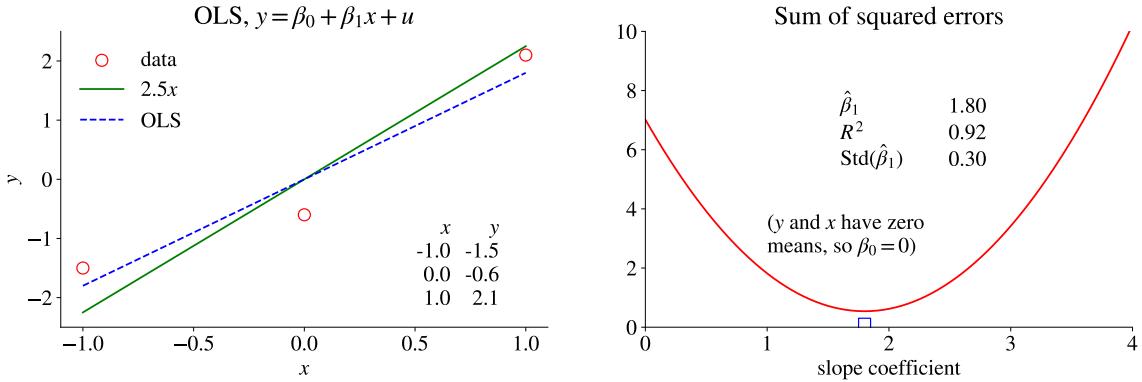


Figure 2.4: Example of OLS estimation

sample mean (see Remark 2.4).

Remark 2.4 (*Cross moments and covariance*) A covariance is defined as

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

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.

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

$$\begin{aligned}\sum_{t=1}^T y_t x_t &= \hat{\beta}_1 \sum_{t=1}^T x_t x_t \text{ or} \\ \hat{\beta}_1 &= \frac{\sum_{t=1}^T y_t x_t / T}{\sum_{t=1}^T x_t x_t / T}.\end{aligned}\tag{2.7}$$

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—just more tedious to show it).

See Table 2.1 for an empirical example.

	HiTec	Utils
constant	-0.13 (-0.94)	0.25 (1.73)
market return	1.27 (34.81)	0.51 (12.87)
R^2	0.75	0.32
obs	576.00	576.00

Table 2.1: CAPM regressions, monthly returns, %, US data 1970:01-2017:12. Numbers in parentheses are t-stats.

Example 2.5 (Simple regression) Consider the simple regression model (PSLS1). Suppose we have the following data

t	x	y
1	-1	-1.5
2	0	-0.6
3	1	2.1

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

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

The fitted residuals are

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

The fitted residuals indeed obey the first order condition (2.6) since

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

See Figure 2.4 for an illustration.

Example 2.6 Using the same data as in Example 2.5 we can also calculate the sums of squared residuals for different values of the slope coefficient. With $\beta_1 = 1.6$ we get

t	\underline{u}_t	\underline{u}_1^2
1	$-1.5 - \mathbf{1.6} \times (-1) = 0.1$	0.01
2	$-0.6 - \mathbf{1.6} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{1.6} \times 1 = 0.5$	0.25
sum	0	0.62

With $\beta = 1.8$ and $\beta = 2.0$ we instead get

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

t	\underline{u}_t	\underline{u}_1^2
1	$-1.5 - \mathbf{2.0} \times (-1) = 0.5$	0.25
2	$-0.6 - \mathbf{2.0} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{2.0} \times 1 = 0.1$	0.01
sum	0	0.62

Among these alternatives, $\beta = 1.8$ has the lowest sum of squared residuals (it is actually the optimum). See Figure 2.4.

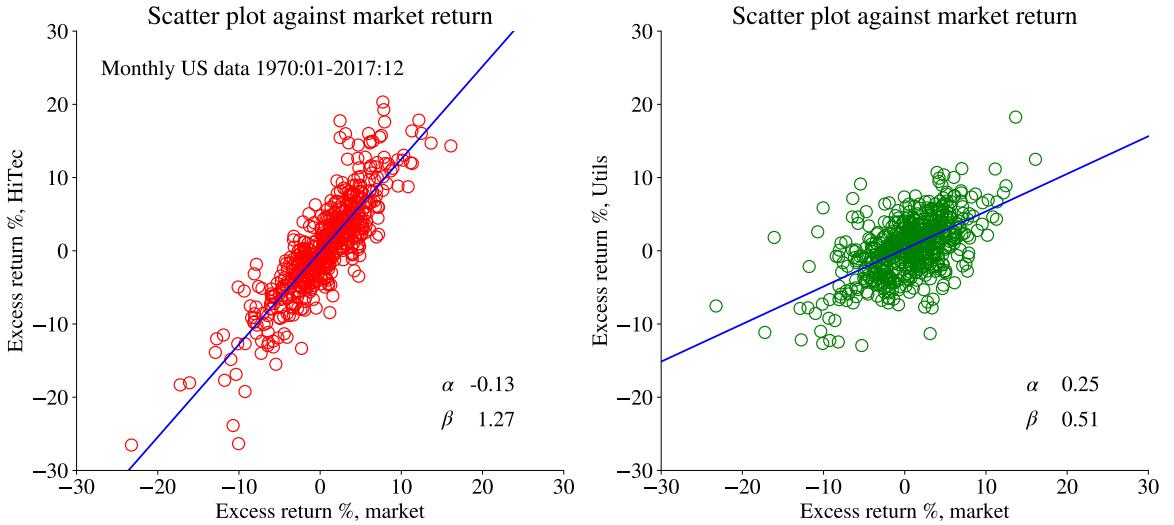


Figure 2.5: Scatter plot against market return

2.1.2 Multiple Regression

All the previous results still hold 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 (see Appendix 2.4 for a summary of matrix algebra). In this expression, one of the elements of x_t is typically a constant equal to one (and its coefficient the intercept).

Remark 2.7 (*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

$$\mathbf{0}_{kx1} = \sum_{t=1}^T x_t(y_t - x_t' \hat{\beta}) \text{ or } \sum_{t=1}^T x_t y_t = \sum_{t=1}^T x_t x_t' \hat{\beta}. \quad (2.10)$$

Solve this as

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

Remark 2.8 (*Alternative 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 $T \times 1$ where element t is y_t . Then, (2.11) can also be written $\hat{\beta} = (X'X)^{-1} X'Y$.

Example 2.9 With 2 regressors ($k = 2$), (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 2.10 (*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 2.5, 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 $\sum_{t=1}^T x_t x'_t$ 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

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

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 . For instance, with (2.1) it would be $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$. 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

$$R^2 = \frac{\text{Var}(\hat{y}_t)}{\text{Var}(y_t)} = 1 - \frac{\text{Var}(\hat{u}_t)}{\text{Var}(y_t)}, \quad (2.12)$$

which can also be rewritten as the squared correlation of the actual and fitted values

$$R^2 = \text{Corr}(y_t, \hat{y}_t)^2. \quad (2.13)$$

Notice that we must have constant in regression (unless both y_t and x_t have zero means) for R^2 to make sense.

Example 2.11 (R^2) From Example 2.5 we have $\text{Var}(\hat{u}_t) = 0.18$ and $\text{Var}(y_t) = 2.34$, so

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

See Figure 2.4.

Proof. (of (2.12)–(2.13)) 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

$$\text{Var}(y_t) = \text{Var}(\hat{y}_t) + \text{Var}(\hat{u}_t).$$

R^2 is defined as the fraction of $\text{Var}(y_t)$ that is explained by the model

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

Equivalently, we can rewrite R^2 by noting that

$$\text{Cov}(y_t, \hat{y}_t) = \text{Cov}(\hat{y}_t + \hat{u}_t, \hat{y}_t) = \text{Var}(\hat{y}_t).$$

Use this in the denominator of R^2 and multiply by $\text{Cov}(y_t, \hat{y}_t) / \text{Var}(\hat{y}_t) = 1$

$$R^2 = \frac{\text{Cov}(y_t, \hat{y}_t)^2}{\text{Var}(y_t) \text{Var}(\hat{y}_t)} = \text{Corr}(y_t, \hat{y}_t)^2.$$

■

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 $\hat{\beta}_1 = 0$. As a consequence $\hat{y}_t = \hat{\beta}_0$, which is a constant. This means that R^2 in (2.12) 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 (2.13) is also zero, since a constant is

always uncorrelated with anything else (as correlations measure comovements around the means). See Figure 2.6 for an example.

Remark 2.12 (R^2 from simple regression*) Suppose $\hat{y}_t = \beta_0 + \beta_1 x_t$, so (2.13) becomes

$$R^2 = \frac{\text{Cov}(y_t, \beta_0 + \beta_1 x_t)^2}{\text{Var}(y_t) \text{Var}(\beta_0 + \beta_1 x_t)} = \frac{\text{Cov}(y_t, x_t)^2}{\text{Var}(y_t) \text{Var}(x_t)} = \text{Corr}(y_t, x_t)^2.$$

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, $\bar{R}^2 = 1 - (1 - R^2)(T - 1)/(T - k)$, where k is the number of regressors (including the constant). This measure can be negative.

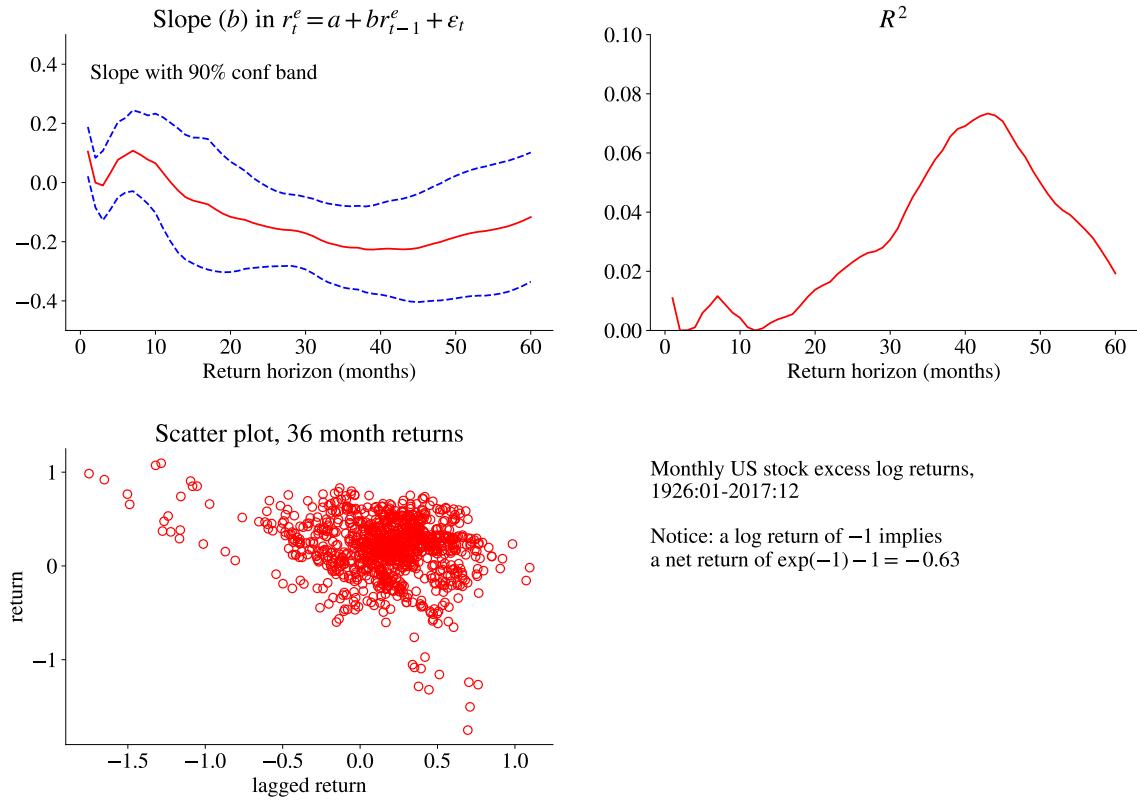


Figure 2.6: Prediction equations for US stock returns

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

Note that the estimated coefficients are random variables since they depend on which particular sample that has been “drawn.” This means that we cannot be sure that the estimated

coefficients are equal to the true coefficients (β_0 and β_1 in (2.1)). We can calculate an estimate of this uncertainty in the form of variances and covariances of $\hat{\beta}_0$ and $\hat{\beta}_1$. These can be used for testing hypotheses about the coefficients, for instance, that $\beta_1 = 0$.

To see where the uncertainty comes from, consider the simple case of only one regressor and a zero constant in (2.7). Use (2.1) to substitute for y_t (recall $\beta_0 = 0$)

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^T x_t (\beta_1 x_t + u_t) / T}{\sum_{t=1}^T x_t x_t / T} \\ &= \beta_1 + \frac{\sum_{t=1}^T x_t u_t / T}{\sum_{t=1}^T x_t x_t / T},\end{aligned}\tag{2.14}$$

so the OLS estimate, $\hat{\beta}_1$, equals the true value, β_1 , plus the sample covariance of x_t and u_t divided by the sample variance of x_t . Since u_t is a random variable, $\hat{\beta}_1$ is too. Clearly, we do not know the true value β_1 , so this decomposition is just conceptual.

One of the basic assumptions in (2.1) is that the covariance of the regressor and the residual is zero. This should hold in a very large sample (or else OLS cannot be used to estimate β_1), but in a small sample it may be different from zero. Only as the sample gets very large can we be (almost) sure that the second term in (2.14) vanishes.

Equation (2.14) will give different values of $\hat{\beta}$ when we use different samples, that is different draws of the random variables x_t and y_t (and thus u_t). Since the true value, β , is a fixed constant, the distribution of these estimates across samples would describe the uncertainty we should have about the true value after having obtained a specific estimated value. However, we cannot observe this distribution directly (we do not have a lot of different samples). However, we can use the idea of this distribution to discuss the general properties of OLS—and we can (with some added assumptions) provide a good estimate of how that distribution could look like.

The first conclusion from (2.14) 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 $\hat{\beta}_1$ —in contrast to a large sample where the randomness in $\sum_{t=1}^T x_t u_t / T$ is averaged out more effectively (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 at least independent of the residuals in all periods) and then assume something about the

distribution of the residuals; or (iii) use the asymptotic (large sample) distribution as an approximation. The asymptotic distribution can often be derived, in contrast to the exact distribution in a sample of a given size. If the actual sample is large, then the asymptotic distribution may be a good 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 we are looking at, including the specific parameter values). See Figures 2.7, 2.9 and 2.11 for examples.

In contrast, asymptotic theory give more general results—but arriving there is hard. Treating the regressors as constants is easier—and is often good enough for illustrating the main properties of the estimation method.

When there are several regressors (x_t is a vector with k elements), then (2.14) becomes an expression for the vector

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

where $\Sigma x_t x_t'$ is a $k \times k$ matrix and $\Sigma x_t u_t$ is a $k \times 1$ vector.

The typical outcome of all three approaches will (under strong assumptions) be that

$$\hat{\beta} \sim N \left[\beta, \left(\sum_{t=1}^T x_t x_t' \right)^{-1} \sigma^2 \right], \quad (2.16)$$

which allows for x_t to be a vector with k elements. Clearly, with $k = 1$, x_t is a scalar (and $x_t' = x_t$). In practice, we calculate/estimate both $\sum_{t=1}^T x_t x_t'$ and σ^2 from the available data (the latter as the variance of the fitted residuals). See Table 2.1 for an empirical example and Figure 2.7 for an illustration of how the results depend on σ and the standard deviation of x_t .

Remark 2.13 (*Alternative 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 (2.16) can also be written $(X'X)^{-1}\sigma^2$.

An alternative way of expressing the distribution (often used in conjunction with asymptotic) theory is

$$\sqrt{T}(\hat{\beta} - \beta) \sim N \left[0, \left(\sum_{t=1}^T x_t x_t' / T \right)^{-1} \sigma^2 \right]. \quad (2.17)$$

This is the same as (2.16). (To see that, divide the LHS of (2.17) by \sqrt{T} . Then, the

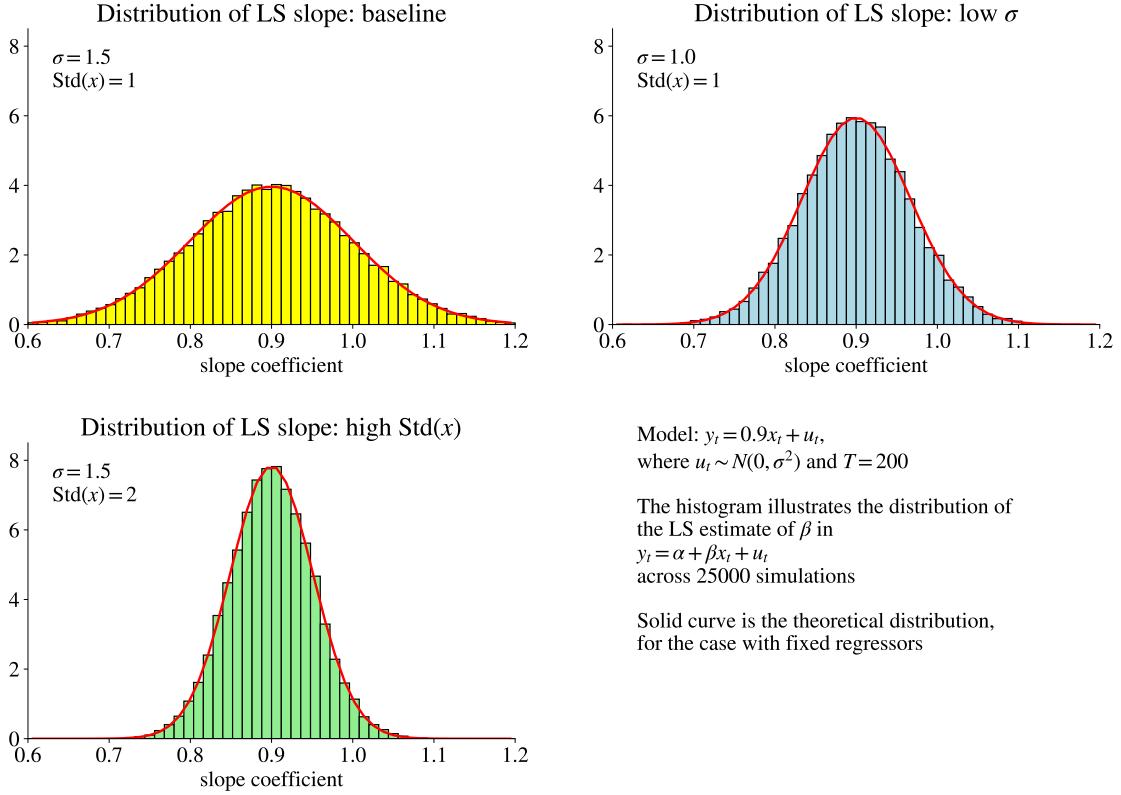


Figure 2.7: Distribution of OLS estimate, from simulation and theory

variance on the RHS must be divided by T , which gives the same variance as in (2.16). Then, add β to the LHS, which changes the mean on the RHS to β . We then have (2.16).)

Example 2.14 (*Distribution of slope coefficient*) From Example 2.5 we have $\text{Var}(\hat{u}_t) = \sigma^2 = 0.18$ 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 2.15 (*Covariance matrix of b_1 and b_2*) From Example 2.10

$$\begin{aligned} \sum_{t=1}^T x_t x_t' &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \text{ and } \sigma^2 = 0.18, \text{ then} \\ \text{Var}\left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}\right) &= \begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) \end{bmatrix} \\ &= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} 0.18 = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.09 \end{bmatrix}. \end{aligned}$$

The standard deviations (also called standard errors) are therefore

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

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

The assumption of fixed regressors makes a lot of sense in controlled experiments, where we actually can generate different samples with the same values of the regressors (the heat or whatever). It makes much less sense in econometrics. However, it is easy to derive results for this case—and those results happen to be very similar to what asymptotic theory gives.

The results we derive 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 zero mean iid variables*. (As an alternative to assuming fixed regressors (as we do here), it can instead be assumed that the residuals and regressors are independent. This delivers very similar results.) We will also assume that the residuals are normally distributed (not part of the typical Gauss-Markov assumptions).

For notational convenience, write (2.15) 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 non-random, the expected value of this expression is

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

since we always assume that the residuals have zero means (see (2.1)). The interpretation is that we can expect OLS to give (on average) a correct answer. That is, if we could draw many different samples and estimate the slope coefficient in each of them, then the average of those estimates would be the correct number (β). Clearly, this is something we want from an estimation method (a method that was systematically wrong would not be very attractive).

Remark 2.16 (*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 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 (with fixed regressors x_t and S_{xx}^{-1} can be treated as constants). It is straightforward to see that the mean of this normal distribution is β (the true value), since the rest is a linear combination of the residuals—and they all have a zero mean.

Finding the variance-covariance matrix of $\hat{\beta}$ is just slightly more complicated. Remember that we treat x_t as fixed numbers (“constants”) 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 2.17 (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} .

There are three main ways of getting a low uncertainty (low $\text{Var}(\hat{\beta})$). For simplicity, focus on the case with just one regressor. We then have the following results. 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 $S_{xx} = \sum_{t=1}^T x_t x'_t$) 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 Figure 2.8.

Example 2.18 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.)

Example 2.19 When the regressor is a zero mean variable, then we have

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

The variance is increasing in σ^2 , but decreasing in both T and $\text{Var}(x_t)$.

Example 2.20 When the regressor is just a constant (equal to one) and one variable regressor with zero mean, f_t , so $x_t = [1, f_t]'$, then we have

$$\begin{aligned} \sum_{t=1}^T x_t x'_t &= \sum_{t=1}^T \begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} = T \begin{bmatrix} 1 & 0 \\ 0 & \text{Var}(f_t) \end{bmatrix}, \text{ so} \\ \text{Var} \left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \right) &= \sigma^2 \left(\sum_{t=1}^T x_t x'_t \right)^{-1} = \begin{bmatrix} \sigma^2 / T & 0 \\ 0 & \sigma^2 / [\text{Var}(f_t)T] \end{bmatrix}. \end{aligned}$$

A combination of the two previous examples.

A key assumption in regression analysis is that our sample is “representative” of the population. In practice, this means that we can estimate both S_{xx} and σ^2 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). It is perhaps a stronger assumption when the sample is a time series of data. Then we basically 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 for you.

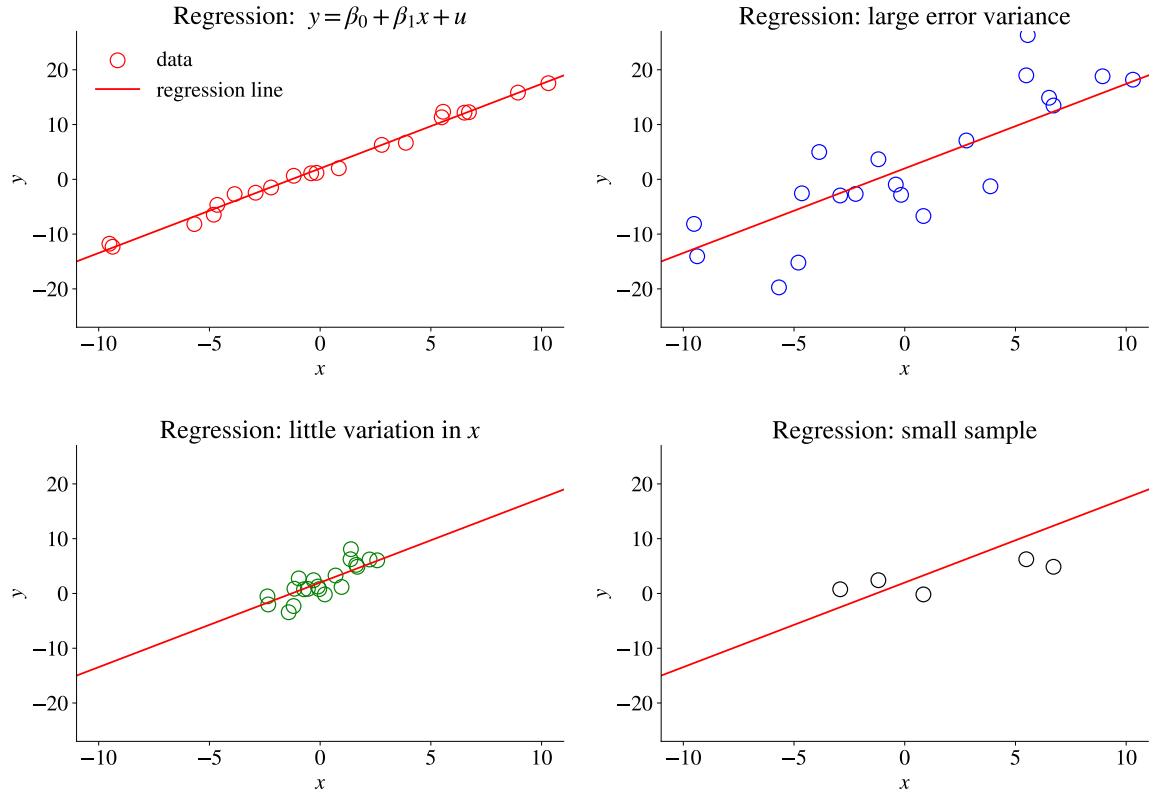


Figure 2.8: Regressions: importance of error variance and variation of regressor

2.2.2 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 of

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

where the new term is the (squared) correlation. 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.

Proof. (of 2.23*). 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

$$\begin{bmatrix} \sum_{t=1}^T x_{1t}^2 & \sum_{t=1}^T x_{1t}x_{2t} \\ \sum_{t=1}^T x_{1t}x_{2t} & \sum_{t=1}^T x_{2t}^2 \end{bmatrix}^{-1} = \frac{1}{\sum_{t=1}^T x_{1t}^2 \sum_{t=1}^T x_{2t}^2 - (\sum_{t=1}^T x_{1t}x_{2t})^2} \begin{bmatrix} \sum_{t=1}^T x_{2t}^2 & -\sum_{t=1}^T x_{1t}x_{2t} \\ -\sum_{t=1}^T x_{1t}x_{2t} & \sum_{t=1}^T x_{1t}^2 \end{bmatrix}.$$

The variance of the second slope coefficient is σ^2 time the lower right element of this matrix. Multiply and divide by T to get

$$\begin{aligned} \text{Var}(\hat{\beta}_2) &= \frac{\sigma^2}{T} \frac{\sum_{t=1}^T x_{1t}^2 / T}{\sum_{t=1}^T \frac{1}{T} x_{1t}^2 \sum_{t=1}^T \frac{1}{T} x_{2t}^2 - \left(\sum_{t=1}^T \frac{1}{T} x_{1t} x_{2t} \right)^2} \\ &= \frac{\sigma^2}{T} \frac{\text{Var}(x_{1t})}{\text{Var}(x_{1t}) \text{Var}(x_{2t}) - \text{Cov}(x_{1t}, x_{2t})^2} \\ &= \frac{\sigma^2}{T} \frac{1 / \text{Var}(x_{2t})}{1 - \frac{\text{Cov}(x_{1t}, x_{2t})^2}{\text{Var}(x_{1t}) \text{Var}(x_{2t})}}, \end{aligned}$$

which is the same as (2.23). ■

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

2.3.1 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.

First, the Gauss-Markov assumptions of iid residuals (constant volatility and no correlation across observations) are likely to be false in many cases. These issues (heteroskedasticity and autocorrelation) are therefore discussed at length later on.

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. This can make OLS biased in small samples, although the OLS estimate might converge to the true values (so OLS is “consistent”) as the sample size increases.

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). This is discussed in some detail below.

The next few sections discuss the 2nd and 3rd issue, while the first issue is treated in a later chapter.

2.3.2 Bias

If an estimation method is *biased*, then it produces systematically wrong (say, too low) coefficients in small samples.

Figure 2.9 illustrates some simulation results from estimating an AR(1)

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t \quad (2.24)$$

on artificially generated samples where “data” follow

$$y_t = 0.9y_{t-1} + u_t, \text{ where } u_t \text{ is iid.} \quad (2.25)$$

The figure suggests that the estimates are biased (not centered on the true value) in small samples.

To understand these results, recall that (2.14) says that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^T x_t u_t / T}{\sum_{t=1}^T x_t x_t / T}, \quad (2.26)$$

where u_t are the true residuals. We will never observe the true residuals, so (10.11) can only be used for a conceptual discussion.

To get *unbiased estimates*, ($E \hat{\beta}_1 = \beta_1$), the second term of the right hand side of (10.11) 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

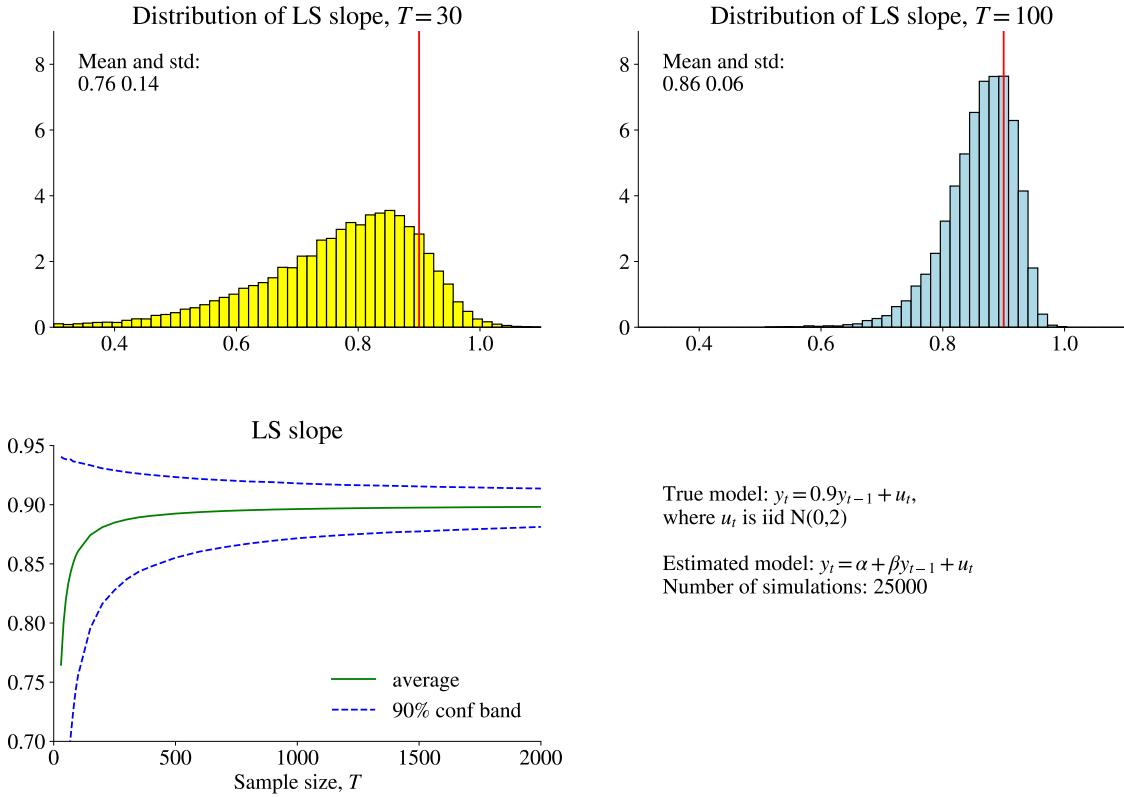


Figure 2.9: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

the numerator and denominator. This is probably most easily investigated by (Monte Carlo) simulations. In many cases, the bias decreases rapidly as the sample size increases (see the discussion of “consistency”).

2.3.3 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 2.9 suggests that the problem with the AR(1) estimation vanishes as the sample size increases. This suggests the importance of doing simulations (to understand the properties of the estimation method)—and of using large data sets.

To get *consistent estimates* (which is defined as the bias and the variance of $\hat{\beta}_1$ go to zero as $T \rightarrow \infty$), then it is enough if x_t and u_t (in the same period) are uncorrelated. This is indeed the case in the AR(1) simulations discussed before. To see this from (10.11),

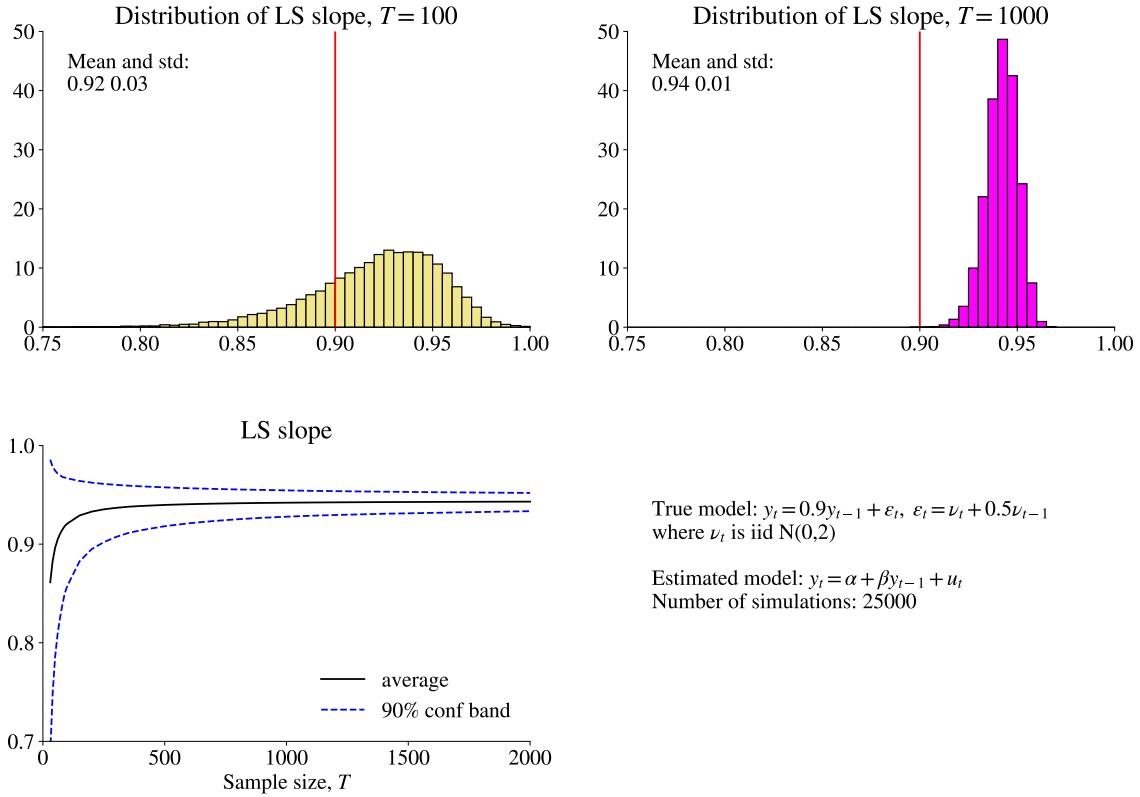


Figure 2.10: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

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.)

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 $\hat{\beta}_1$ 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 2.10. It estimates the same AR(1) as in (2.24) but where the simulated “data” now follows

$$y_t = 0.9y_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t = v_t + 0.5v_{t-1} \text{ where } v_t \text{ is iid.} \quad (2.27)$$

In this case, the residuals (here called ε_t) are themselves autocorrelated. The figure clearly shows that the OLS estimate of the slope β_1 in (2.24) does *not* converge to 0.9 as the sample sizes increases: OLS is inconsistent. The reason in this case is that ε_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 be sure 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.

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.

2.3.4 Normality

If the regressors x_t are fixed numbers and u_t is normally distributed, then the second term in (10.11) shows that the normality carries over to $\hat{\beta}$. Actually, we can relax the assumption about the regressors (to allow them to be random) as long as we assume that x_t and u_t are independent (the same assumption as needed for unbiasedness). We can test the assumption of normally distributed residuals by using a Bera-Jarque test

$$BJ = \frac{T}{6} \text{skewness}^2 + \frac{T}{24} (\text{kurtosis} - 3)^2, \quad (2.28)$$

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

Even if the normality test fails, we can often still hope for a (close to) normal distribution of $\hat{\beta}_1$ if the sample is large—due to a central limit theorem. This is illustrated in Figure 2.11. It is based on simulations where the residual is drawn from a very non-normal distribution. For a small sample, this carries over to $\hat{\beta}_1$ and the t -stat for the hypothesis that $\beta_1 = 0$. However, already a moderately sized sample tend to give an almost normal distribution.

To understand the theory of this rewrite (10.11) by subtracting β_1 from both sides and then multiply both sides by \sqrt{T}

$$\sqrt{T}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{T}\sum_{t=1}^T x_t u_t / T}{\sum_{t=1}^T x_t x_t / T}. \quad (2.29)$$

The denominator is the sample average of $x_t x_t$ which will converge to a fixed number (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 $x_t u_t$. Recall that, under weak conditions, a central limit theorem applies so \sqrt{T} times a sample average: it typically converges to a normal distribution.

This shows that $\sqrt{T}\hat{\beta}_1$ has an *asymptotic normal distribution*. This often holds as a reasonable approximation also in moderately sized samples.

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 properties of this distribution are quite similar to those that we derived by assuming that the regressors were fixed numbers.

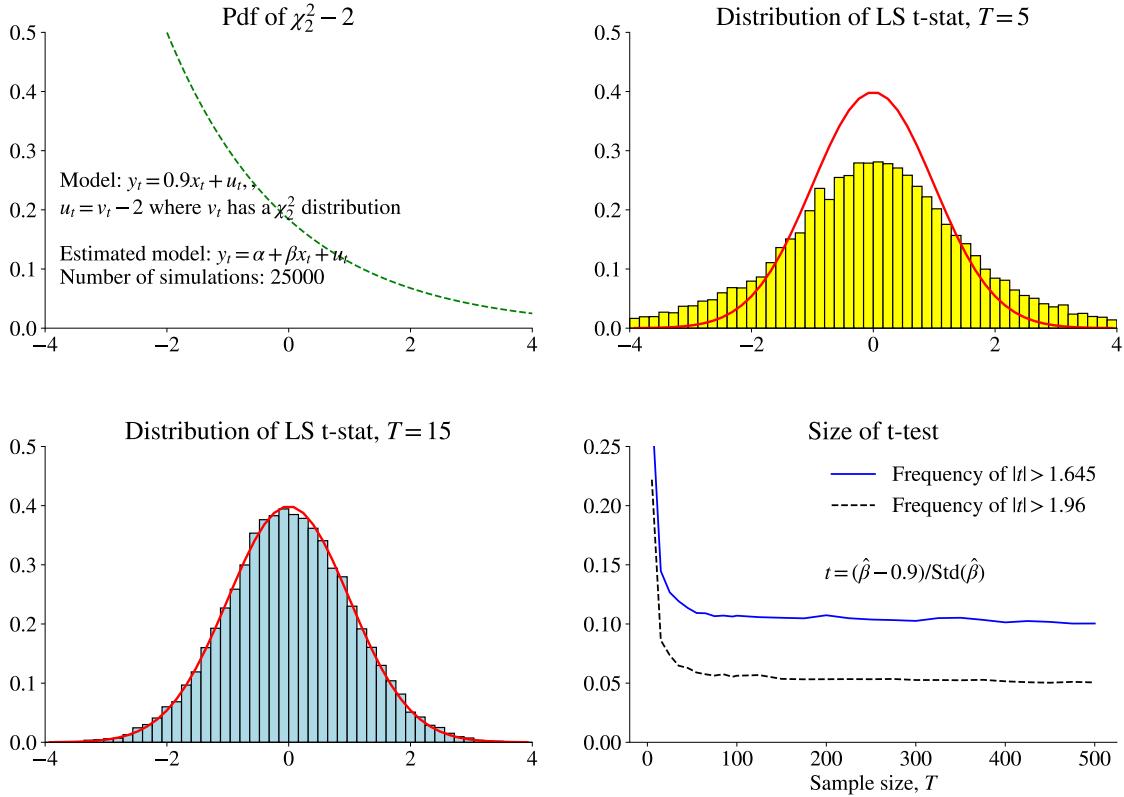


Figure 2.11: Distribution of LS estimator when residuals have a non-normal distribution

2.4 Appendix: A Primer in Matrix Algebra

Let c be a scalar and define the matrices

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

Adding/subtracting a scalar to a matrix or multiplying a matrix by a scalar are both element by element

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

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

Example 2.21

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

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

Matrix *addition* (or subtraction) 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 2.22 (Matrix addition and subtraction)

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$$

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

To turn a column into a row vector, use the *transpose* operator like in x'

$$x' = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}' = [x_1 \ x_2].$$

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 2.23 (Matrix transpose)

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' = [10 \ 11]$$

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

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 2.24 (*Matrix multiplication*)

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

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

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

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

and xz' the *outer product*

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

(Notice that xz does not work). If x is a column vector and A a square matrix, then the product $x'Ax$ is a *quadratic form*.

Example 2.25 (*Inner product, outer product and quadratic form*)

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

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

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

A matrix *inverse* is the closest we get to “dividing” by a 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 zeroes elsewhere). The matrix inverse is useful for solving systems of linear equations, $y = Ax$ as $x = A^{-1}y$.

Example 2.26 (*Matrix inverse*) We have

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

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

Chapter 3

Index Models

Reference: Elton, Gruber, Brown, and Goetzmann (2010) 7–8, 11

3.1 Single-Index Models

The *single-index model* is a way to cut down on the number of parameters that we need to estimate in order to construct the covariance matrix of assets. The model assumes that the co-movement between assets is due to a single common influence (here denoted R_{mt})

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

In this equation, R_{it} stands for the return of asset i in period t , while R_{mt} is the market return in the same period. This regression may use the net returns (as indicated above), or the returns in excess of a riskfree rate. The results for the β (which is the focus here) are typically very similar.

The first two assumptions are the standard assumptions for using Least Squares: the residual has a zero mean and is uncorrelated with the non-constant regressor. (Together they imply that the residuals are orthogonal to both regressors, which is the standard assumption in econometrics.) Hence, these two properties will be automatically satisfied if (3.1) is estimated by Least Squares.

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.

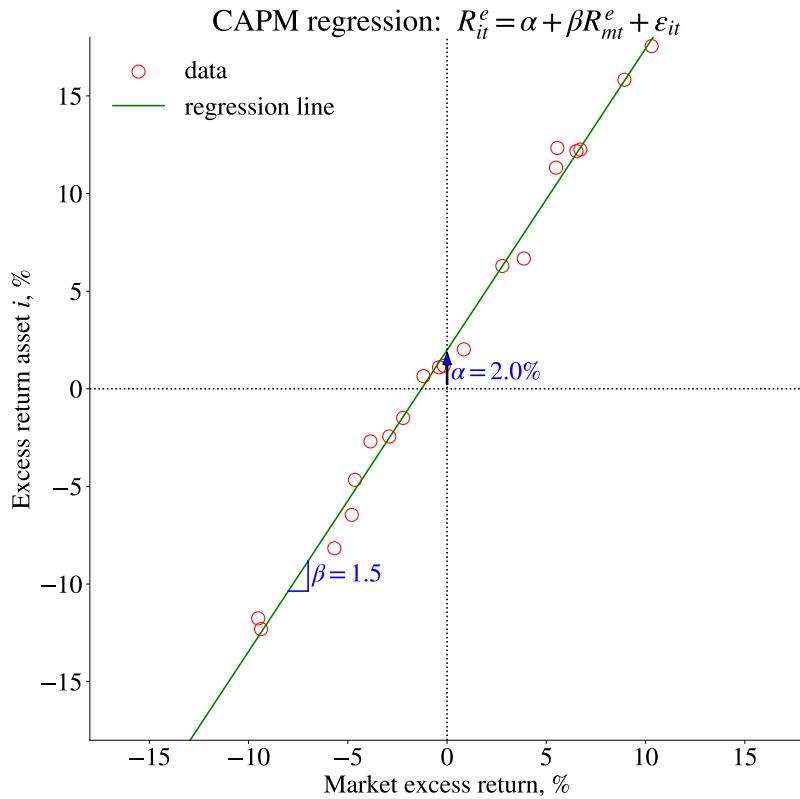


Figure 3.1: CAPM regression

See Figures 3.1 – 3.3 for illustrations.

The key point of the model, however, is the third assumption: the residuals for different assets are uncorrelated. This means that all comovements of two assets (R_{it} and R_{jt} , say) are due to movements in the common “index” R_{mt} . This is not at all guaranteed by running LS regressions—just an assumption. It is likely to be false—but may be a reasonable approximation. In any case, it simplifies the construction of the covariance matrix of the assets enormously—as demonstrated below.

Remark 3.2 (*The market model*) *The market model is (3.1) without the assumption that $\text{Cov}(\varepsilon_{it}, \varepsilon_{jt}) = 0$. This model does not simplify the calculation of a portfolio variance—but will turn out to be important when we want to test CAPM.*

If (3.1) is true, then the variance of asset i and the covariance of assets i and j are

$$\sigma_{ii} = \beta_i^2 \text{Var}(R_{mt}) + \text{Var}(\varepsilon_{it}) \quad (3.2)$$

$$\sigma_{ij} = \beta_i \beta_j \text{Var}(R_{mt}). \quad (3.3)$$

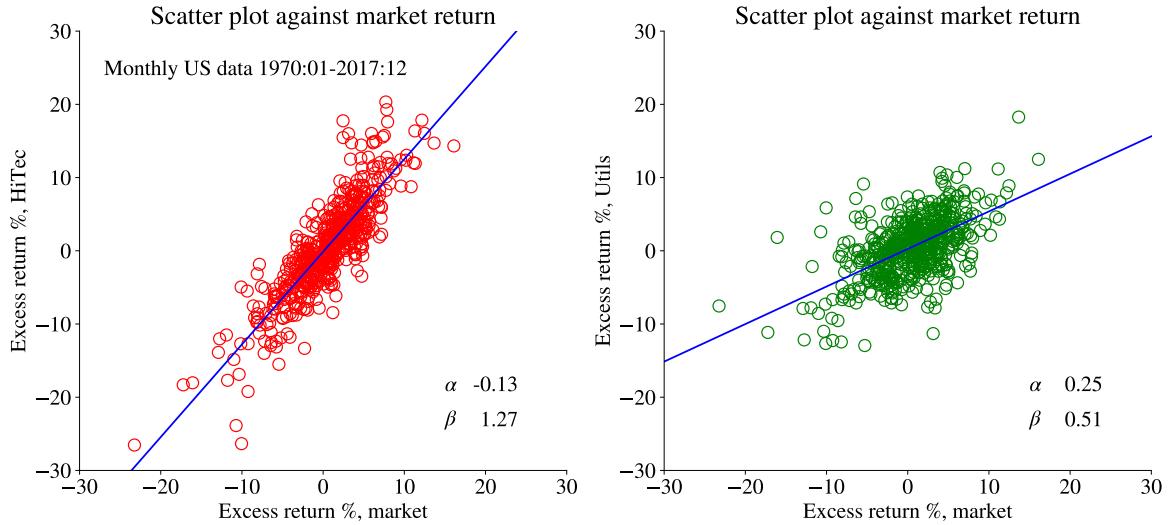


Figure 3.2: Scatter plot against market return

Together, these equations show that we can calculate the whole covariance matrix by having just the variance of the index (to get $\text{Var}(R_{mt})$) and the output from n regressions (to get β_i and $\text{Var}(\varepsilon_{it})$ for each asset). This is, in many cases, much easier to obtain than direct estimates of the covariance matrix. For instance, a new asset does not have a return history, but it may be possible to make intelligent guesses about its beta and residual variance (for instance, from knowing the industry and size of the firm).

This gives the covariance matrix (for two assets)

$$\text{Cov} \begin{pmatrix} R_{it} \\ R_{jt} \end{pmatrix} = \begin{bmatrix} \beta_i^2 & \beta_i \beta_j \\ \beta_i \beta_j & \beta_j^2 \end{bmatrix} \text{Var}(R_{mt}) + \begin{bmatrix} \text{Var}(\varepsilon_{it}) & 0 \\ 0 & \text{Var}(\varepsilon_{jt}) \end{bmatrix}, \text{ or} \quad (3.4)$$

$$= \begin{bmatrix} \beta_i \\ \beta_j \end{bmatrix} \begin{bmatrix} \beta_i & \beta_j \end{bmatrix} \text{Var}(R_{mt}) + \begin{bmatrix} \text{Var}(\varepsilon_{it}) & 0 \\ 0 & \text{Var}(\varepsilon_{jt}) \end{bmatrix} \quad (3.5)$$

More generally, with n assets we can define β to be an $n \times 1$ vector of all the betas and Σ to be an $n \times n$ matrix with the variances of the residuals along the diagonal. We can then write the covariance matrix of the $n \times 1$ vector of the returns as

$$\text{Cov}(R_t) = \beta \beta' \text{Var}(R_{mt}) + \Sigma. \quad (3.6)$$

See Figure 3.4 for an example based on the Fama-French portfolios detailed in Table 3.2.

	HiTec	Utils
constant	-0.13 (-0.94)	0.25 (1.73)
market return	1.27 (34.81)	0.51 (12.87)
R^2	0.75	0.32
Autocorr	0.39	0.79
White	0.03	0.00
All Slopes	0.00	0.00
obs	576.00	576.00

Table 3.1: CAPM regressions, monthly returns, %, US data 1970:01-2017:12. 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.

Remark 3.3 (*Fama-French portfolios*) *The portfolios in Table 3.2 are calculated by annual rebalancing (June/July). The US stock market is divided into 5×5 portfolios as follows. First, split up the stock market into 5 groups based on the book value/market value: put the lowest 20% in the first group, the next 20% in the second group etc. Second, split up the stock market into 5 groups based on size: put the smallest 20% in the first group etc. Then, form portfolios based on the intersections of these groups (also called double sorting). For instance, in Table 3.2 the portfolio in row 2, column 3 (portfolio 8) belong to the 20%-40% largest firms and the 40%-60% firms with the highest book value/market value.*

		Book value/Market value				
		1	2	3	4	5
1	1	1	2	3	4	5
	2	6	7	8	9	10
	3	11	12	13	14	15
	4	16	17	18	19	20
	5	21	22	23	24	25

Table 3.2: Numbering of the FF portfolios.

Proof. (of (3.2)–(3.3)) By using (3.1) and recalling that $\text{Cov}(R_{mt}, \varepsilon_{it}) = 0$ direct

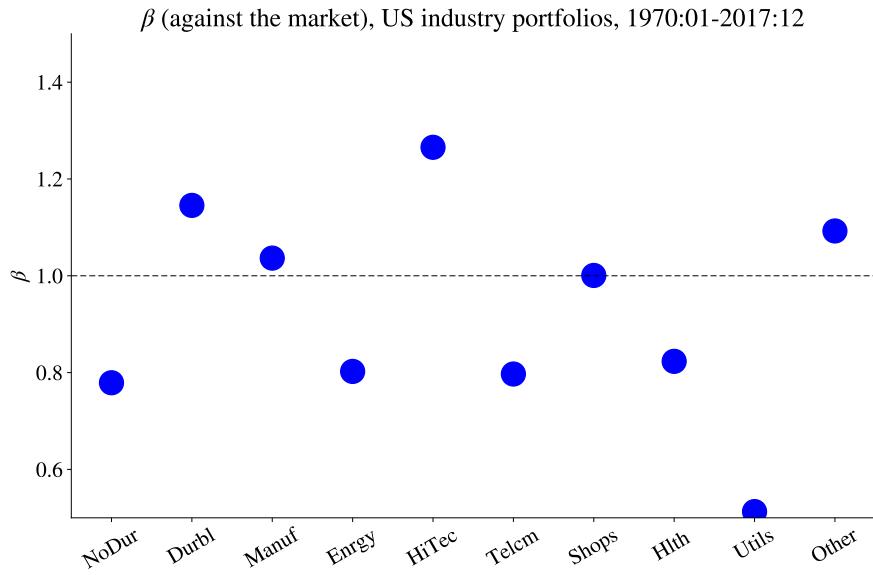


Figure 3.3: β s of US industry portfolios

calculations give

$$\begin{aligned}
 \sigma_{ii} &= \text{Var}(R_{it}) \\
 &= \text{Var}(\alpha_i + \beta_i R_{mt} + \varepsilon_{it}) \\
 &= \text{Var}(\beta_i R_{mt}) + \text{Var}(\varepsilon_{it}) + 2 \times 0 \\
 &= \beta_i^2 \text{Var}(R_{mt}) + \text{Var}(\varepsilon_{it}).
 \end{aligned}$$

Similarly, the covariance of assets i and j is (recalling also that $\text{Cov}(\varepsilon_{it}, \varepsilon_{jt}) = 0$)

$$\begin{aligned}
 \sigma_{ij} &= \text{Cov}(R_{it}, R_{jt}) \\
 &= \text{Cov}(\alpha_i + \beta_i R_{mt} + \varepsilon_{it}, \alpha_j + \beta_j R_{mt} + \varepsilon_{jt}) \\
 &= \beta_i \beta_j \text{Var}(R_{mt}) + 0 \\
 &= \beta_i \beta_j \text{Var}(R_{mt}).
 \end{aligned}$$

■

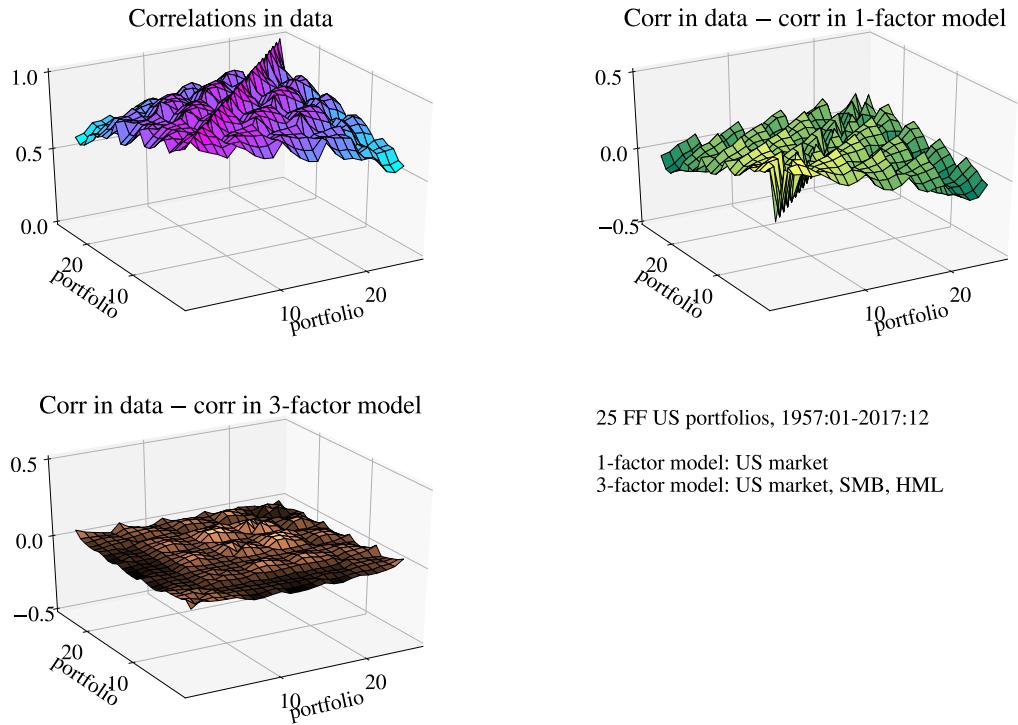


Figure 3.4: Correlations of US portfolios

3.2 Estimating Beta

3.2.1 Estimating Historical Beta: OLS and Other Approaches

Least Squares (LS) is typically used to estimate α_i , β_i and $\text{Std}(\varepsilon_{it})$ in (3.1)—and the R^2 is used to assess the quality of the regression.

Remark 3.4 (R^2 of market model) R^2 of (3.1) measures the fraction of the variance (of R_{it}) that is due to the systematic part of the regression, that is, relative importance of market risk as compared to idiosyncratic noise ($1 - R^2$ is the fraction due to the idiosyncratic noise)

$$R^2 = \frac{\text{Var}(\alpha_i + \beta_i R_{mt})}{\text{Var}(R_{it})} = \frac{\beta_i^2 \sigma_m^2}{\beta_i^2 \sigma_m^2 + \sigma_{ei}^2}.$$

To assess the accuracy of historical betas, Blume (1971) and others estimate betas for non-overlapping samples (periods)—and then compare the betas across samples. They find that the correlation of betas across samples is moderate for individual assets, but relatively high for diversified portfolios. It is also found that betas tend to “regress” towards

one: an extreme (high or low) historical beta is likely to be followed by a beta that is closer to one. There are several suggestions for how to deal with this problem.

To use *Blume's ad-hoc technique*, let $\hat{\beta}_{i1}$ be the estimate of β_i from an early sample, and $\hat{\beta}_{i2}$ the estimate from a later sample. Then regress

$$\hat{\beta}_{i2} = \gamma_0 + \gamma_1 \hat{\beta}_{i1} + \nu_i \quad (3.7)$$

and use it for forecasting the beta for yet another sample. Blume found $(\hat{\gamma}_0, \hat{\gamma}_1) = (0.343, 0.677)$ for his data.

Other authors have suggested averaging the OLS estimate ($\hat{\beta}_i$) with some average beta. For instance, $(\hat{\beta}_i + 1)/2$ since the average beta must be unity or $(\hat{\beta}_i + \sum_{j=1}^n \hat{\beta}_j / n)/2$, which will typically be similar since $\sum_{j=1}^n \hat{\beta}_j / n$ is likely to be close to one.

The *Bayesian approach* is another (more formal) way of adjusting the OLS estimate. It also uses a weighted average of the OLS estimate, $\hat{\beta}_i$, and some other number, β_0 , $(1 - F)\hat{\beta}_i + F\beta_0$ where F depends on the precision of the OLS estimator. The general idea of a Bayesian approach (Greene (2003) 16) is to treat both R_{it} and β_i as random. In this case a Bayesian analysis could go as follows. First, suppose our prior beliefs (before having data) about β_i is that it is normally distributed, $N(\beta_0, \sigma_0^2)$, where (β_0, σ_0^2) are some numbers. Second, run a LS regression of (3.1). If the residuals are normally distributed, so is the estimator—it is $N(\hat{\beta}_i, \sigma_{\beta_1}^2)$, where we have taken the point estimate to be the mean. If we treat the variance of the LS estimator (σ_{β}^2) as known, then the Bayesian estimator of beta is

$$b = (1 - F)\hat{\beta}_i + F\beta_0, \text{ where} \\ F = \frac{1/\sigma_0^2}{1/\sigma_0^2 + 1/\sigma_{\beta}^2} = \frac{\sigma_{\beta_1}^2}{\sigma_0^2 + \sigma_{\beta}^2}. \quad (3.8)$$

When the prior beliefs are very precise ($\sigma_0^2 \rightarrow 0$), then $F \rightarrow 1$ so the Bayesian estimator is the same as the prior mean. Effectively, when the prior beliefs are so precise, there is no room for data to add any information. In contrast, when the prior beliefs are very imprecise ($\sigma_0^2 \rightarrow \infty$), then $F \rightarrow 0$, so the Bayesian estimator is the same as OLS. Effectively, the prior beliefs do not add any information. In the current setting, $\beta_0 = 1$ and σ_0^2 taken from a previous (econometric) study might make sense.

3.2.2 Fundamental Betas

Another way to improve the forecasts of the beta over a future period is to bring in 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

$$R_{it} = a_i + b'_i I_t + \varepsilon_{it}, \text{ where} \quad (3.9)$$

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

As an example, there could be two indices: the stock market return and an interest rate. An ad-hoc approach is to first try a single-index model and then test if the residuals are approximately uncorrelated. If not, then adding a second index might improve the model.

It is often found that it takes several indices to get a reasonable approximation—but that a single-index model is equally good (or better) at “forecasting” the covariance over a future period. This is much like the classical trade-off between in-sample fit (requires a large model) and forecasting (often better with a small model).

The types of indices vary, but one common set captures the “business cycle” and includes things like the market return, interest rate (or some measure of the yield curve slope), GDP growth, inflation, and so forth. Another common set of indices are industry indices.

It turns out (see below) that the calculations of the covariance matrix are simpler if the indices are transformed to be uncorrelated.

Remark 3.5 (*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 riskfree return), since they are long-short portfolios.

3.3.2 Using the Multi-Index Model

If Ω is the covariance matrix of the indices, then the covariance of assets i and j is

$$\sigma_{ij} = b_i' \Omega b_j, \quad (3.10)$$

where b_i is the vector of slope coefficients obtained from regressing R_{it} on the vector of factors (I_t or I_t^*) as in (3.9). To get the variance of asset i , use the same formula but set $j = i$ and add the variance of the residuals, $\text{Var}(\varepsilon_{it})$.

In case the factors are uncorrelated, then Ω is diagonal so (3.10) can be simplified as

$$\sigma_{ij} = \sum_{k=1}^K b_{ik} b_{jk} \text{Var}(I_{kt}), \quad (3.11)$$

where b_{ik} is the coefficient on factor k in the regression of R_{it} .

See Figure 3.4 for an example.

3.3.3 Multi-Index Model as a Method for Portfolio Choice

The factor loadings (betas) can be used for more than just constructing the covariance matrix. In fact, the factor loadings 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 to understand the broad features of even complicated portfolios. Combined this with the fact that many analysts and investors have fairly little direct information about individual assets, but are often willing to form opinions about the future relative performance of different asset classes (small vs large firms, equity vs bonds, etc)—and the role for factor loadings becomes clear.

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 combinations of z_t that account for as much of the variability as possible—and its variance is denoted 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

$$W = [w_1 \ \cdots \ w_n]. \quad (3.12)$$

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

$$pc_t = W' z_t. \quad (3.13)$$

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

$$z_t = W pc_t. \quad (3.14)$$

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 $-pc_{jt}$).

Example 3.7 (*PCA with 2 series*) Let $w_i^{(j)}$ be the t th element in the i th eigenvector. With

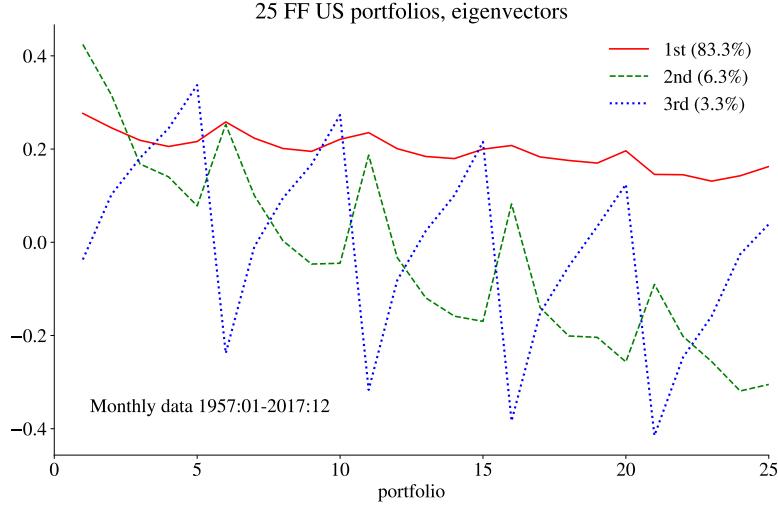


Figure 3.5: Eigenvectors for US portfolio returns

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 3.8 (Data in matrices*) Transpose (3.13) 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 (3.14) 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

$$\text{relative importance of } pc_j = \lambda_j / \sum_{i=1}^n \lambda_i. \quad (3.15)$$

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

Chapter 4

Least Squares: Testing

Reference: Verbeek (2012) 2 and 4

More advanced material is denoted by a star (*). It is not required reading.

4.1 Hypothesis Testing

4.1.1 Testing a Single Coefficient: A t -test

We are here interested in testing the null hypothesis that $\beta = q$, where q is a number of interest. A null hypothesis is often denoted H_0 . (Econometric programs often automatically report results for $H_0: \beta = 0$.)

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

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

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 (recall that $E\hat{\beta}$ equals the true value). Dividing by the standard error (deviation) of $\hat{\beta}$, we should have

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

In case $|t|$ is very large (say, 1.64 or larger), then $\hat{\beta}$ is a very unlikely outcome if β is indeed q . We therefore draw the conclusion that the true coefficient is not q , that is, we reject the null hypothesis. See Tables 4.1 and 4.2 for examples.

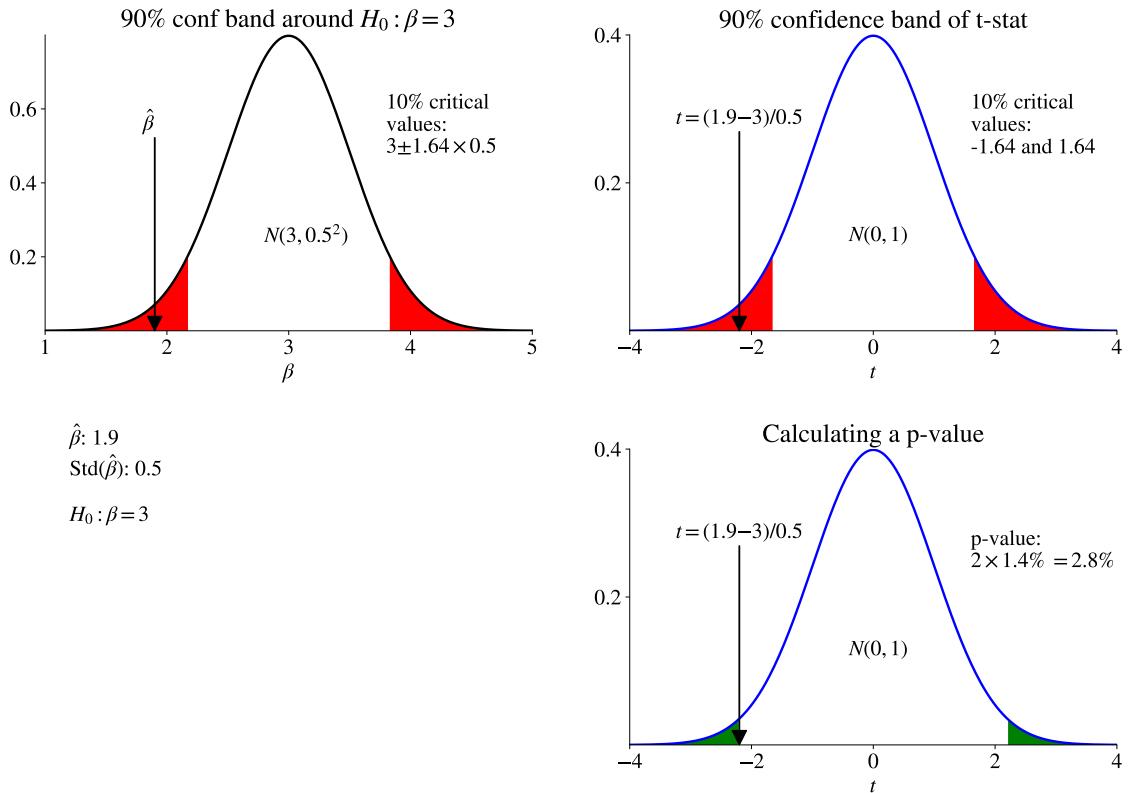


Figure 4.1: Confidence band and t-stat

4.1.2 Confidence Band around the Null Hypothesis

The logic of a hypothesis test is perhaps best described by a 90% confidence band of the t value (4.2)

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

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

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

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

The idea of the t -test is thus that we reject the null hypothesis when the estimate $\hat{\beta}$ is far from the null hypothesis (q), and where “far” depends on the significance level (here 10%) and how uncertain $\hat{\beta}$ is (as measured by $\text{Std}(\hat{\beta})$). In a sense, we then have sufficiently strong evidence against the null hypothesis. In contrast, when t and $\hat{\beta}$ are

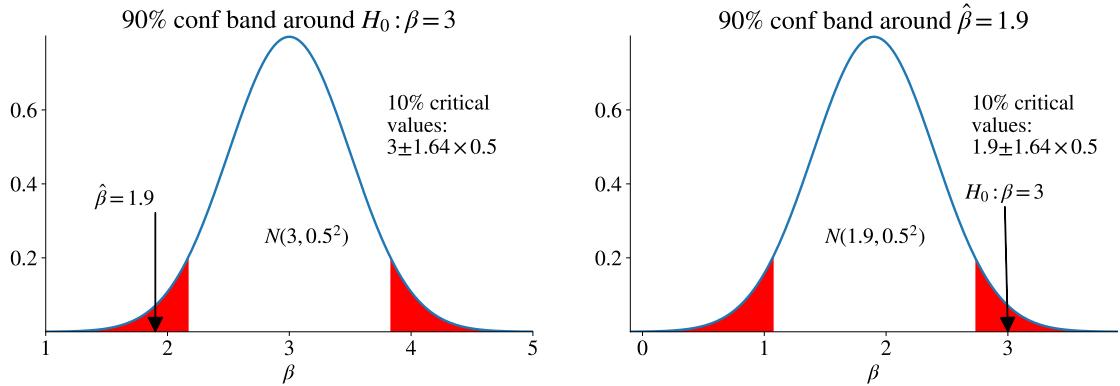


Figure 4.2: Confidence band around the null hypothesis or around the point estimate

	HiTec	Utils
constant	-0.13 (-0.94)	0.25 (1.73)
market return	1.27 (34.81)	0.51 (12.87)
R^2	0.75	0.32
Autocorr	0.39	0.79
White	0.03	0.00
All Slopes	0.00	0.00
obs	576.00	576.00

Table 4.1: CAPM regressions, monthly returns, %, US data 1970:01-2017:12. 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.

inside their confidence intervals, then the evidence is not sufficient, so we do not reject the null hypothesis. (You may compare with a court of law where the null hypothesis is that the accused is not guilty.) See Figures 4.1 and 4.2 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

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

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

■

Example 4.1 (*t-test*) With $\text{Std}(\hat{\beta}) = \sqrt{0.25}$ and $q = 3$, the 90% confidence band is $3 \pm 1.64 \times \sqrt{0.25}$, that is, $[2.175, 3.825]$. Notice that $\hat{\beta} = 1.95$ is outside this band, so we reject the null hypothesis. Equivalently, $t = (1.95 - 3)/\sqrt{0.25} = -2.1$ is outside the band $[-1.64, 1.64]$.

Using a 90% confidence band means that you are using a 10% significance level. If you want a more conservative test (that is, making it harder to reject the null hypothesis), then you may change the *critical value* (the cut-off point in the decision) from 1.64 to 1.96. This gives a 95% confidence band, so the significance level is 5%. See Figure 4.3 for an illustration.

The *p-value* is a related concept. It is the lowest significance level at which we can reject the null hypothesis. See Figure 4.1 for an illustration.

Example 4.2 (*p-value*) With $\text{Std}(\hat{\beta}) = \sqrt{0.25}$, $\hat{\beta} = 1.95$ and $q = 3$, we have $t = -2.1$. According to a $N(0,1)$ distribution, the probability of -2.1 or lower is 1.8%, so the *p-value* is 3.6%. 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 a $N(0, 1)$, especially when the sample is short. 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.

4.1.3 Confidence Band around the Point Estimate

Yet another way to construct a confidence band is to centre the band on the *point estimate*. In short,

$$t \text{ is outside } [-1.64, 1.64] \text{ when (and only when)} \quad (4.6)$$

$$q \text{ is outside } [\hat{\beta} - 1.64 \text{ Std}(\hat{\beta}), \hat{\beta} + 1.64 \text{ Std}(\hat{\beta})]. \quad (4.7)$$

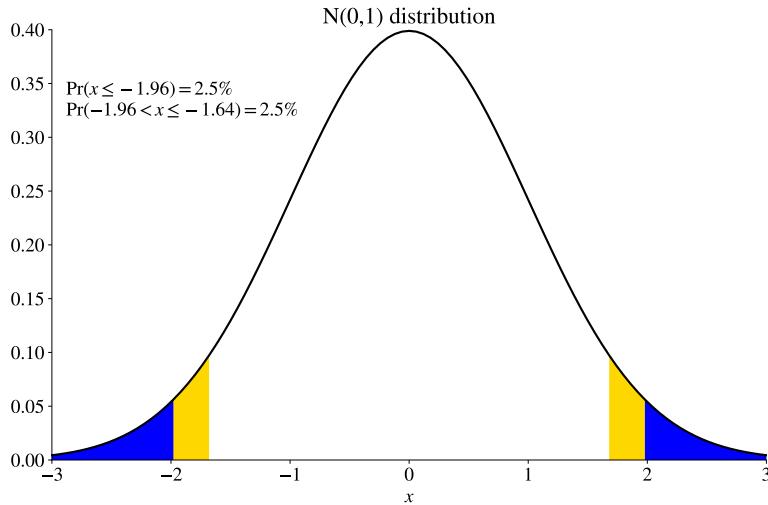


Figure 4.3: Density function of a standard normal distribution

In this case, we are 90% sure that the true value will be inside the band. If the value q (say, $q = 3$) is not in this band, then this is the same thing as rejecting (on the 10% significance level) the null hypothesis that coefficient equals q . (As before, change 1.64 to 1.96 to get a 95% confidence band.) See Figure 4.2 for an illustration.

4.1.4 Power and Size*

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

The *power* is the probability of rejecting a false H_0 . 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, which 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

$$\text{power} = \Pr(t \leq -1.64) + \Pr(t \geq 1.64). \quad (4.8)$$

(1-power is sometimes called the type II error. This is the probability of not rejecting a false H_0 .)

	HiTec	Utils
constant	0.11 (0.95)	0.10 (0.74)
market return	1.13 (32.73)	0.61 (15.74)
SMB	0.21 (4.05)	-0.19 (-3.77)
HML	-0.54 (-9.70)	0.33 (5.11)
R^2	0.81	0.40
Autocorr	0.96	0.99
White	0.00	0.00
obs	576.00	576.00

Table 4.2: Fama-French regressions, monthly returns, %, US data 1970:01-2017:12. Numbers in parentheses are t-stats. Autocorr the p-value for no autocorrelation; White is the p-value for homoskedasticity.

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

$$t = \frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} \sim N\left(\frac{\beta - q}{\text{Std}(\hat{\beta})}, 1\right). \quad (4.9)$$

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 (4.9). Clearly, the results depend on what the true value β really is. See Figure 4.4.

Example 4.3 If $\beta = 1.6$, $q = 1$ and $\text{Std}(\hat{\beta}) = 1/3$, then the power is 0.56.

4.1.5 Joint Test of Several Coefficients: Chi-Square Test

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. It might well be the case that we cannot reject any of the hypotheses (that $\beta_1 = 1$ and $\beta_3 = 0$), but that a joint test might be able to reject it. Intuitively, a joint test is like

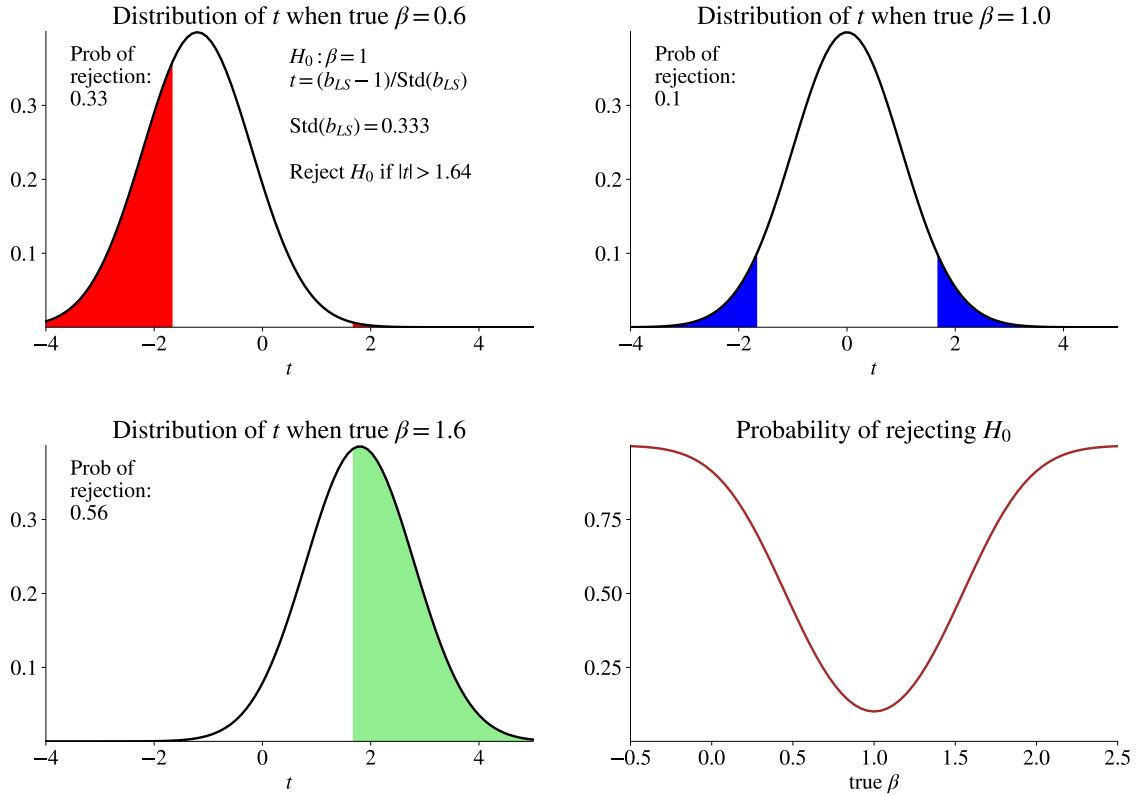


Figure 4.4: Power of t-test, assuming different true parameter values

exploiting the power of repeated sampling.

A joint test makes use of the following remark.

Remark 4.4 (Chi-square distribution) If v is a zero mean vector with n elements which 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 only has one element. In this case, the quadratic form can be written $[v / \text{Std}(v)]^2$, which is the square of a t-statistic.

Example 4.5 (Quadratic form with a chi-square distribution) If the 2×1 vector v has the following normal distribution

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

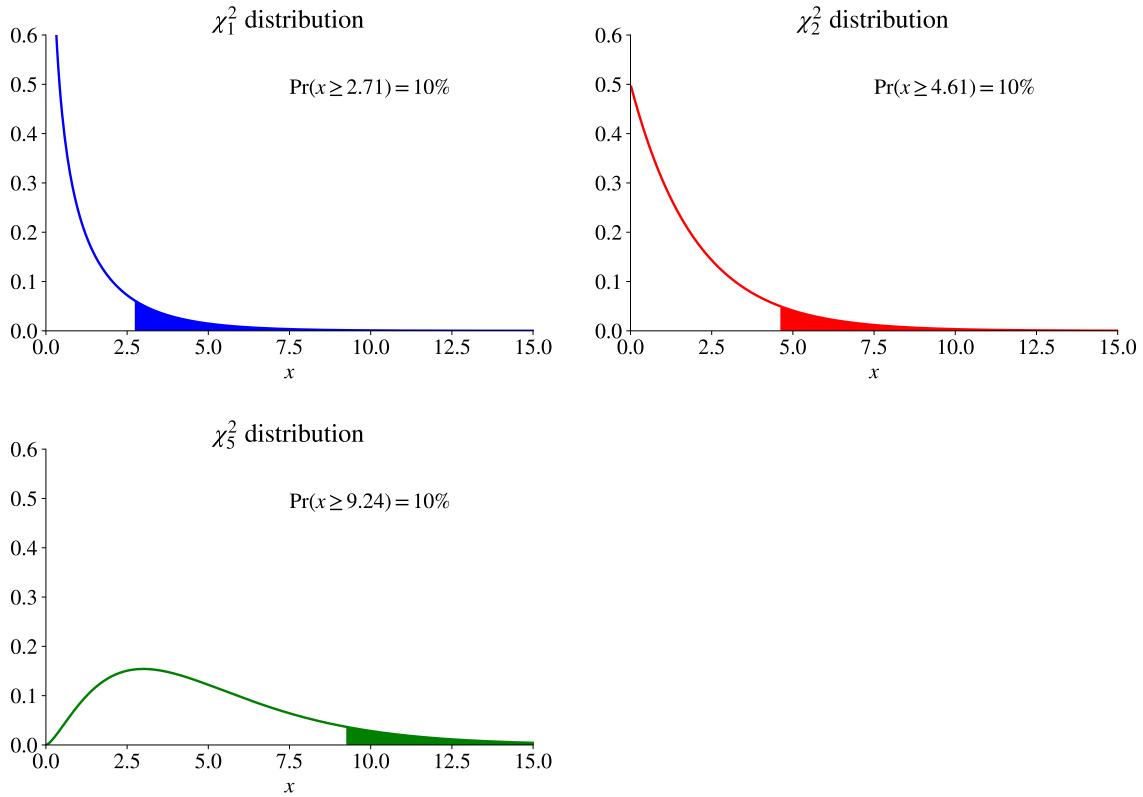


Figure 4.5: Density functions of χ^2 distributions with different degrees of freedom

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.)

For instance, suppose we have estimated a model with three coefficients and the null hypothesis is

$$H_0 : \beta_1 = 1 \text{ and } \beta_3 = 0. \quad (4.10)$$

It is convenient to write this on matrix form as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ or more generally} \quad (4.11)$$

$$R\beta = q, \quad (4.12)$$

where q has J (here 2) rows. Notice that the covariance matrix of these linear combinations is then

$$\text{Var}(R\hat{\beta}) = RV(\hat{\beta})R', \quad (4.13)$$

where $V(\hat{\beta})$ denotes the covariance matrix of the coefficients. Putting together these results we have the test static (a scalar)

$$(R\hat{\beta} - q)'[RV(\hat{\beta})R']^{-1}(R\hat{\beta} - q) \sim \chi_J^2. \quad (4.14)$$

This test statistic is compared to the critical values of a χ_J^2 distribution. (Alternatively, it can be put in the form of an F statistics, which is a small sample refinement.)

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. It can be shown that the test statistics for this hypothesis is (assuming your regression also contains an intercept)

$$TR^2/(1 - R^2) \sim \chi_{k-1}^2. \quad (4.15)$$

See Tables 4.1 and 4.2 for examples of this test.

Example 4.6 (*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

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

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

(We assume $V(\hat{\beta})$ is diagonal just because it makes it easier to invert.) Then, (4.14) 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 4.7 (*Power and size of a joint test**) Suppose $v \sim N(v_0, \Sigma)$, where v_0 might be non-zero. Then $v' \Sigma^{-1} v \sim \chi_n^2(\lambda)$ with $\lambda = v_0' \Sigma^{-1} v_0$ and where $\chi_n^2(\lambda)$ is a non-central chi-square distribution with non-centrality parameter λ . (This distribution coincides with the traditional chi-square when $\lambda = 0$.) In particular, if $R\beta - q = q_0$ (instead of zero), then the test static in (4.14) would have a $\chi_J^2(\lambda)$ distribution with $\lambda = q_0' [RV(\hat{\beta})R']^{-1} q_0$. We could then calculate the power of the test in (4.14) for different values of q_0 .

Proof. (of (4.15)) Recall that $R^2 = \text{Var}(\hat{y}_t) / \text{Var}(y_t) = 1 - \text{Var}(\hat{u}_t) / \text{Var}(y_t)$, where $\hat{y}_t = x_t' \hat{\beta}$ and \hat{u}_t are the fitted value and residual respectively. We therefore get $TR^2/(1 - R^2) = T \text{Var}(\hat{y}_t) / \text{Var}(\hat{u}_t)$. 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 $TR^2/(1 - R^2) = T \hat{\beta}' \text{Var}(x_t) \hat{\beta} / \sigma^2$, where $\sigma^2 = \text{Var}(\hat{u}_t)$. If the iid assumptions are correct, then the variance-covariance matrix of $\hat{\beta}$ is $V(\hat{\beta}) = [T \text{Var}(x_t)]^{-1} \sigma^2$, so we get

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

This has the same form as (4.14) with $R = I$ and $q = \mathbf{0}$ and J equal to the number of slope coefficients. ■

4.1.6 Confidence Bands around a Forecast and a Forecast Error*

Suppose we have estimated the linear model

$$y_t = x_t' \beta + u_t. \quad (4.16)$$

For a given (known) vector x_s , our *forecast* of y_s is

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

For a given x_s , this is just a linear combination of the estimated coefficients, so the result in (4.13) holds, but with x_s' replacing R

$$\text{Var}[E(y_s | x_s)] = x_s' V(\hat{\beta}) x_s. \quad (4.17)$$

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

$$y_s - E(y_s | x_s) = x_s' (\beta - \hat{\beta}) + u_s, \quad (4.18)$$

then we have to add the uncertainty of u_s

$$\text{Var}[y_s - \mathbb{E}(y_s|x_s)] = x'_s V(\hat{\beta}) x_s + \sigma^2. \quad (4.19)$$

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

4.1.7 A Joint Test of Several Coefficients: F-test*

The joint test can also be cast in *terms of the F distribution* (which may have better small sample properties).

Divide (4.14) by J and replace $V(\hat{\beta})$ by the estimated covariance matrix $\hat{V}(\hat{\beta})$. This is, for instance, $\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum_{t=1}^T x_t x'_t \right)^{-1}$, 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

$$\frac{(R\hat{\beta} - q)' [R\hat{V}(\hat{\beta})R']^{-1} (R\hat{\beta} - q)}{J} \sim F_{J,T-k}, \text{ where} \quad (4.20)$$

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum_{t=1}^T x_t x'_t \right)^{-1}.$$

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)

$$\frac{R^2/(k-1)}{(1-R^2)/(T-k)} \sim F_{k-1,T-k}. \quad (4.21)$$

Proof. (of (4.20)) Equation (4.20) can also be written

$$\frac{(R\hat{\beta} - q)' [R\sigma^2 \left(\sum_{t=1}^T x_t x'_t \right)^{-1} R']^{-1} (R\hat{\beta} - q) / J}{\hat{\sigma}^2 / \sigma^2}.$$

The numerator is a χ_J^2 variable divided by J . The denominator can be written $\Sigma_{t=1}^T (\hat{u}_t / \sigma)^2 / (T - k)$. If the residuals are normally distributed (and independent across time), then this is a χ_{T-k}^2 variable (not χ_T^2 since we have estimated k parameters which influence \hat{u}_t) divided by $T - k$. In addition, if the numerator and denominator are independent (which requires that the residuals are independent of the regressors), then the ratio has an $F_{J,T-k}$ distribution. ■

Example 4.8 (*Joint F test*) Continuing Example 4.6, and assuming that $\hat{V}(\hat{\beta}) = V(\hat{\beta})$,

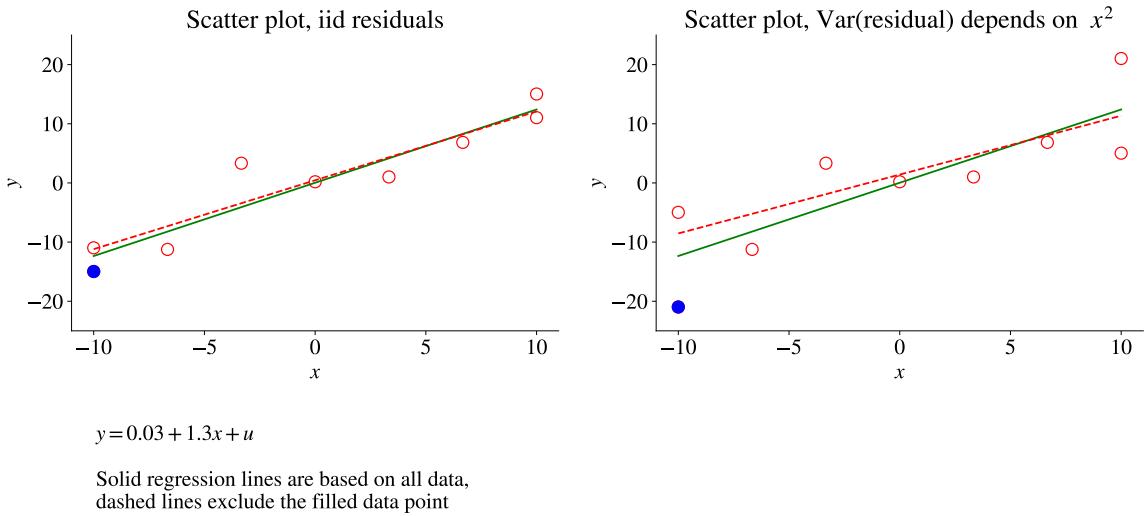


Figure 4.6: Effect of heteroskedasticity on uncertainty about regression line

we have a test statistic of $10/2 = 5$. Assume $T - k = 50$, then the 10% critical value (from an $F_{2,50}$ distribution) is 2.4, so the null hypothesis is rejected at the 10% level.

4.2 Heteroskedasticity

Suppose we have a regression model

$$y_t = x_t' b + u_t, \text{ where } E u_t = 0 \text{ and } \text{Cov}(x_{it}, u_t) = 0. \quad (4.22)$$

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).

In case the residuals actually are heteroskedastic, least squares (LS) is nevertheless a useful estimator: it is still consistent (we get the correct values as the sample becomes really large). However, the standard expression for the standard errors of the coefficients is, except in a special case, not correct. This is illustrated in Table 11.3, which shows results from simulations.

To test for heteroskedasticity, we can use *White's test of heteroskedasticity*. The null hypothesis is homoskedasticity, and the alternative hypothesis is the kind of heteroskedasticity which can be explained by the levels, squares, and cross products of the regressors—clearly a special form of heteroskedasticity. The reason for this specification is that if the squared residuals are uncorrelated with the squared regressors, then the usual LS covari-

ance matrix applies—even if the residuals have some other sort of heteroskedasticity (this is the special case mentioned before).

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

$$\hat{u}_t^2 = w_t' \gamma + v_t, \quad (4.23)$$

and to test if all the slope coefficients (not the intercept) in γ are zero. This can be done by using the fact that $TR^2/(1 - R^2) \sim \chi_p^2$, $p = \dim(w_t) - 1$.

Example 4.9 (White's test) If the regressors include $(1, x_{1t}, x_{2t})$ then w_t in (4.23) is the vector $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$.

$\alpha :$	$\underline{\gamma = 0}$		$\underline{\gamma = 1}$	
	0	1	0	1
Simulated	7.1	19.2	13.5	24.9
OLS formula	7.1	13.3	13.4	19.3
White's	7.0	18.5	13.3	24.3
Bootstrap	7.1	18.5	13.4	24.4
Bootstrap 2	7.0	18.5	13.3	24.3

Table 4.3: Standard error of OLS slope (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. Number of simulations: 25000. 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 (4.22) with an ARCH structure of the residuals—and estimating the whole thing with maximum likelihood (MLE). As a by-product we get the correct standard errors—provided the assumed distribution (in the likelihood function) is correct. Second, we could stick to OLS, but use another expression for the variance of the coefficients: a heteroskedasticity consistent covariance matrix, 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 $\hat{\beta}_1$ is found from

$$\hat{\beta} = \beta + S_{xx}^{-1} (x_1 u_1 + x_2 u_2 + \dots x_T u_T), \quad (4.24)$$

where $S_{xx} = \sum_{t=1}^T x_t x'_t$. If we assume that the residuals are uncorrelated with each other, then

$$\begin{aligned} \text{Var}(\hat{\beta}) &= S_{xx}^{-1} (x_1 x'_1 \sigma_1^2 + x_2 x'_2 \sigma_2^2 + \dots x_T x'_T \sigma_T^2) S_{xx}^{-1} \\ &= S_{xx}^{-1} \underbrace{\sum_{t=1}^T x_t x'_t \sigma_t^2}_S S_{xx}^{-1}. \end{aligned} \quad (4.25)$$

(Notice that S_{xx} and S denote very different things.) This expression cannot be simplified further since σ_t is not constant—and also related to x_t^2 . The idea of White's estimator is to estimate S by

$$\hat{S} = \sum_{t=1}^T x_t x'_t \hat{u}_t^2. \quad (4.26)$$

(This also allows for the case with several elements in x_t , that is, several regressors.)

It is straightforward to show that the standard expression for the variance underestimates the true variance when there is a positive relation between x_t^2 and σ_t^2 (and vice versa). 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 4.6 and Table 11.3.

White's covariance matrix should be applied when White's test (4.23) 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 (4.26) risks introducing more noise.

Remark 4.10 (*Standard OLS vs White's variance*) For simplicity, consider the case of only one regressor. If x_t^2 is not related to σ_t^2 , then we could write the last term in (4.25) as

$$\begin{aligned} \sum_{t=1}^T x_t^2 \sigma_t^2 &= \frac{1}{T} \sum_{t=1}^T \sigma_t^2 \sum_{t=1}^T x_t^2 \\ &= \bar{\sigma}^2 \sum_{t=1}^T x_t^2 \end{aligned}$$

where $\bar{\sigma}^2$ is the average variance, typically estimated as $\sum_{t=1}^T u_t^2 / T$. That is, it is the

same as for standard OLS. In addition, notice that

$$\sum_{t=1}^T x_t^2 \sigma_t^2 > \frac{1}{T} \sum_{t=1}^T \sigma_t^2 \sum_{t=1}^T x_t^2$$

if x_t^2 is positively related to σ_t^2 (and vice versa). For instance, with $(x_1^2, x_2^2) = (10, 1)$ and $(\sigma_1^2, \sigma_2^2) = (5, 2)$, $\sum_{t=1}^T x_t^2 \sigma_t^2 = 10 \times 5 + 1 \times 2 = 52$ while $\frac{1}{T} \sum_{t=1}^T \sigma_t^2 \sum_{t=1}^T x_t^2 = \frac{1}{2}(5+2)(10+1) = 38.5$.

Remark 4.11 (GLS*) With heteroskedasticity and/or autocorrelation, OLS is still consistent and we can adjust the covariance matrix of the coefficients. However, OLS is less efficient (higher uncertainty of the coefficients) than GLS (Generalized Least Squares) is. The basic idea of GLS is transform regression equation so

$$y_i^* = x_i^{*\prime} \beta + \varepsilon_i^*,$$

have iid residuals. Estimating β with LS on this transformation is efficient (called GLS) and the traditional expressions of the covariance matrix of the coefficients can be used. For instance, with heteroskedasticity, the transformation is

$$\frac{y_i}{\sigma_i} = \frac{x_i'}{\sigma_i} \beta + \frac{\varepsilon_i}{\sigma_i}.$$

Notice that ε_i/σ_i has a constant variance (equal to one). Notice also that all regressors are scaled by $1/\sigma_i$ (the constant 1 becomes $1/\sigma_i$). With first-order autocorrelation, ($\varepsilon_i = \rho \varepsilon_{i-1} + v_i$, v_i is iid), the transformation is a “quasi-difference” of regression equation

$$y_i - \rho y_{i-1} = (x_i - \rho x_{i-1})' \beta + (\varepsilon_i - \rho \varepsilon_{i-1}).$$

This new residual, $\varepsilon_i - \rho \varepsilon_{i-1}$, is iid. In practice we don't know σ_i or ρ , so we first estimate them (the method is then called “feasible” GLS, FGLS.) Notice, however, that FGLS may improve the efficiency, but can be dangerous if we model the heteroskedasticity/autocorrelation wrongly. Many applications tend to generate unstable (non-robust) results from FGLS.

4.3 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 $\hat{\beta}$. In this case, LS is

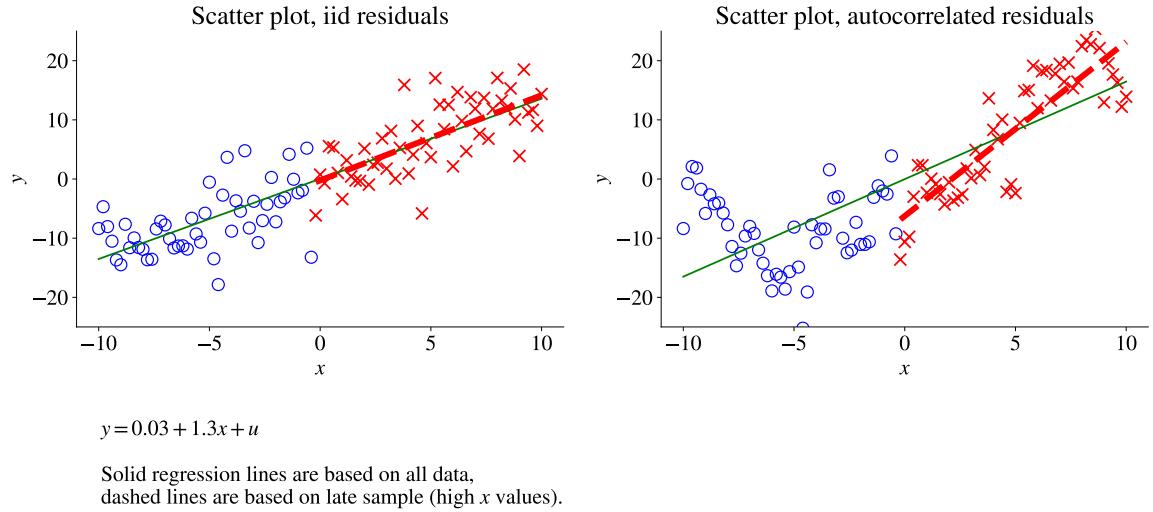


Figure 4.7: Effect of autocorrelation on uncertainty about regression line

typically still consistent (exceptions: 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 standard error calculated on the iid assumptions. This is perhaps easiest to understand in the case of estimating the mean of a data series, that is, when regressing a data series on a constant only. If the residual is positively autocorrelated (have long swings), then the sample mean can deviate from the true mean for an extended period of time—perhaps for most of a sample: the estimate is imprecise. See Figure 4.7 for an illustration.

There are several straightforward tests of autocorrelation—all based on using the fitted residuals. The null hypothesis is no autocorrelation. First, estimate the autocorrelations of the fitted residuals as

$$\rho_s = \text{Corr}(\hat{u}_t, \hat{u}_{t-s}), s = 1, \dots, L. \quad (4.27)$$

Second, test autocorrelation s by using the fact that $\sqrt{T}\hat{\rho}_s$ has a standard normal distribution (in large samples)

$$\sqrt{T}\hat{\rho}_s \sim N(0, 1). \quad (4.28)$$

To extend (4.28) to higher-order autocorrelation, use the Box-Pierce test

$$Q_L = T \sum_{s=1}^L \hat{\rho}_s^2 \xrightarrow{d} \chi_L^2. \quad (4.29)$$

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

$$DW \approx 2 - 2\hat{\rho}_1, \quad (4.30)$$

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).

$\rho :$	0.0	0.75
Simulated	5.8	23.1
OLS formula	5.8	8.7
Newey-West	5.7	16.3
VARHAC	5.7	22.4
Bootstrapped	5.5	19.6

Table 4.4: 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 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

$\rho :$	$\kappa = 0.0$		$\kappa = 0.75$	
	0.0	0.75	0.0	0.75
Simulated	5.8	8.7	3.9	11.0
OLS formula	5.8	8.6	3.9	5.8
Newey-West	5.7	8.4	3.8	8.9
VARHAC	5.7	8.5	3.8	10.5
Bootstrapped	5.8	8.5	3.8	10.1

Table 4.5: 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 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

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—for instance, the “*Newey-West covariance matrix*.”

To understand the Newey-West covariance matrix, notice that the variance of $\text{Var}(\hat{\beta})$ is

$$\text{Var}(\hat{\beta}) = S_{xx}^{-1} \underbrace{\text{Var}(x_1 u_1 + x_2 u_2 + \dots + x_T u_T)}_S S_{xx}^{-1}. \quad (4.31)$$

However, there might be correlation across time periods, so the S term in the middle needs to account for terms like $\text{Cov}(x_t u_t, x_{t-s} u_{t-s})$. For instance, for $T = 3$ the S term is

$$S = \text{Var}(x_1 u_1) + \text{Var}(x_2 u_2) + \text{Var}(x_3 u_3) + \\ 2 \text{Cov}(x_2 u_2, x_1 u_1) + 2 \text{Cov}(x_3 u_3, x_2 u_2) + 2 \text{Cov}(x_3 u_3, x_1 u_1). \quad (4.32)$$

When data is uncorrelated across time (observations), then all the covariance terms are zero. With autocorrelation, they may not be. For a general T , the S term is

$$S = \sum_{t=1}^T \text{Var}(x_t u_t) + 2 \sum_{s=1}^m \sum_{t=s+1}^T \text{Cov}(x_t u_t, x_{t-s} u_{t-s}), \quad (4.33)$$

where m denotes the number of covariance terms that might be non-zero (at most, $m = T - 1$).

The idea of the Newey-West estimator is to estimate S by (with several regressors)

$$\hat{S} = \sum_{t=1}^T x_t x'_t \hat{u}_t^2 + 2 \sum_{s=1}^m \sum_{t=s+1}^T \left(1 - \frac{|s|}{m+1}\right) (x_t x'_{t-s} + x_{t-s} x'_t) \hat{u}_t \hat{u}_{t-s}, \quad (4.34)$$

where the weights $1 - |s|/(m+1)$ are close to 1 for small lags, but decline linearly (tent shaped weights). The point of using such weights is to make sure that the \hat{S} matrix remains invertible (to show this is somewhat involved). This suggests that m should be somewhat larger than last lag with significant autocorrelation. However, a common rule of thumb is to use round $m = \text{floor}(0.75T^{1/3})$, where $\text{floor}()$ means rounding down to nearest integer (and alternative rule is $m = \text{floor}(4(T/100)^{2/9})$).

For instance, with only one lag ($m = 1$) the calculation is (with several regressors)

$$\hat{S} = \sum_{t=1}^T x_t x'_t \hat{u}_t^2 + 2 \sum_{t=2}^T \left(1 - \frac{1}{2}\right) (x_t x'_{t-1} + x_{t-1} x'_t) \hat{u}_t \hat{u}_{t-1}, \quad (4.35)$$

and by excluding all lags (setting $m = 0$), the Newey-West estimator coincides with White's estimator

$$\hat{S} = \sum_{t=1}^T x_t x'_t \hat{u}_t^2. \quad (4.36)$$

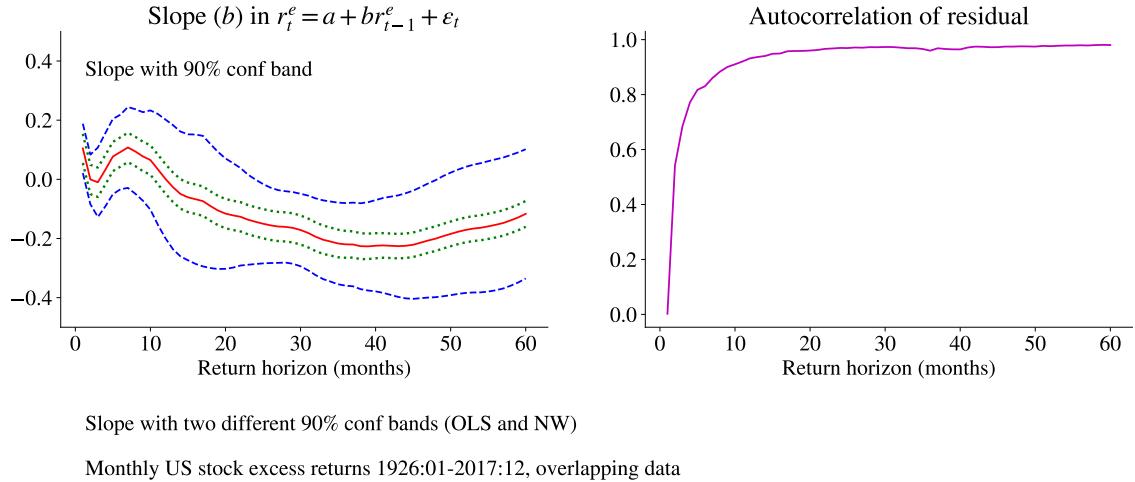


Figure 4.8: Slope coefficient, LS vs Newey-West standard errors

Hence, Newey-West estimator handles also heteroskedasticity.

Remark 4.12 (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 (use $\hat{\varepsilon}_t$ instead of $x_t \hat{u}_t$ in (4.34)).

It is clear from this expression that what really counts is not so much the autocorrelation in u_t per se, 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, $x_t = 1$ is extremely positively autocorrelated. In contrast, when the regressor has no autocorrelation, then the product does not either. This is illustrated in Tables 18.4–18.5.

Figures 4.8–4.9 are empirical examples of the importance of using the Newey-West method rather than relying of the iid assumptions. In both cases, the residuals have strong positive autocorrelation.

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 and uncertainty.

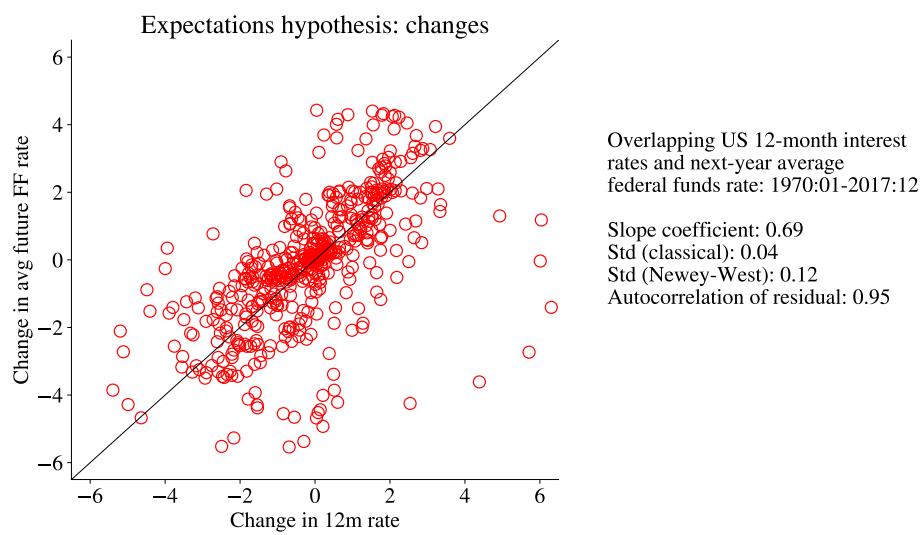


Figure 4.9: US 12-month interest and average federal funds rate (next 12 months)

Chapter 5

Testing CAPM and Multifactor Models

Reference: Elton, Gruber, Brown, and Goetzmann (2010) 15

More advanced material is denoted by a star (*). It is not required reading.

5.1 Market Model

Let $R_{it}^e = R_{it} - R_{ft}$ be the excess return on asset i in excess over the riskfree 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

$$E R_{it}^e = \beta_i E R_{mt}^e, \text{ where } \beta_i = \frac{\text{Cov}(R_i, R_m)}{\text{Var}(R_m)}. \quad (5.1)$$

To test this, consider the regression

$$\begin{aligned} R_{it}^e &= \alpha_i + b_i R_{mt}^e + \varepsilon_{it}, \text{ where} \\ E \varepsilon_{it} &= 0 \text{ and } \text{Cov}(R_{mt}^e, \varepsilon_{it}) = 0. \end{aligned} \quad (5.2)$$

The two last conditions are automatically imposed by LS. Take expectations of the regression (assuming we know the coefficients) to get

$$E R_{it}^e = \alpha_i + b_i E R_{mt}^e. \quad (5.3)$$

Notice that the LS estimate of b_i is the sample analogue to β_i in (5.1). 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} is the

correct benchmark (the tangency portfolio for which (5.1) is true by definition), then it is a test of whether asset R_{it} is correctly priced. This is typically the perspective in performance analysis of mutual funds. Alternatively, if we assume that R_{it} is correctly priced, then it is a test of the mean-variance efficiency of R_{mt} . 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

$$\frac{\hat{\alpha}_i}{\text{Std}(\hat{\alpha}_i)} \xrightarrow{d} N(0, 1) \text{ under } H_0 : \alpha_i = 0. \quad (5.4)$$

We get $\text{Std}(\hat{\alpha}_i)$ from the OLS regression (possibly with an adjustment due to autocorrelation and/or heteroskedasticity). Note that this is the distribution under the null hypothesis that the true value of the intercept is zero, that is, that CAPM is correct.

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 extremely 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 can possibly suggest how the model needs to be changed.

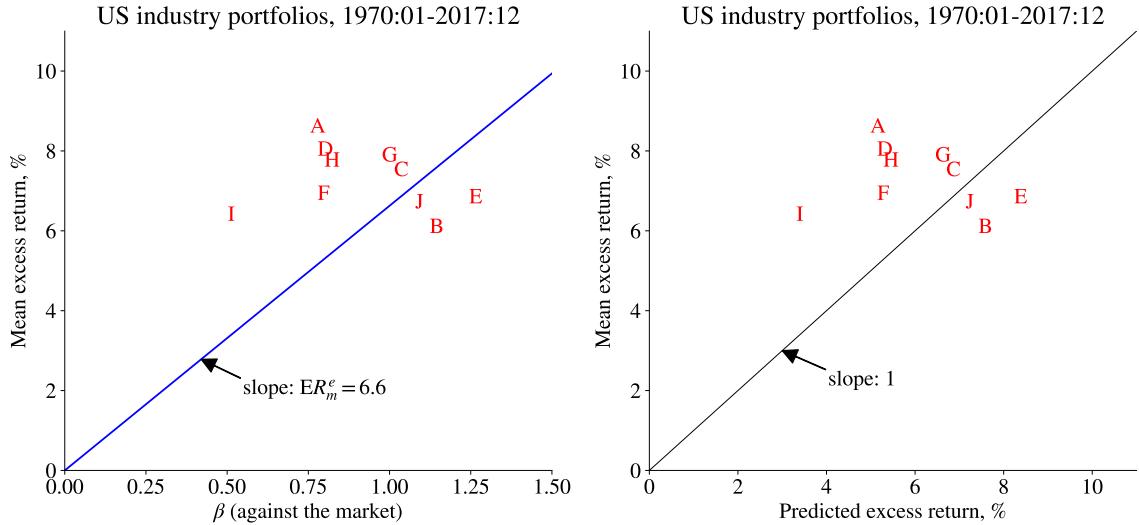
The empirical results from such tests vary with the test assets used. For US portfolios, CAPM seems to work reasonably well for some types of portfolios (for instance, portfolios based on firm size or industry), but much worse for other types of portfolios (for instance, portfolios based on firm dividend yield or book value/market value ratio). Figure 5.1 shows some results for US industry portfolios.

In Figure 5.1, the results are presented in two different ways:

	horizontal axis	vertical axis
1 :	β_i	$\sum_{t=1}^T R_{it}^e / T$
2 :	$\beta_i \sum_{t=1}^T R_{mt}^e / T$	$\sum_{t=1}^T R_{it}^e / T$

(5.5)

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, $\sum_{t=1}^T R_m^e / T$.



	α	t-stat	σ
A (NoDur)	3.42	2.75	8.58
B (Durbl)	-1.52	-0.79	13.26
C (Manuf)	0.63	0.70	6.15
D (Enrgy)	2.69	1.25	14.79
E (HiTec)	-1.57	-0.94	11.52
F (Telcm)	1.62	1.04	10.75
G (Shops)	1.23	0.92	9.19
H (Hlth)	2.28	1.41	11.11
I (Utils)	2.98	1.76	11.62
J (Other)	-0.55	-0.56	6.81

CAPM: $R_i^e = \alpha_i + \beta_i R_m^e + e_i$
 Predicted excess return: $\beta_i ER_m^e$
 Factor: US market
 α and σ (std of residual) are in annualized %
 p-val for testing if all $\alpha_i = 0$: 0.105

Figure 5.1: CAPM regressions on US industry indices

In the second approach, CAPM says that all data points should cluster around a 45-degree line. In either case, the vertical distance to the line is α_i (which should be zero according to CAPM).

5.1.1 Econometric Properties of the CAPM Test

A common finding from Monte Carlo simulations is that these tests tend to reject a true null hypothesis too often when the critical values from the asymptotic distribution are used. The practical consequence is that we should either use adjusted critical values (from Monte Carlo or bootstrap simulations)—or more pragmatically, that we should only believe in strong rejections of the null hypothesis.

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 independent of the market return) are correct. Then, it is straightforward to show that the variance of Jensen's alpha is

$$\text{Var}(\hat{\alpha}_i) = [1 + (SR_m)^2]\sigma^2/T, \quad (5.6)$$

where σ^2 is the variance of the residual in (5.2) 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 5.1 Suppose we have monthly data with $\hat{\alpha}_i = 0.2\%$ (that is, $0.2\% \times 12 = 2.4\%$ per year), $\sigma = 3\%$ (that is, $3\% \times \sqrt{12} \approx 10\%$ per year) and a market Sharpe ratio of 0.15 (that is, $0.15 \times \sqrt{12} \approx 0.5$ per year). (This corresponds well to US CAPM regressions for industry portfolios.) A significance level of 10% requires a t -statistic (5.4) of at least 1.64, so

$$\frac{0.2}{\sqrt{1 + 0.15^2} 3 / \sqrt{T}} \geq 1.64 \text{ or } T \geq 626.$$

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 just 2% (7% per year). Notice that cumulating a 0.3% return over 25 years means almost 2.5 times the initial value.

Proof. (*Proof of (5.6)) Consider the regression equation $y_t = x'_t b + \varepsilon_t$. With iid errors that are independent of all regressors (also across observations), the LS estimator, \hat{b}_{LS} , is asymptotically distributed as

$$\sqrt{T}(\hat{b}_{LS} - b) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Sigma_{xx}^{-1}), \text{ where } \sigma^2 = \text{Var}(\varepsilon_t) \text{ and } \Sigma_{xx} = \text{plim} \Sigma_{t=1}^T x_t x'_t / T.$$

When the regressors are just a constant (equal to one) and one variable regressor, f_t , so $x_t = [1, f_t]'$, then we have

$$\begin{aligned} \Sigma_{xx} &= E \sum_{t=1}^T x_t x'_t / T = E \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} = \begin{bmatrix} 1 & E f_t \\ E f_t & E f_t^2 \end{bmatrix}, \text{ so} \\ \sigma^2 \Sigma_{xx}^{-1} &= \frac{\sigma^2}{E f_t^2 - (E f_t)^2} \begin{bmatrix} E f_t^2 & -E f_t \\ -E f_t & 1 \end{bmatrix} = \frac{\sigma^2}{\text{Var}(f_t)} \begin{bmatrix} \text{Var}(f_t) + (E f_t)^2 & -E f_t \\ -E f_t & 1 \end{bmatrix}. \end{aligned}$$

(In the last line we use $\text{Var}(f_t) = \mathbb{E} f_t^2 - (\mathbb{E} f_t)^2$.) ■

5.1.2 Interpretation of the CAPM Test

Instead of a t-test, we can use the equivalent chi-square test

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} \xrightarrow{d} \chi_1^2 \text{ under } H_0: \alpha_i = 0. \quad (5.7)$$

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 (5.7) can be written

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} = \frac{(SR_c)^2 - (SR_m)^2}{[1 + (SR_m)^2]/T}, \quad (5.8)$$

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.

Proof. (*Proof of (5.8)) From the CAPM regression (5.2) we have

$$\text{Cov} \begin{bmatrix} R_{it}^e \\ R_{mt}^e \end{bmatrix} = \begin{bmatrix} \beta_i^2 \sigma_m^2 + \text{Var}(\varepsilon_{it}) & \beta_i \sigma_m^2 \\ \beta_i \sigma_m^2 & \sigma_m^2 \end{bmatrix}, \text{ and } \begin{bmatrix} \mu_i^e \\ \mu_m^e \end{bmatrix} = \begin{bmatrix} \alpha_i + \beta_i \mu_m^e \\ \mu_m^e \end{bmatrix}.$$

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

$$\left(\frac{\mu_c^e}{\sigma_c} \right)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + \left(\frac{\mu_m^e}{\sigma_m} \right)^2,$$

which we can write as

$$(SR_c)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + (SR_m)^2.$$

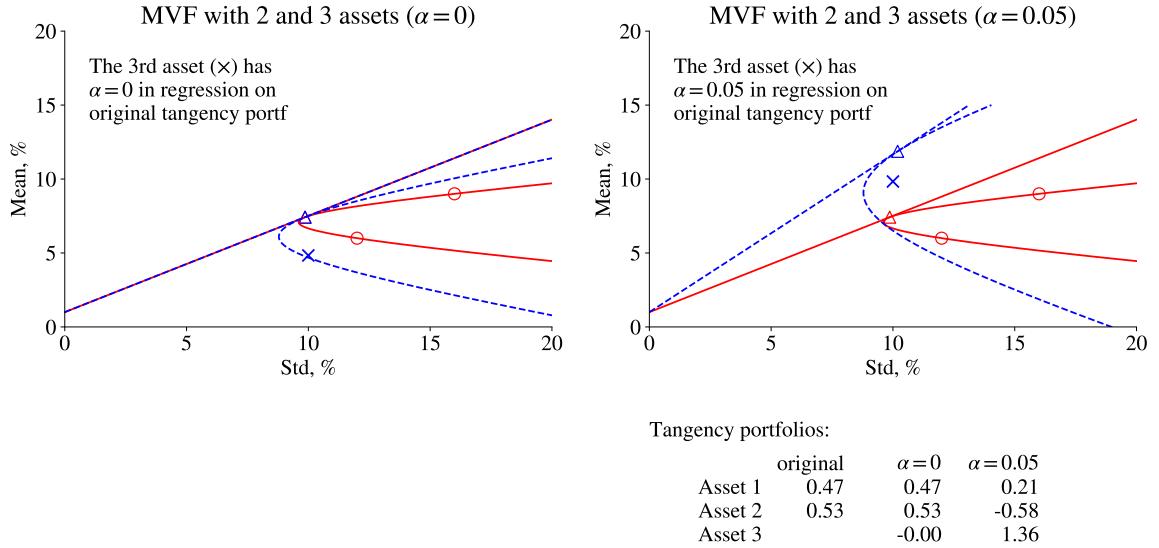


Figure 5.2: Effect on MV frontier of adding assets

Combine this with (5.6) which shows that $\text{Var}(\hat{\alpha}_i) = [1 + (SR_m)^2] \text{Var}(\varepsilon_{it})/T$. ■

This is illustrated in Figure 5.2 which shows the effect of adding an asset to the investment opportunity set. The basic point is that testing the alphas is the same as testing if the new assets move the location of the tangency portfolio. 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.

5.1.3 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.

While it is straightforward to construct such a test, it is also a bit messy. As a quick way out, the following will work fairly well, although it does not provide a formal test. First, test each asset individually. Second, form a few different portfolios of the test assets (equally weighted, value weighted) and test these portfolios. Although this does not deliver one single test statistic, it provides plenty of information to base a judgment

on. Alternatively, we can apply a Bonferroni correction of the individual t-stats: reject CAPM at the 5% significance level only if the largest t-stat (in absolute terms) exceeds the critical value at the $0.05/n$ significance level. For instance, with $n = 25$, the critical value from a standard normal distribution would be 3.09 instead of 1.96. For another approach, see Section 5.1.4.

A quite different approach to study a cross-section of assets is to first estimate CAPM regressions (5.2) for each of the assets ($i = 1, 2, \dots, n$) and then the following cross-sectional regression

$$\bar{R}_i^e = \gamma + \lambda \hat{\beta}_i + u_i, \quad (5.9)$$

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 typically tends to bias the slope coefficient (λ) towards zero. To get the intuition for this bias, consider an extremely noisy measurement of the regressor: it would be virtually uncorrelated with the dependent variable (noise isn't correlated with anything), so the estimated slope coefficient would be close to zero.

If we could overcome this bias (and we can by being careful), then the testable implications of CAPM is that $\gamma = 0$ and that λ equals the average market excess return. We also want (5.9) to have a high R^2 —since it should be unity in a very large sample (if CAPM holds).

5.1.4 Several Assets: SURE Approach

This section outlines how we can set up a formal test of CAPM when there are several test assets. We first estimate (5.2) for each of the assets, and then test the null hypothesis that all the alphas are zero.

For simplicity, suppose we have two test assets. The null hypothesis

$$H_0 : \alpha_1 = 0 \text{ and } \alpha_2 = 0. \quad (5.10)$$

In a large sample, the estimator is normally distributed. Therefore, under the null hypothesis we have the following result.

$$\begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} [1 + (SR_m)^2]/T \right). \quad (5.11)$$

Notice that the zero means in (5.11) come from the null hypothesis: the distribution is (as usual) constructed by pretending that the null hypothesis is true. In practice we use the sample moments for the covariance matrix. We can now construct a chi-square test by using the following fact.

Remark 5.2 If the $n \times 1$ vector $v \sim N(0, \Sigma)$, then $v' \Sigma^{-1} v \sim \chi_n^2$.

To apply this, form the test static

$$T \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix}' [1 + (SR_m)^2]^{-1} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \sim \chi_2^2. \quad (5.12)$$

This can also be transformed into an F test, which might have better small sample properties.

Proof. (of (5.11)) Stack (5.2) for the two assets

$$\begin{aligned} R_{1t}^e &= \alpha_1 + b_1 R_{mt}^e + \varepsilon_{1t}, \\ R_{2t}^e &= \alpha_2 + b_2 R_{mt}^e + \varepsilon_{2t} \end{aligned}$$

where $E \varepsilon_{it} = 0$ and $\text{Cov}(R_{mt}^e, \varepsilon_{it}) = 0$. This is a system of seemingly unrelated regressions (SURE)—with the same regressor (see, for instance, Wooldridge (2002) 7.7). In this case, the efficient estimator (GLS) is LS on each equation separately. Moreover, the covariance matrix of the coefficients is particularly simple. To see what the covariances of the coefficients are, write the regression equation for asset 1 on a traditional form

$$R_{1t}^e = x_t' \beta_1 + \varepsilon_{1t}, \text{ where } x_t = \begin{bmatrix} 1 \\ R_{mt}^e \end{bmatrix}, \beta_1 = \begin{bmatrix} \alpha_1 \\ b_1 \end{bmatrix},$$

and similarly for the second asset (and any further assets). Define

$$\hat{\Sigma}_{xx} = \sum_{t=1}^T x_t x_t' / T, \text{ and } \hat{\sigma}_{ij} = \sum_{t=1}^T \hat{\varepsilon}_{it} \hat{\varepsilon}_{jt} / T,$$

where $\hat{\varepsilon}_{it}$ is the fitted residual of asset i . The key result is then that the (estimated) asymptotic covariance matrix of the vectors $\hat{\beta}_i$ and $\hat{\beta}_j$ (for assets i and j) is

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \hat{\sigma}_{ij} \hat{\Sigma}_{xx}^{-1} / T.$$

(In many text books, this is written $\hat{\sigma}_{ij} (X' X)^{-1}$.) From (5.6) we know that the upper left element of Σ_{xx}^{-1} / T equals $[1 + (SR_m)^2] / T$. ■

5.1.5 Representative Results of the CAPM Test

One of the more interesting empirical studies is Fama and French (1993) (see also Fama and French (1996)). They 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 (so this is a *double-sort*)

$$\begin{bmatrix} \text{small size, low B/M} & \dots & \dots & \dots & \text{small size, high B/M} \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \vdots & & & & \ddots \\ \text{large size, low B/M} & & & & \text{large size, high B/M} \end{bmatrix}$$

They run a traditional CAPM regression on each of the 25 portfolios (monthly data 1963–1991)—and then study if the expected excess returns are related to the betas as they should according to CAPM (recall that CAPM implies $E R_{it}^e = \beta_i \lambda$ where λ is the risk premium (excess return) on the market portfolio).

However, it is found that there is almost no relation between $E R_{it}^e$ and β_i (there is a cloud in the $\beta_i \times E R_{it}^e$ space). This is due to the combination of two features of the data. First, *within a BE/ME quintile*, there is a positive relation (across size quintiles) between $E R_{it}^e$ and β_i —as predicted by CAPM. Second, *within a size quintile* there is a negative relation (across BE/ME quintiles) between $E R_{it}^e$ and β_i —in stark contrast to CAPM. Figure 5.1 shows some results for US industry portfolios and Figures 5.3–5.4 for US size/book-to-market portfolios.

5.1.6 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).

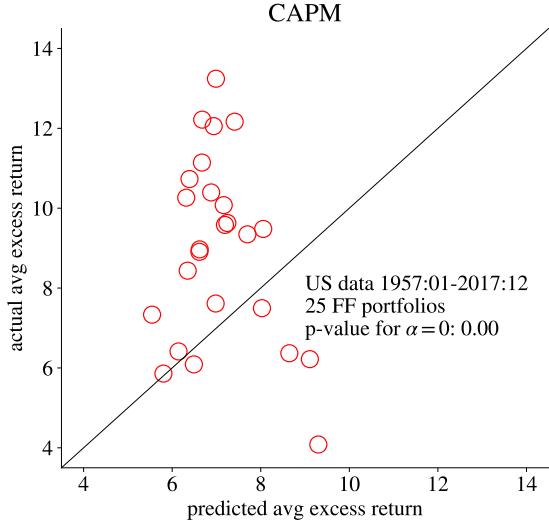


Figure 5.3: CAPM, FF portfolios

5.2 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 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)

$$\bar{R}_{jt}^e = \frac{1}{N_j} \sum_{i \in \text{Group } j} R_{it}^e, \quad (5.13)$$

where N_j is the number of individuals in group j .

Then, we run a factor model

$$\bar{R}_{jt}^e = x_t' \beta_j + v_{jt}, \text{ for } j = 1, 2, \dots, M \quad (5.14)$$

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.

Example 5.3 (*CalTime with two investor groups*) With two investor groups, estimate the

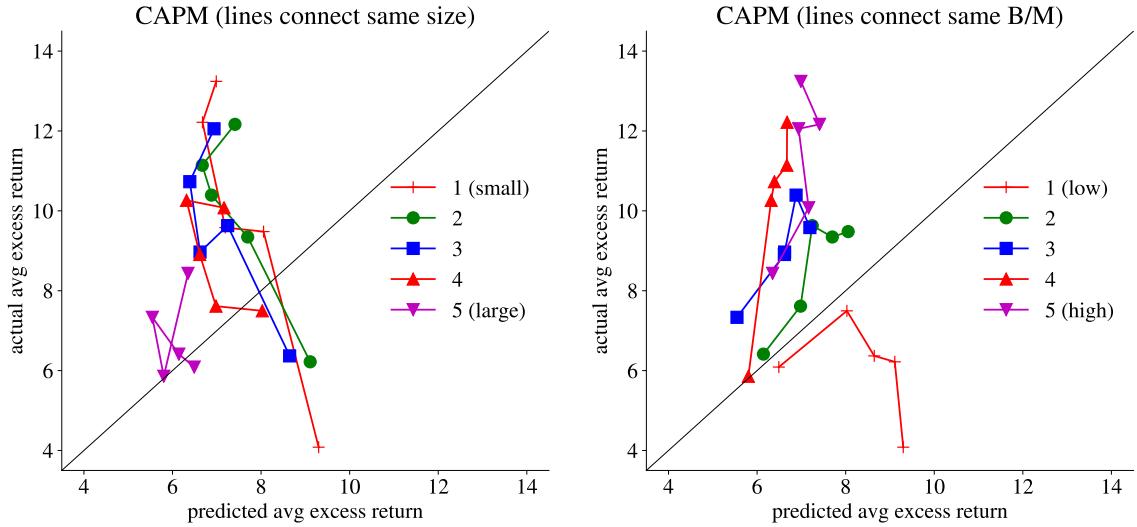


Figure 5.4: CAPM, FF portfolios

following SURE system

$$\begin{aligned}\bar{R}_{1t}^e &= x_t' \beta_1 + v_{1t}, \\ \bar{R}_{2t}^e &= x_t' \beta_2 + v_{2t}.\end{aligned}$$

The CalTime approach is straightforward and the cross-sectional correlations are fairly easy to handle (in the SURE approach). However, it forces us to define discrete investor groups—which makes it hard to handle several different types of investor characteristics (for instance, age, trading activity and income) at the same time.

5.3 Several Factors

In multifactor models, (5.2) is still valid—provided we reinterpret b_i and R_{mt}^e as vectors, so $b_i R_{mt}^e$ stands for $b_{io} R_{ot}^e + b_{ip} R_{pt}^e + \dots$

$$R_{it}^e = \alpha + b_{io} R_{ot}^e + b_{ip} R_{pt}^e + \dots + \varepsilon_{it}. \quad (5.15)$$

In this case, (5.2) is a multiple regression, but the test (5.4) 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 (two more factors are needed to also fit the seven bond portfolios that they use). 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 5.4 (*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 other gross or net returns, you add the riskfree rate: $R_i - R_j + R_f$, implicitly assuming that your portfolio includes a riskfree investment of the same size as your long-short exposure (x).

Chen, Roll, and Ross (1986) use a number of macro variables as factors—along with traditional market indices. They find that industrial production and inflation surprises are priced factors, while the market index might not be.

Figure 5.5 shows some results for the Fama-French model on US industry portfolios and Figures 5.6–5.7 on the 25 Fama-French portfolios.

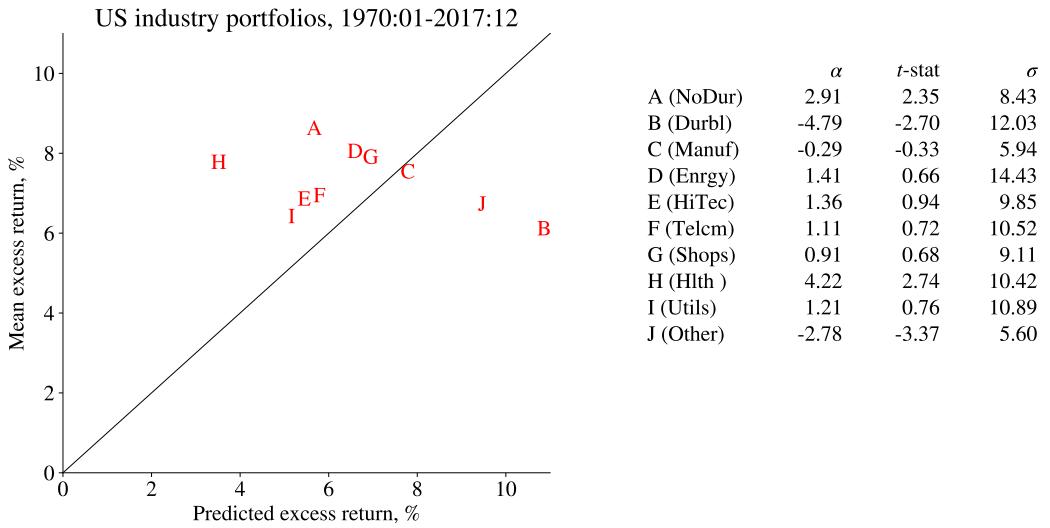
5.4 Fama-MacBeth^{*}

Reference: Cochrane (2001) 12.3; Campbell, Lo, and MacKinlay (1997) 5.8; Fama and MacBeth (1973)

The Fama and MacBeth (1973) approach is a bit different from the regression approaches discussed so far. The method has three steps, described below.

- First, estimate the betas β_i ($i = 1, \dots, n$) from (5.2) (this is a time-series regression). This is often done on the whole sample—assuming the betas are constant. Sometimes, the betas are estimated separately for different sub samples (so we could let $\hat{\beta}_i$ carry a time subscript in the equations below).
- Second, run a cross sectional regression for every t . That is, for period t , estimate λ_t from the cross section (across the assets $i = 1, \dots, n$) regression

$$R_{it}^e = \lambda'_t \hat{\beta}_i + \varepsilon_{it}, \quad (5.16)$$



Fama-French model

Predicted excess return: $\beta_m R_m^e + \beta_{SMB} R_{SMB} + \beta_{HML} R_{HML}$

Factors: US market, SMB (size), and HML (book-to-market)

α and σ (std of residual) are in annualized %
p-val for testing if all $\alpha_i = 0$: 0.0

Figure 5.5: Fama-French regressions on US industry indices

where $\hat{\beta}_i$ are the regressors. (Note the difference to the traditional cross-sectional approach discussed in (5.9), where the second stage regression regressed $E R_{it}^e$ on $\hat{\beta}_i$, while the Fama-French approach runs one regression for every time period.)

- Third, estimate the time averages

$$\hat{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_{it} \text{ for } i = 1, \dots, n, \text{ (for every asset)} \quad (5.17)$$

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^T \hat{\lambda}_t. \quad (5.18)$$

The second step, using $\hat{\beta}_i$ as regressors, creates an errors-in-variables problem since $\hat{\beta}_i$ are estimated, that is, measured with an error. The effect of this is typically to bias the estimator of λ_t towards zero (and any intercept, or mean of the residual, is biased upward). One way to minimize this problem, used by Fama and MacBeth (1973), is to let the assets be portfolios of assets, for which we can expect some of the individual noise in the first-

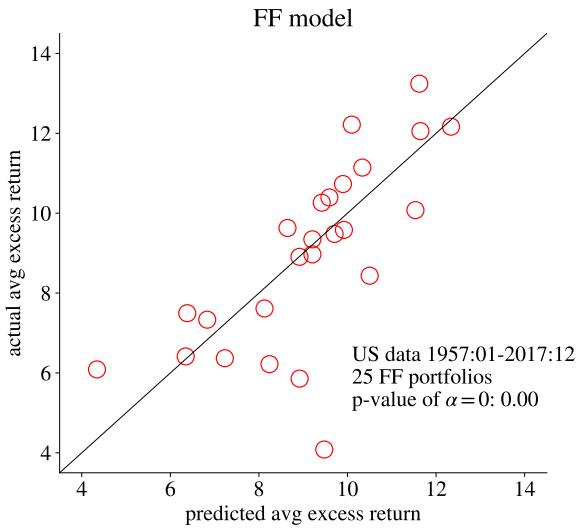


Figure 5.6: FF, FF portfolios

step regressions to average out—and thereby make the measurement error in $\hat{\beta}_i$ smaller. If CAPM is true, then the return of an asset is a linear function of the market return and an error which should be uncorrelated with the errors of other assets—otherwise some factor is missing. If the portfolio consists of 20 assets with equal error variance in a CAPM regression, then we should expect the portfolio to have an error variance which is 1/20th as large.

We clearly want portfolios which have different betas, or else the second step regression (5.16) does not work. Fama and MacBeth (1973) choose to construct portfolios according to some initial estimate of asset specific betas. Another way to deal with the errors-in-variables problem is to adjust the tests.

We can test the model by studying if $\varepsilon_i = 0$ (recall from (5.17) that ε_i is the time average of the residual for asset i , ε_{it}), by forming a t-test $\hat{\varepsilon}_i / \text{Std}(\hat{\varepsilon}_i)$. Fama and MacBeth (1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\varepsilon}_{it}$. In particular, they suggest that the variance of $\hat{\varepsilon}_{it}$ (not $\hat{\varepsilon}_i$) can be estimated by the (average) squared variation around its mean

$$\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T} \sum_{t=1}^T (\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2. \quad (5.19)$$

Since $\hat{\varepsilon}_i$ is the sample average of $\hat{\varepsilon}_{it}$, the variance of the former is the variance of the latter

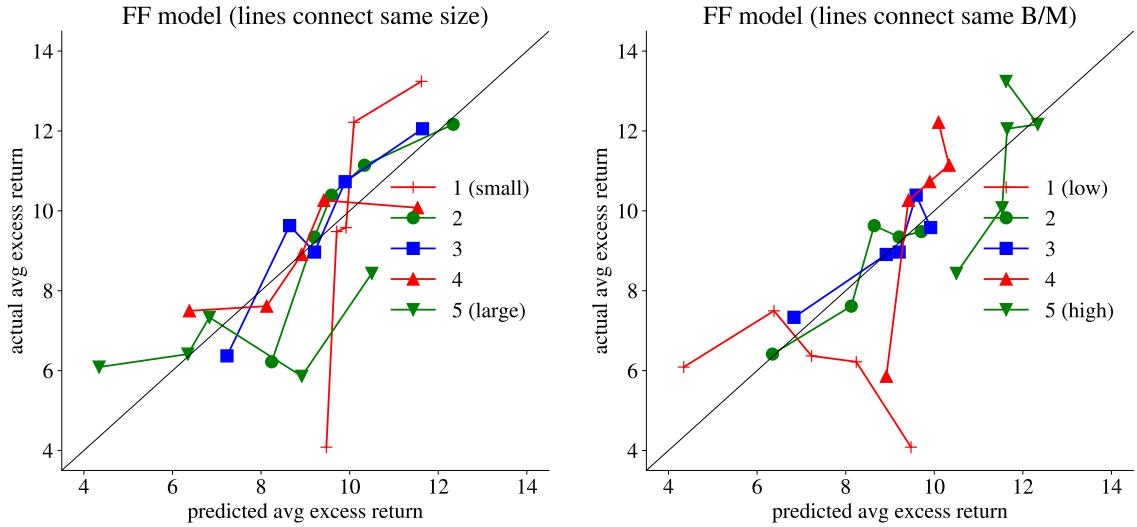


Figure 5.7: FF, FF portfolios

divided by T (the sample size)—provided $\hat{\varepsilon}_{it}$ is iid. That is,

$$\text{Var}(\hat{\varepsilon}_i) = \frac{1}{T} \text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2. \quad (5.20)$$

A similar argument leads to the variance of $\hat{\lambda}$

$$\text{Var}(\hat{\lambda}) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\lambda}_t - \hat{\lambda})^2. \quad (5.21)$$

Fama and MacBeth (1973) found, among other things, that the squared beta is not significant in the second step regression, nor is a measure of non-systematic risk.

Chapter 6

Model Selection and Other Topics

6.1 Model Selection I

Excluding a relevant regressor will cause a bias of all coefficients (unless those 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 “punish” models with too many parameters.

Remember that R^2 can never decrease by adding more regressors—not really a good guide. To avoid overfitting, we could consider \bar{R}^2 instead

$$\bar{R}^2 = 1 - (1 - R^2) \frac{T - 1}{T - k}, \quad (6.1)$$

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) and the Bayesian information criterion (BIC). They are

$$AIC = \ln \sigma^2 + 2 \frac{k}{T} \quad (6.2)$$

$$BIC = \ln \sigma^2 + \frac{k}{T} \ln T, \quad (6.3)$$

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). Choose the model with the *highest* \bar{R}^2 or *lowest* AIC or BIC. It can be shown (by using $R^2 = 1 - \sigma^2 / \text{Var}(y_t)$ so $\sigma^2 = \text{Var}(y_t)(1 - R^2)$) that AIC and BIC can be rewritten as

$$AIC = \ln \text{Var}(y_t) + \ln(1 - R^2) + 2\frac{k}{T} \quad (6.4)$$

$$BIC = \ln \text{Var}(y_t) + \ln(1 - R^2) + \frac{k}{T} \ln T. \quad (6.5)$$

This shows that both are decreasing in R^2 (which is good), 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 . Recall that the model should always include a constant.

Example 6.1 (*Empirical application of model selection*) See Table 6.1 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 possible regressors are lags of the dependent variable, the VIX index and the the S&P 500 returns. Similarly, Table 6.2 for the the best specification according to AIC. Notice that AIC tend to favour fairly large models with many regressors.

6.2 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 models with many predictors tend to fail out of sample. It then becomes crucial to apply some model selection technique, that is, 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 the AIC and BIC in (6.2)–(6.3) can be used, see Table 6.2. Otherwise, we need some sort of sequential approach.

Example 6.2 (*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) ; $(1, x_1, x_2)$.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RV _{t-1}	0.73 (11.21)						0.19 (1.79)
RV _{t-2}		0.56 (9.72)					0.06 (0.53)
VIX _{t-1}			0.95 (9.85)				0.86 (5.09)
VIX _{t-2}				0.71 (9.15)			-0.26 (-1.33)
R _{t-1}					-0.96 (-3.25)		-0.20 (-1.70)
R _{t-2}						-0.57 (-2.24)	-0.11 (-1.09)
constant	4.06 (4.78)	6.65 (8.02)	-3.15 (-1.97)	1.44 (1.16)	15.78 (16.59)	15.50 (15.84)	0.17 (0.18)
R ²	0.53	0.31	0.62	0.34	0.19	0.07	0.65
\bar{R}^2	0.53	0.31	0.62	0.34	0.18	0.06	0.64
obs	317.00	317.00	317.00	317.00	317.00	317.00	317.00

Table 6.1: Regression of monthly realized S&P 500 return volatility 1990:02-2017:12. 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
 - (2) add the variable that improves the fit the most
 - (3) repeat (2) until the fit does not improve much
- (6.6)

To specify a stopping rule, first define the residual sum of squares (for a given vector of coefficients, β) as

$$RSS(\beta) = \sum_{t=1}^T (y_t - x'_t \beta)^2. \quad (6.7)$$

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

$$\frac{RSS(\hat{\beta}_{\text{old}}) - RSS(\hat{\beta}_{\text{new}})}{RSS(\hat{\beta}_{\text{new}})/(T - k - 1)} < c_{1,T-k-1}, \quad (6.8)$$

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 $F_{1,T-k-1}$ distri-

	(1)	(2)	(3)	(4)
RV _{t-1}	0.20 (2.19)	0.23 (2.24)		0.23 (2.17)
RV _{t-2}				
VIX _{t-1}	0.66 (7.52)	0.96 (7.40)	0.88 (12.13)	0.81 (5.07)
VIX _{t-2}		-0.31 (-2.23)		-0.18 (-1.10)
R _{t-1}	-0.34 (-2.71)		-0.32 (-2.68)	-0.21 (-1.81)
R _{t-2}				
constant	-0.44 (-0.52)	-0.90 (-1.08)	-1.60 (-1.39)	-0.37 (-0.44)
R ²	0.65	0.65	0.64	0.65
BIC	3.46	3.46	3.46	3.47
obs	317.00	317.00	317.00	317.00

Table 6.2: Regression of monthly realized S&P 500 return volatility 1990:02-2017:12. Ordered from best (1) according to BIC to fourth best (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

bution. For instance, the 90% critical value of $F_{1,100}$ equals 2.76.

As an alternative to the RSS based rule in (6.7)–(6.8), we could instead use t-stats: in step (2) add the variable with the highest $|t\text{-stat}|$ and in step (3) stop adding variables when that $|t\text{-stat}|$ is lower than 1.64 (or 1.96).

Example 6.3 (*Forward stepwise selection*) Applying the forward step selection approach (based on t-stats) to the regression discussed in Example 6.1 gives a sequence of larger and larger models shown in Table 6.3.

6.2.1 The Lasso Method*

An alternative approach to model selection is the *lasso method*, which minimizes the sum of squared residuals (just like OLS), but under a restriction that the sum of the absolute value of the coefficients should not exceed a threshold t . In short, it solves the following

	(1)	(2)	(3)	(4)
RV _{t-1}			0.20 (2.19)	0.23 (2.17)
RV _{t-2}				
VIX _{t-1}	0.95 (9.85)	0.88 (12.13)	0.66 (7.52)	0.81 (5.07)
VIX _{t-2}				-0.18 (-1.10)
R _{t-1}		-0.32 (-2.68)	-0.34 (-2.71)	-0.21 (-1.81)
R _{t-2}				
constant	-3.15 (-1.97)	-1.60 (-1.39)	-0.44 (-0.52)	-0.37 (-0.44)
R ²	0.62	0.64	0.65	0.65
obs	317.00	317.00	317.00	317.00

Table 6.3: Best four regressions of monthly realized S&P 500 return volatility according to a forward step selection (based on t-stats), 1990:02-2017:12. Ordered from smallest model (1) to fourth smallest model (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

constrained optimization problem

$$\min_b \sum_{t=1}^T (y_t - \alpha - x_t' b)^2 \text{ subject to } \sum_{i=1}^K |b_i| \leq t, \quad (6.9)$$

where the value of t is chosen a priori. (This problem can be solved by brute force minimization if there are few regressors. Otherwise, the “lars” algorithm by Efron, Hasti, Johnstone, and Tibshirani (2004) is very efficient and can handle large problems.)

Clearly, when $t \geq \sum_{i=1}^K |\hat{b}_i|$ where \hat{b}_i are the OLS estimates, then the lasso approach reproduces the OLS estimates. For smaller values of t , the lasso will give smaller coefficients: some b_i will be zero and others tend to be closer to zero than OLS would suggest (like other “shrinkage” methods like a ridge estimation).

The lasso method can be used as a model selection technique by estimating a sequence of models with increasingly higher t values. With a sufficiently low t , only one coefficient is non-zero—for a somewhat higher t value, two coefficients are non-zero and so on. (If we solve (6.9) with brute force, then this might involve some tweaking of the sequence of the t values. However, the lars algorithm does this automatically.) Once the L (five, say)

	(1)	(2)	(3)	(4)
RV _{t-1}		0.17 (2.02)	0.20 (2.19)	0.18 (1.72)
RV _{t-2}				
VIX _{t-1}	0.95 (9.85)	0.77 (9.02)	0.66 (7.52)	0.68 (7.13)
VIX _{t-2}				
R _{t-1}			-0.34 (-2.71)	-0.35 (-2.79)
R _{t-2}				-0.06 (-0.63)
constant	-3.15 (-1.97)	-2.26 (-1.87)	-0.44 (-0.52)	-0.28 (-0.34)
R ²	0.62	0.63	0.65	0.65
$\sum b_i $	0.79	0.81	0.91	0.93
obs	317.00	317.00	317.00	317.00

Table 6.4: Best four regressions of monthly realized S&P 500 return volatility where the model are selected by lasso, but then estimated with OLS, 1990:02-2017:12. Ordered from smallest model (1) to fourth smallest model (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags. The $\sum |b_i|$ is for regression using standardized variables.

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).)

Example 6.4 (*lasso regression*) Applying the lasso approach to the regression discussed in Example 6.1 gives a sequence of larger and larger models. Re-estimating the four smallest of those models with OLS gives the results in Table 6.4.

Remark 6.5 (*Ridge regression**) The ridge regression solves $\min_b \sum_{t=1}^T (y_t - \alpha - x'_t b)^2 + \lambda \sum_{i=1}^K b_i^2$, where $\lambda > 0$, so it forms a compromise between OLS and zero coefficients. This is easiest to see if y_t and x_t are demeaned so $\alpha = 0$. Then, the first order conditions for minimization are $\sum_{t=1}^T x_t (y_t - x'_t \tilde{b}) - \lambda \tilde{b} = 0$, so $\tilde{b} = (\frac{1}{T} \sum_{t=1}^T x_t x'_t + \lambda)^{-1} \frac{1}{T} \sum_{t=1}^T x_t y_t$. Notice that $\lambda = 0$ gives OLS, while $\lambda = \infty$ gives $\tilde{b} = \mathbf{0}$.

Remark 6.6 (*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. In some cases, they also calculate $b_i \sqrt{T}$ instead of b_i .

6.3 Comparing Non-Nested Models

Consider two competing models

$$\text{Model A: } y_t = x'_t \beta + \varepsilon_t \quad (6.10)$$

$$\text{Model B: } y_t = z'_t \gamma + v_t. \quad (6.11)$$

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

$$y_t = z'_t \gamma + x'_{2t} \delta_A + v_t, \quad (6.12)$$

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 .

6.4 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*

$$y_t = \alpha + \delta x_t^{3.4} + \varepsilon_t, \quad (6.13)$$

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*

$$y_t = \beta_1 + \beta_2 x_t^{\beta_3} + u_t \quad (6.14)$$

cannot be estimated by OLS. Do nonlinear LS (NLS) instead. This requires the use of a numerical minimization routine to minimize the sum of squared residuals, $\sum_{t=1}^T u_t^2$.

To *test the functional form* (...is a linear specification really correct?), estimate non-

linear extension and test if they are significant. Alternatively, do a RESET test

$$y_t = x'_t \beta + \alpha_2 \hat{y}_t^2 + v_t, \quad (6.15)$$

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.

6.5 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—to see if there are extreme data points.

Since the loss function defining OLS is quadratic, a few outliers can easily have a very large influence on the estimated coefficients. For instance, suppose the true model is $y_t = 0.75x_t + u_t$, and that the residual is very large for some time period s .

Consider the loss function at a slope of 0.75 (the true value, actually): it would be large due to the u_s^2 term. The loss function value will probably be lower if the coefficient is changed to pick up the y_s observation—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 6.1.

There is of course nothing sacred about the 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. This is illustrated in Figure 6.1.

However, LS is by far the most popular choice. There are two main reasons: LS is very easy to compute and it is fairly straightforward to construct standard errors and confidence intervals for the estimator. (From an econometric point of view you may want to add that LS coincides with maximum likelihood when the errors are normally distributed.)

As complement, 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 $\hat{\beta}^{(s)}$ 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 $\hat{\beta}^{(s)} - \hat{\beta}$, $\hat{y}_s^{(s)} - \hat{y}_t$ or $\hat{u}_s^{(s)} / \hat{\sigma}$.

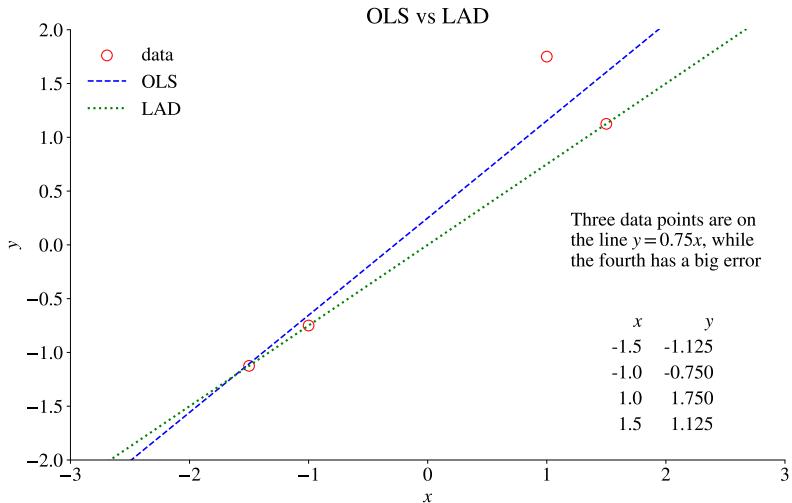


Figure 6.1: Data and regression line from OLS and LAD

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).

6.6 Estimation on Subsamples

To *test for a structural break* of (one or more) coefficients, 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 subset of x_t)

$$y_t = x_t' \beta + g_t z_t' \delta + \varepsilon_t, \text{ where} \quad (6.16)$$

$$g_t = \begin{cases} 1 & \text{for some subsample} \\ 0 & \text{else} \end{cases} \quad (6.17)$$

and test $\delta = \mathbf{0}$ (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).

To capture *time-variation in the regression coefficients*, it is fairly common to run the regression

$$y_t = x_t' \beta + \varepsilon_t \quad (6.18)$$

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. See Figure 6.2 for an illustration.

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 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

$$\tilde{x}_t = x_t \lambda^{T-t} \text{ and } \tilde{y}_t = y_t \lambda^{T-t} \quad (6.19)$$

and then estimate

$$\tilde{y}_t = \tilde{x}'_t \beta + \varepsilon_t. \quad (6.20)$$

Notice that also the constant (in x_t) should be scaled in the same way. Again, see Figure 6.2 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). Again, see Figure 6.2 for an illustration.

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. See 6.3 for an illustration.

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.

The recursive estimates can be used to construct another formal test of structural breaks, the *CUSUM test* (see, for instance, Enders (2004)). First, do a regression on the sample $t = 1, 2, \dots, \tau$ and use the estimated coefficients (denoted $\beta^{(\tau)}$) to calculate a “forecast” and “forecast error” for $\tau + 1$ as

$$\hat{y}_{\tau+1} = x'_{\tau+1} \beta^{(\tau)} \text{ and } v_{\tau+1} = y_{\tau+1} - \hat{y}_{\tau+1}. \quad (6.21)$$

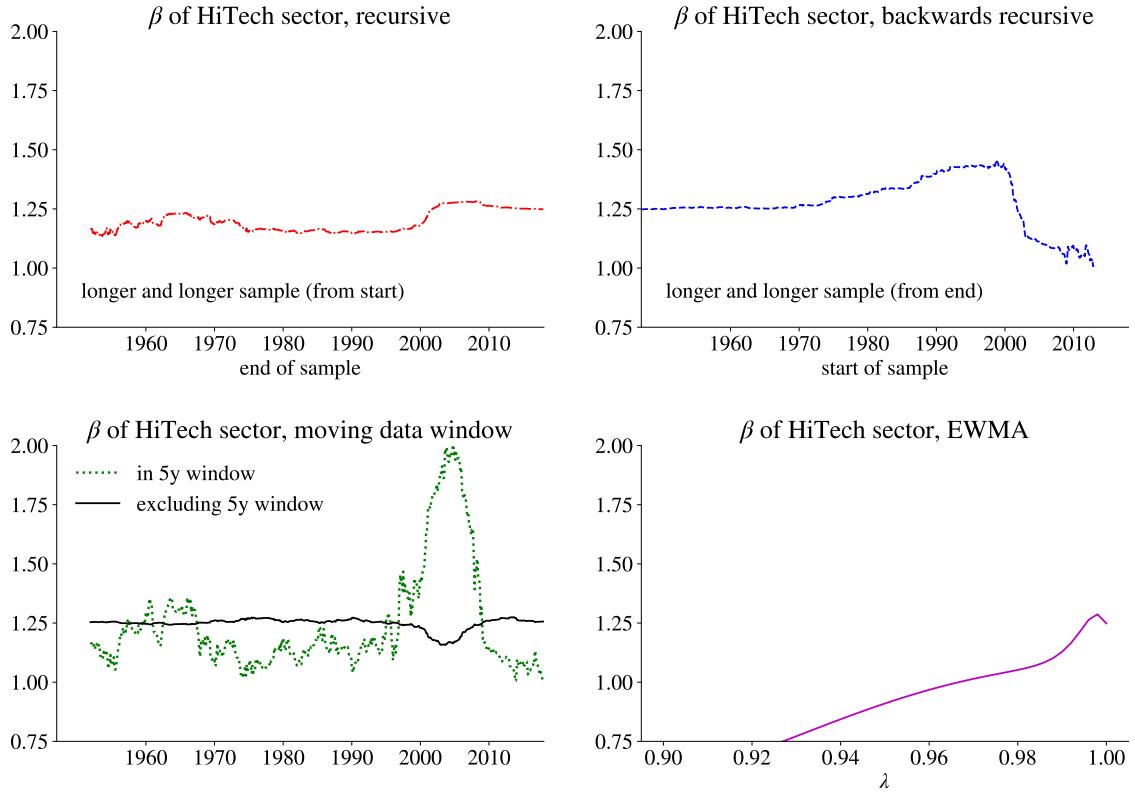


Figure 6.2: Betas of US industry portfolios

Second, do a second estimation on the sample $t = 1, 2, \dots, \tau + 1$ and calculate

$$\hat{y}_{\tau+2} = x'_{\tau+2}\beta^{(\tau+1)} \text{ and } v_{\tau+2} = y_{\tau+2} - \hat{y}_{\tau+2}. \quad (6.22)$$

Third, do the same for all other samples (observation 1 to $\tau + 2$, observation 1 to $\tau + 3$, etc). Forth, calculate the standard deviation of those forecast errors (denoted σ below). Fifth, calculate a corresponding sequence of cumulative sums of standardized residuals

$$W_\tau = \frac{v_{\tau+1}}{\sigma}, W_{\tau+1} = \frac{v_{\tau+1} + v_{\tau+2}}{\sigma} \quad (6.23)$$

and so forth. More generally we have the sequence

$$W_t = \sum_{s=\tau}^t \frac{v_{s+1}}{\sigma}, \text{ for } t = \tau, \dots, T-1. \quad (6.24)$$

Sixth and finally, plot W_t along with a 95% confidence interval: $\pm 0.948(\sqrt{T-\tau} +$

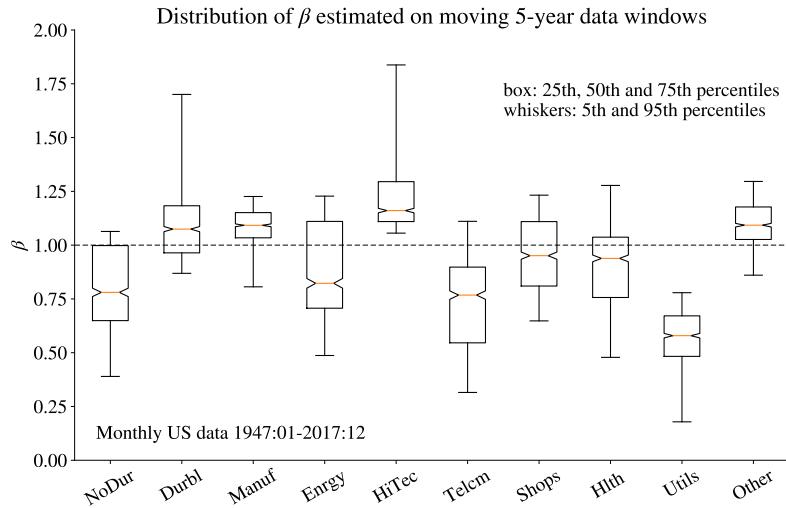


Figure 6.3: Distribution of betas of US industry portfolios (estimated on 5-year data windows)

$2(t - \tau) / \sqrt{T - \tau}$). Reject stability if any observation is outside. See Figure 6.5.

6.7 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 the missing t . 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).

6.8 Robust Estimation*

6.8.1 Robust Means, Variances and Correlations

Outliers and other extreme observations can have very decisive influence on the estimates of the key statistics needed for financial analysis, including mean returns, variances, covariances and also regression coefficients.

The perhaps best way to solve these problems is to carefully analyse the data—and then decide which data points to exclude. Alternatively, robust estimators can be applied instead of the traditional ones.

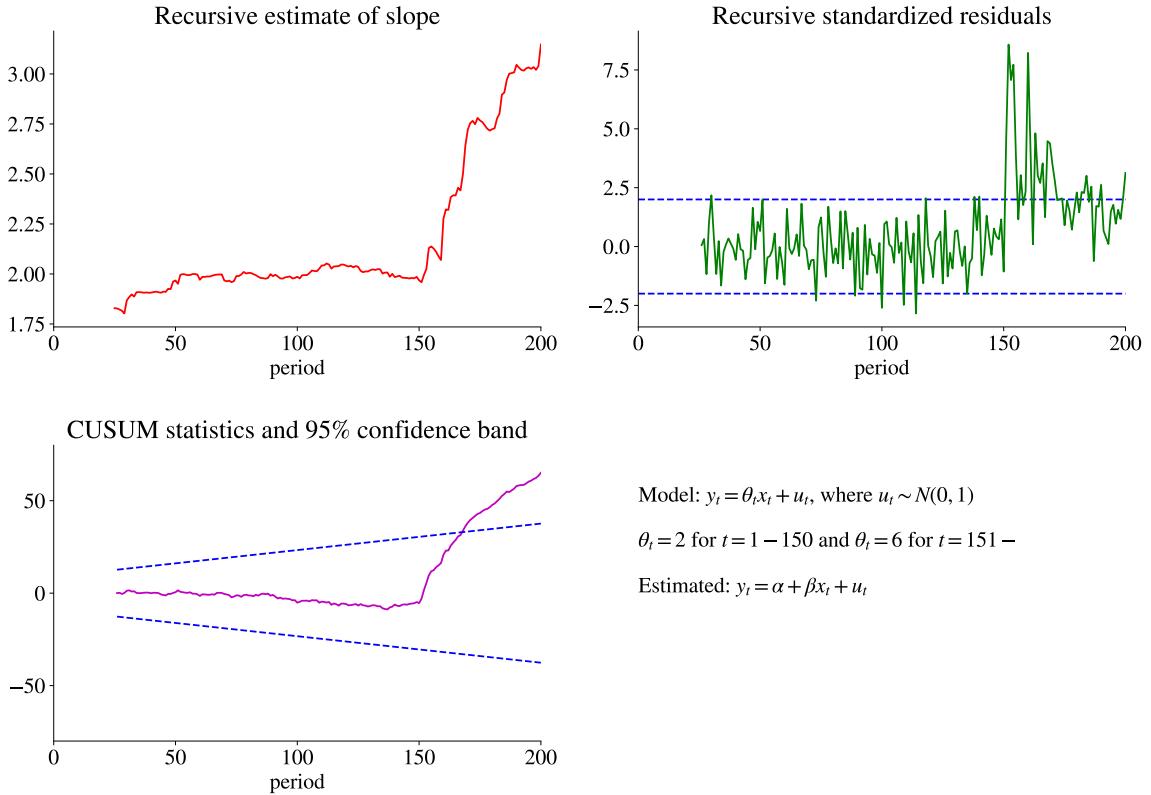


Figure 6.4: Stability test

To estimate the mean, the sample average can be replaced by the *median* or a *trimmed mean* (where the $x\%$ lowest and highest observations are excluded).

Similarly, to estimate the variance, the sample standard deviation can be replaced by the *interquartile range* (the difference between the 75th and the 25th percentiles), divided by 1.35

$$\text{StdRobust} = [\text{quantile}(0.75) - \text{quantile}(0.25)]/1.35, \quad (6.25)$$

or by the *median absolute deviation*

$$\text{StdRobust} = \text{median}(|x_t - \mu|)/0.675. \quad (6.26)$$

Both these would coincide with the standard deviation if data was indeed drawn from a normal distribution without outliers.

A robust covariance can be calculated by using the identity

$$\text{Cov}(x, y) = [\text{Var}(x + y) - \text{Var}(x - y)]/4 \quad (6.27)$$

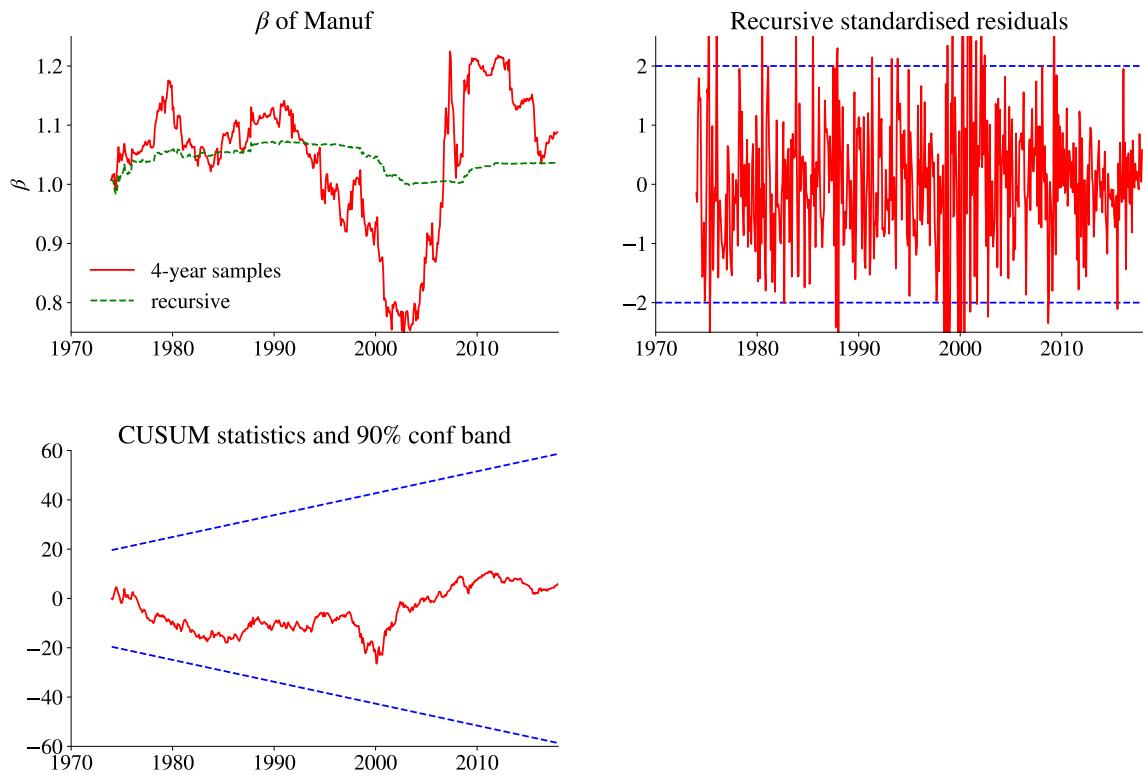


Figure 6.5: CAPM regression on a US industry index

and using a robust estimator of the variances—like the square of (6.25). A robust correlation is then created by dividing the robust covariance with the two robust standard deviations.

See Figures 6.6–6.7 for empirical examples.

6.8.2 Robust Regression Coefficients

Reference: Amemiya (1985) 4.6

The least absolute deviations (LAD) estimator minimizes the sum of absolute residuals (rather than the squared residuals)

$$\hat{\beta}_{LAD} = \arg \min_b \sum_{t=1}^T |y_t - x'_t b| \quad (6.28)$$

This estimator involves non-linearities, but a simple iteration works nicely. It is typically less sensitive to outliers than OLS. (There are also other ways to estimate robust regression

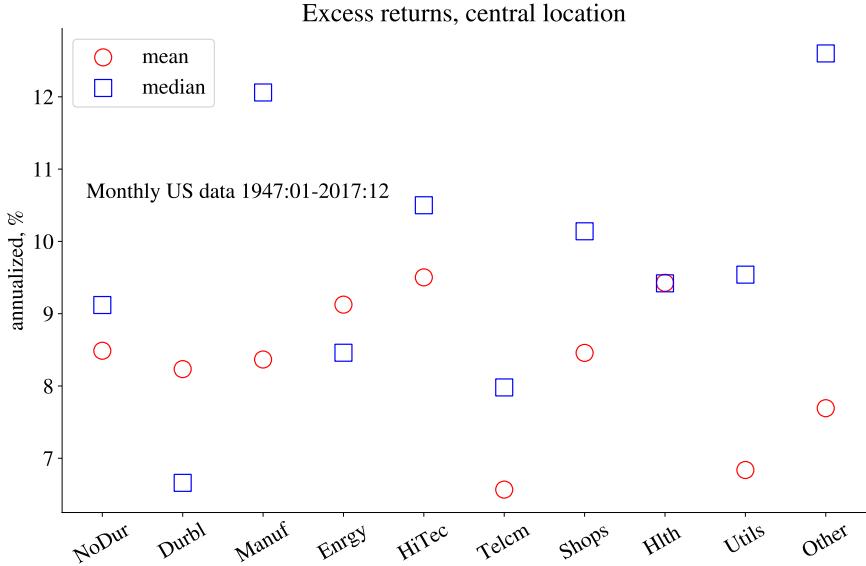


Figure 6.6: Mean excess returns of US industry portfolios

coefficients.) This is illustrated in Figure 6.1. See Figure 6.8 for an empirical example.

If we assume that the median of the true residual, u_t , is zero, then we (typically) have

$$\sqrt{T}(\hat{\beta}_{LAD} - \beta_0) \xrightarrow{d} N[0, f(0)^{-2} \Sigma_{xx}^{-1}/4] \quad (6.29)$$

where Σ_{xx} is (the probability limit of) $\sum_{t=1}^T x_t x'_t / T$ and where $f(0)$ is the value of the pdf of the residual at zero. In practice, Σ_{xx} is estimated by data (as $\sum_{t=1}^T x_t x'_t / T$). Unless we know density function $f()$ (or else we would probably have used MLE instead of LAD), we need to estimate it—for instance with a kernel density method.

Example 6.7 ($N(0, \sigma^2)$) When $u_t \sim N(0, \sigma^2)$, then $f(0) = 1/\sqrt{2\pi\sigma^2}$, so the covariance matrix in (6.29) becomes $\pi\sigma^2 \Sigma_{xx}^{-1}/2$. This is $\pi/2$ times larger than when using LS.

Remark 6.8 (Algorithm for LAD) The LAD estimator can be written

$$\hat{\beta}_{LAD} = \arg \min_{\beta} \sum_{t=1}^T w_t \hat{u}_t(b)^2, \quad w_t = 1/|\hat{u}_t(b)|, \text{ with } \hat{u}_t(b) = y_t - x'_t \hat{b}$$

so it is a weighted least squares where both y_t and x_t are multiplied by $1/|\hat{u}_t(b)|$. It can be shown that iterating on LS with the weights given by $1/|\hat{u}_t(b)|$, where the residuals are from the previous iteration, converges very quickly to the LAD estimator.

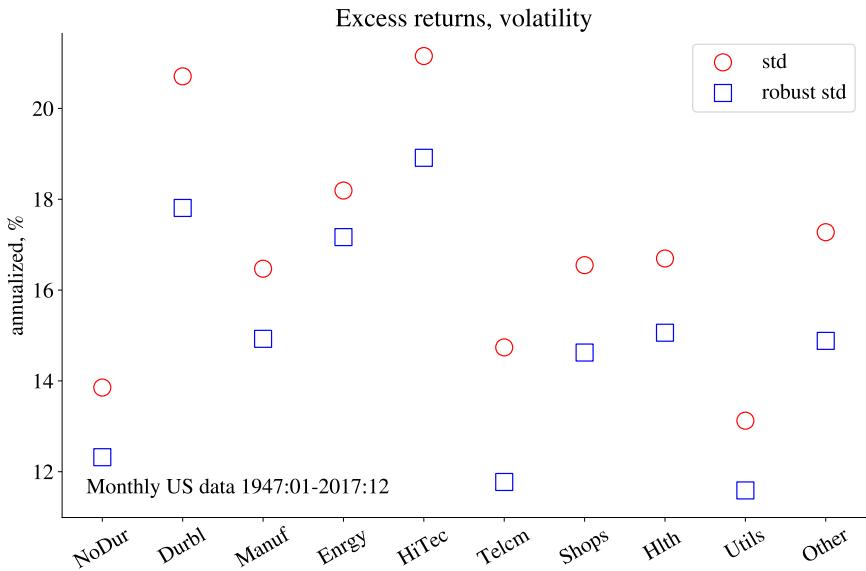


Figure 6.7: Volatility of US industry portfolios

Some alternatives to LAD: least median squares (LMS), and least trimmed squares (LTS) estimators which solve

$$\hat{\beta}_{LMS} = \arg \min_{\beta} [\text{median}(\hat{u}_t^2)], \text{ with } \hat{u}_t = y_t - x_t' \hat{b} \quad (6.30)$$

$$\hat{\beta}_{LTS} = \arg \min_{\beta} \sum_{i=1}^h \hat{u}_i^2, \hat{u}_1^2 \leq \hat{u}_2^2 \leq \dots \text{ and } h \leq T. \quad (6.31)$$

Note that the LTS estimator in (6.31) minimizes the sum of the h smallest squared residuals.

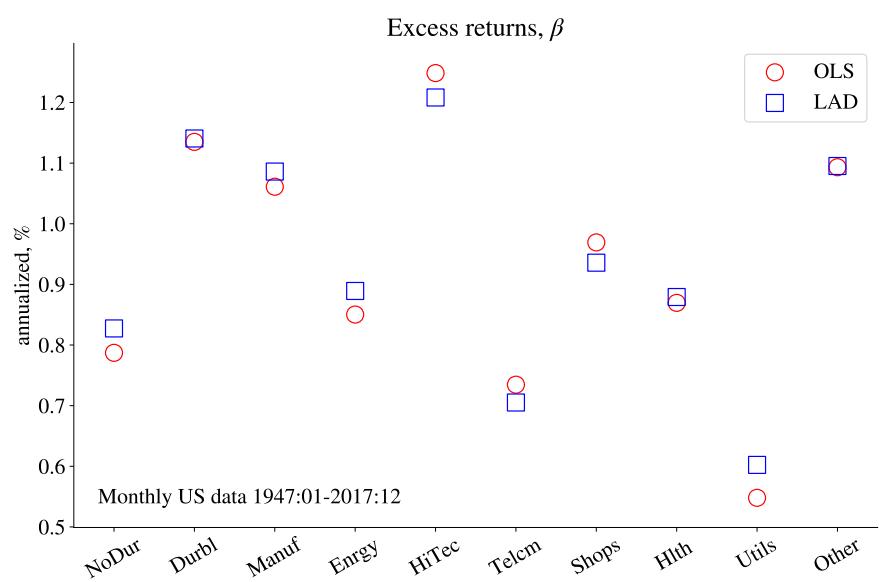


Figure 6.8: Betas of US industry portfolios

Chapter 7

Asymptotic Results on OLS*

Reference: Verbeek (2012) 2 and 5

7.1 Motivation of Asymptotics

There are several problems when the standard assumptions about linear regressions are wrong. First, the result that $E\hat{\beta} = \beta$ (unbiased) relies on the assumption that the regressors are fixed or alternatively that $\{u_1, \dots, u_T\}$ and $\{x_1, \dots, x_T\}$ are independent. Otherwise not true (in a finite sample)—see Figure 7.1. 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).

What *is* true when these assumptions are not satisfied? How should we test hypotheses? Two ways to find answers: (a) do computer (Monte Carlo or bootstrap) simulations; (b) find results for $T \rightarrow \infty$ (“asymptotic properties”) and use as approximation.

The results from asymptotic theory are more general (and prettier) than simulations—and can be used as approximation if the sample is large. The basic reasons for this is that most estimators are sample averages and sample averages often have nice properties as $T \rightarrow \infty$. In particular, we can make use of the law of large numbers (LLN) and the central limit theorem (CLT). See Figure 7.2.

However, the asymptotic results are unlikely to be good approximations in small samples. In those cases we need simulations.

7.2 Asymptotics: Consistency

Issue: will our estimator come closer to the truth as the sample size increases? If not, use another estimator (method).

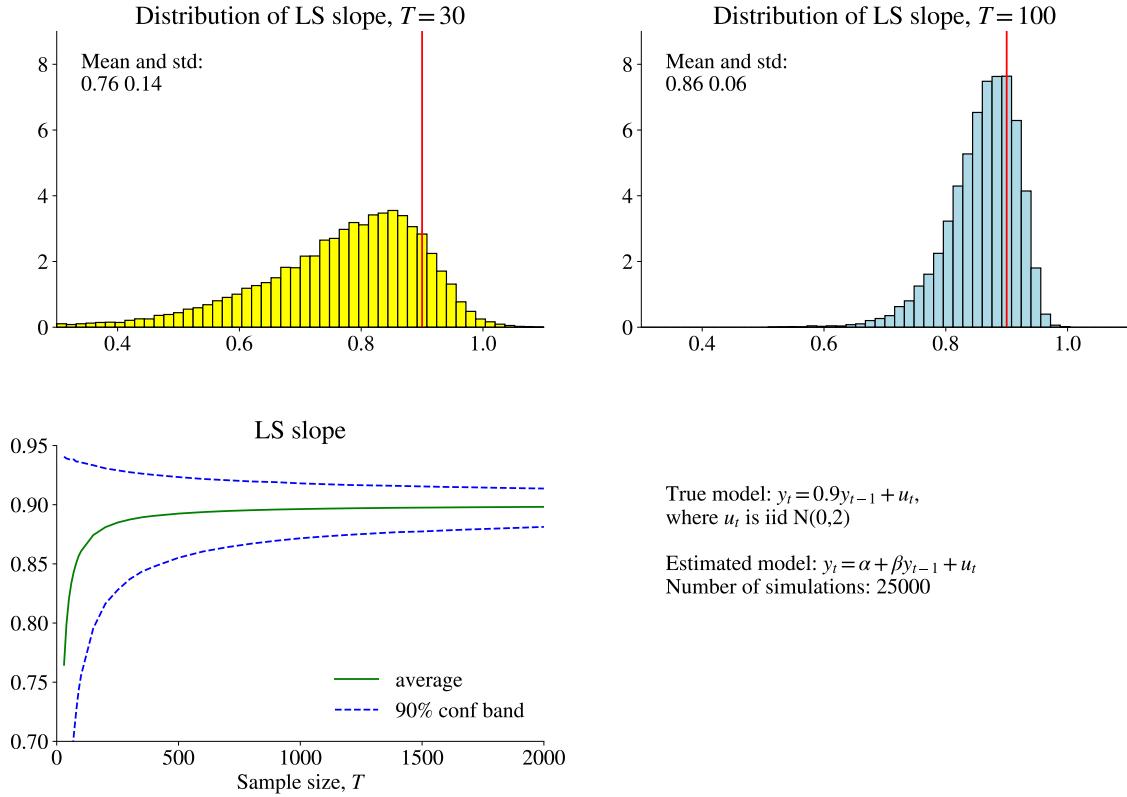


Figure 7.1: Distribution of LS estimator of autoregressive parameter

Remark 7.1 (Consistency) *Consistency means that the estimate $\hat{\beta}$ converges to the true value as the sample size increases (to infinity). In formal terms, $\text{Prob}(\hat{\beta} \text{ deviates much from } \beta) \rightarrow 0$ as $T \rightarrow \infty$. Notation: $\text{plim } \hat{\beta} = \beta$.*

Remark 7.2 (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 means that the sample average is a consistent estimator of the population mean. Notation: $\text{plim}(\bar{x}) = E(x)$.*

The OLS estimate of a slope coefficient is

$$\hat{\beta} = \beta + \underbrace{\left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)^{-1}}_{\rightarrow \Sigma_{xx}^{-1}} \underbrace{\frac{1}{T} \sum_{t=1}^T x_t u_t}_{\rightarrow E(x,u)} \quad (7.1)$$

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

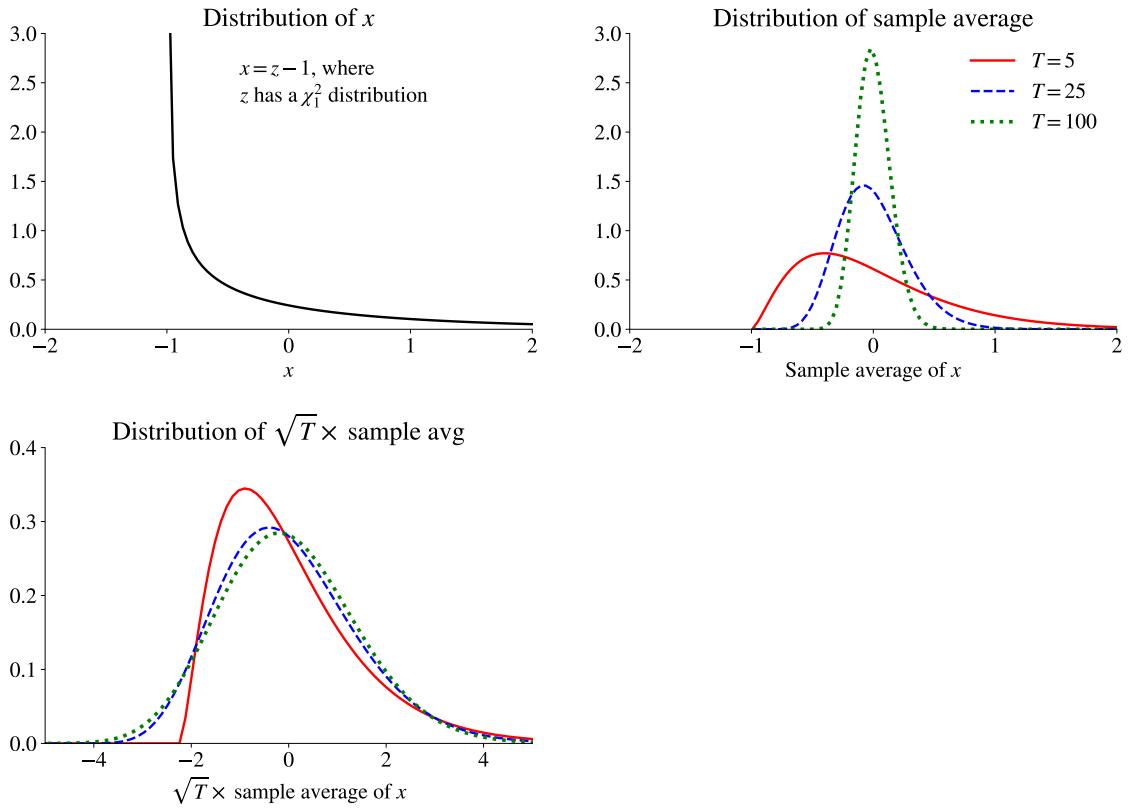


Figure 7.2: Distribution of sample averages

different terms converge to (according to a LLN) as the sample size increases. In short, the probability limit is

$$\text{plim } \hat{\beta} = \beta + \Sigma_{xx}^{-1} E(x_t u_t). \quad (7.2)$$

The term Σ_{xx}^{-1} is a matrix of constants. Clearly, for the estimate $\hat{\beta}$ to converge to the true values (β) , $E(x_t u_t) = 0$ is needed. If $E u_t = 0$ (which is a basic assumption in most regression analysis) $E(x_t u_t) = \text{Cov}(x_t, u_t)$, so consistency of $\hat{\beta}$ requires the regressors and the (true) residuals to be uncorrelated.

Some observations:

1. We can not (easily) test this, since OLS creates $\hat{\beta}$ and the fitted residuals \hat{u}_t such that $\sum_{t=1}^T x_t \hat{u}_t / T = 0$.
2. The standard regression assumption that u_t and x_j are independent implies that $E(x_t u_t) = 0$. This means that the standard regression assumptions take it for granted that OLS is consistent.

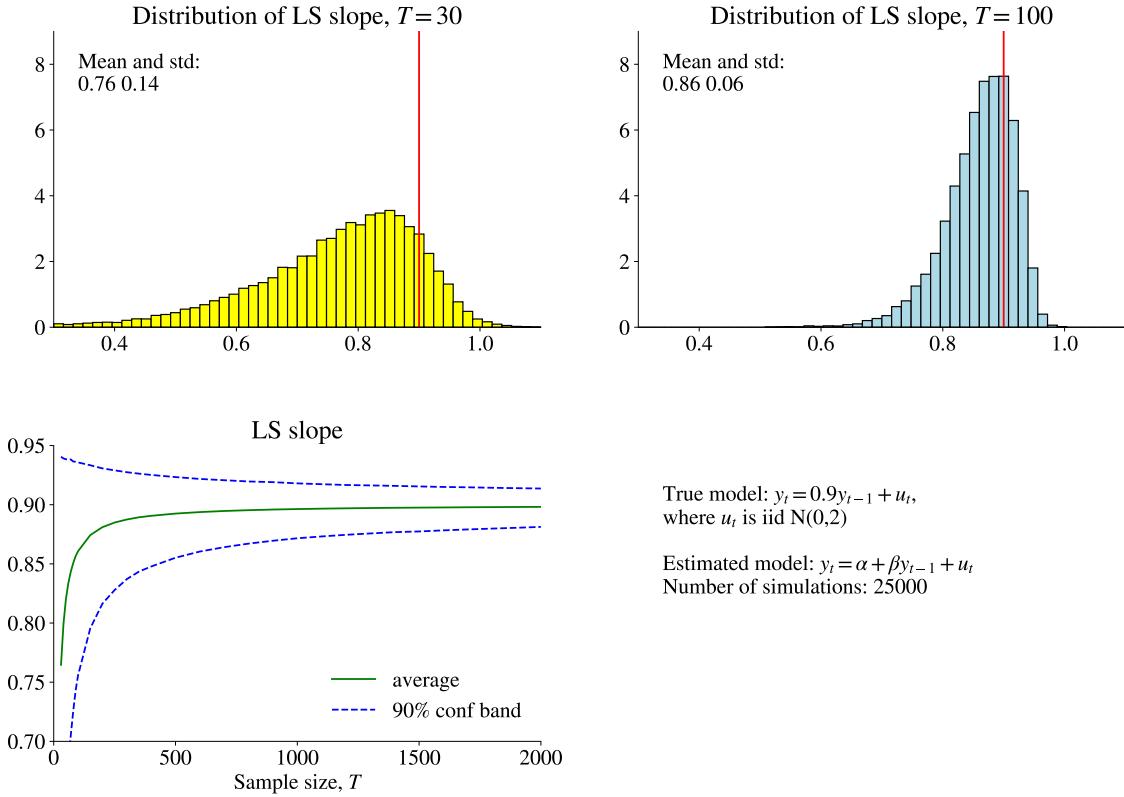


Figure 7.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

3. OLS can be biased, but still be consistent. This means that OLS is systematically wrong in any small sample, but the problem vanishes in large samples. See Figure 18.2. 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.
4. There are cases when $E(x_t u_t) = 0$ doesn't make sense. More on this later.
5. See Figure 18.2 for an example of where OLS is consistent, and Figure 18.3 when it is not.

What have we learned? Well,...under what conditions ($E(x_t u_t) = 0$) OLS comes closer to the truth as T increases.

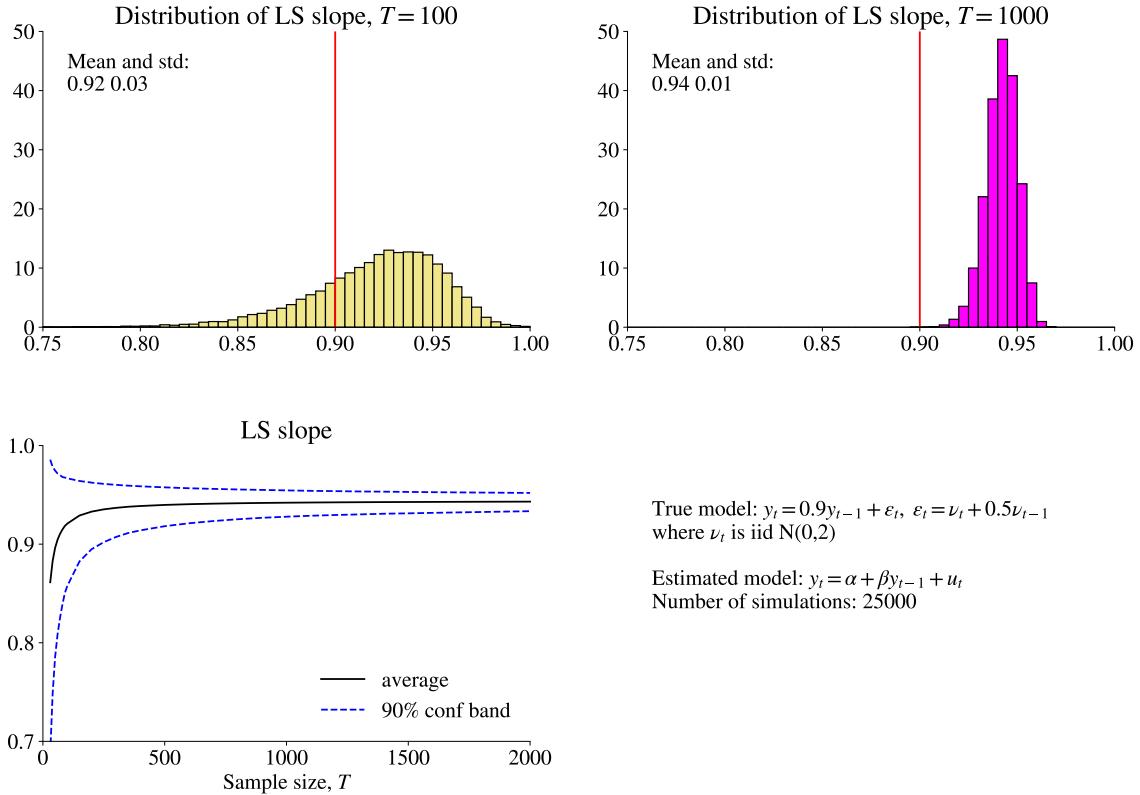


Figure 7.4: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

7.3 When LS Cannot be Saved

...not even in large samples (since it's inconsistent)

Q. When do we have $\text{Cov}(x_t, u_t) \neq 0$?

A. Need to think hard...

But the usual suspects are (i) omitted variables; (ii) autocorrelated errors combined with lagged dependent variable; (iii) measurement errors in regressors; and (iv) endogenous regressors.

7.3.1 When LS Cannot be Saved: Omitted Variables

Consider the regression

$$y_t = x'_t \beta + h'_t \gamma + \varepsilon_t, \quad (7.3)$$

Suppose we omit (exclude) the h_t variables and instead estimate

$$y_t = x'_t \beta + u_t. \quad (7.4)$$

This means that $u_t = h'_t \gamma + \varepsilon_t$, that is, incorporates the effect of both the omitted variables and the “true” residual.

Recall that the OLS estimates are

$$\hat{\beta} = \beta + S_{xx}^{-1} \Sigma_{t=1}^T x_t u_t, \quad (7.5)$$

where $S_{xx} = \Sigma_{t=1}^T x_t x'_t$. Since $u_t = h'_t \gamma + \varepsilon_t$, we can write this as

$$\hat{\beta} = \beta + S_{xx}^{-1} \sum_{t=1}^T x_t h'_t \gamma + S_{xx}^{-1} \sum_{t=1}^T x_t \varepsilon_t. \quad (7.6)$$

The last term should vanish as the sample size increases (the residual in (7.3) should not be correlated with any of the regressors), while the middle term can be written

$$[\hat{\theta}_1 \dots \hat{\theta}_L] \gamma \quad (7.7)$$

where $\hat{\theta}_i$ is the (column) vector of coefficients obtained by regressing h_{it} on x_t

$$\hat{\theta}_i = S_{xx}^{-1} \sum_{t=1}^T x_t h_{it}. \quad (7.8)$$

Together this shows that the probability limit of $\hat{\beta}$ is

$$\text{plim } \hat{\beta} = \beta + [\text{plim } \hat{\theta}_1 \dots \text{plim } \hat{\theta}_L] \gamma. \quad (7.9)$$

This analysis shows that $\hat{\beta}$ incorporates how x_t comoves with the h_t . In case they are uncorrelated ($\theta_i = \mathbf{0}$), then omitting the h_t variables does not affect the point estimates of β . However, if they are correlated, then the point estimates $\hat{\beta}$ are inconsistent (and biased) in the sense of being systematically different from the true β values in (7.3).

Notice the following

- $\hat{\beta}$ from (7.4) 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 $\hat{\beta}$ factors in also how x_t predicts u_t , in particular the $h'_t \gamma$ part of u_t .
- $\hat{\beta}$ from (7.4) is not the 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 .

7.3.2 When LS Cannot be Saved: Autocorrelated Errors Combined with Lagged Dependent Variable

As an example of how autocorrelated errors combined with a lagged dependent variable (as regressor) leads to inconsistent OLS estimates, consider

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t, \text{ where} \quad (7.10)$$

$$u_t = v_t + \theta v_{t-1}, v_t \text{ iid.} \quad (7.11)$$

As a special case, $\beta_2 = 0$ gives an ARMA(1,1) model, which is a well known case which cannot be estimated by OLS. See Figure 18.3.

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.

7.3.3 When LS Cannot be Saved: Measurement Errors in a Regressor

As an example of how measurement errors in a regressor gives inconsistent OLS estimates, consider a simple (true) model like

$$y_t = \beta_1 + \beta_2 w_t + v_t. \quad (7.12)$$

However, we estimate with a proxy x_t for w_t

$$y_t = \beta_1 + \beta_2 x_t + u_t, \text{ with} \quad (7.13)$$

$$x_t = w_t + e_t, \quad (7.14)$$

where e_t is a measurement error. This is a common problem in micro data, including corporate finance. This leads to $\text{Cov}(x_t, u_t) \neq 0$ (since both x_t and u_t depend on the measurement error e_t) so OLS is inconsistent for estimating β_2 . See Figure 7.5.

To see the precise source of the inconsistency, solve for $w_t = x_t - e_t$, use in correct model (7.12) to get

$$\begin{aligned} y_t &= \beta_1 + \beta_2 (x_t - e_t) + v_t \\ &= \beta_1 + \beta_2 x_t - \underbrace{\beta_2 e_t + v_t}_{u_t}. \end{aligned} \quad (7.15)$$

From (7.14) we know that x_t is correlated with the measurement error (e_t), which gives $\text{Cov}(x_t, u_t) \neq 0$. In fact, it can be shown that

$$\text{plim } \hat{\beta}_2 = \beta_2 \left(1 - \frac{\text{Var}(e_t)}{\text{Var}(w_t) + \text{Var}(e_t)} \right). \quad (7.16)$$

Notice that $\hat{\beta}_2 \rightarrow 0$ if the measurement error dominates ($\text{Var}(e_t) \rightarrow \infty$), since y_t is not related to the measurement error. In contrast, $\hat{\beta}_2 \rightarrow \beta_2$ as measurement vanishes ($\text{Var}(e_t) \rightarrow 0$): no measurement error.

Proof. (of (7.16)) To simplify, assume that x_t has a zero mean. From (7.2), we then have $\text{plim } \hat{\beta}_2 = \beta_2 + \Sigma_{xx}^{-1} \text{E}(x_t u_t)$. Here, $\Sigma_{xx}^{-1} = 1 / \text{Var}(x_t)$, but notice from (7.14) that $\text{Var}(x_t) = \text{Var}(w_t) + \text{Var}(e_t)$ 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 (7.14) and of u_t in (7.15) gives

$$\text{Cov}(x_t, u_t) = \text{Cov}(w_t + e_t, -\beta_2 e_t + v_t) = -\beta_2 \text{Var}(e_t).$$

Together we get

$$\text{plim } \hat{\beta}_2 = \beta_2 + \Sigma_{xx}^{-1} \text{E}(x_t u_t) = \beta_2 - \beta_2 \frac{\text{Var}(e_t)}{\text{Var}(w_t) + \text{Var}(e_t)},$$

which is (7.16). ■

7.3.4 When LS Cannot be Saved: Endogenous Regressors (System of Simultaneous Equations)

As a final example, consider a simplistic model of quantity bought and sold (z_t), the price (p_t) and a demand shifter (y_t , for instance, income)

$$z_t = \beta_1 + \beta_2 p_t + \beta_3 y_t + u_t \text{ (demand curve)}, \quad (7.17)$$

$$p_t = \alpha_1 + \alpha_2 z_t + \varepsilon_t \text{ (inverse supply curve)}. \quad (7.18)$$

Suppose we want to estimate the demand (elasticity) β_2 by regressing (7.17). This time the problem is that $\text{Cov}(p_t, u_t) \neq 0$, so the OLS estimate of β_2 will not be a consistent estimate of the demand elasticity. Rather, it will be a mix of the demand and supply elasticities.

The basic reason is that a shock to demand (u_t) will affect the quantity (z_t), but also the price (p_t , via the supply equation), so $\text{Cov}(p_t, u_t) \neq 0$.

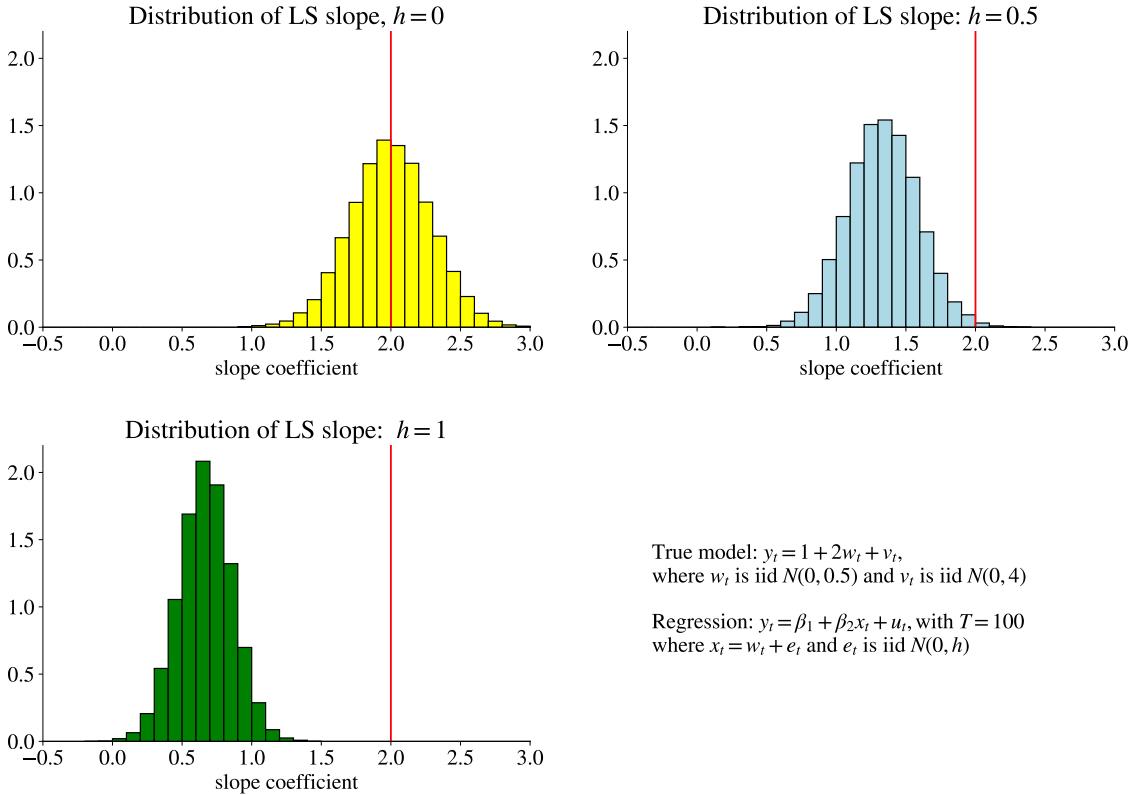


Figure 7.5: Effect of measurement error in regressor, h is the variance of the errors

7.4 Instrumental Variables Method (IV)

When OLS is inconsistent, then we typically apply MLE or the instrumental variables (IV) method. This section describes the latter.

We want to estimate β in

$$y_t = x'_t \beta + u_t, \quad (7.19)$$

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

$$\mathbf{0}_{kx1} = \sum_{t=1}^T x_t (y_t - x'_t \hat{\beta}). \quad (7.20)$$

The IV method replaces this expression with

$$\mathbf{0}_{kx1} = \sum_{t=1}^T z_t (y_t - x'_t \hat{\beta}_{iv}), \quad (7.21)$$

where z_t is a vector of K elements that have two key properties: (1) z_t is uncorrelated

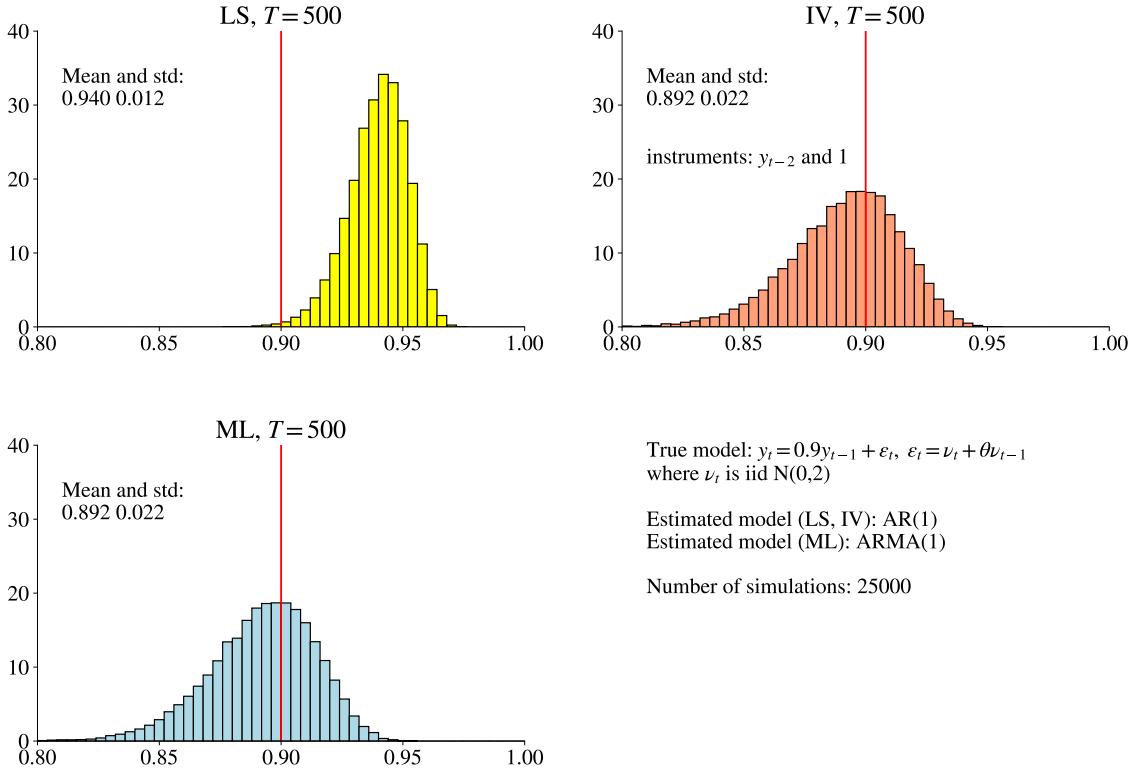


Figure 7.6: Results from a Monte Carlo experiment when data is from an ARMA process.

with the true residual (u_t); but (2) correlated with the regressors (x_t). The first property cannot be directly checked since we never observe the true residuals. Instead, theoretical arguments must be used. In contrast, the second property is easily checked.

Solving (7.21) gives the IV estimator

$$\hat{\beta}_{iv} = \left(\sum_{t=1}^T z_t x_t' \right)^{-1} \sum_{t=1}^T z_t y_t. \quad (7.22)$$

Clearly, this is the same as OLS when $z_t = x_t$.

Notice that we need as many instruments as regressors, but some can be instruments for themselves (if it can be argued that this regressor is not correlated with the true residual).

Figure 7.6 shows an example with an ARMA(1,1) process. 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 IV estimator has often large standard deviations, especially with “weak instruments” (low correlation with regressors)

Remark 7.3 (*Two-stages-least squares, 2SLS*) 2SLS is applied when there are more instruments than regressors. It proceeds in two steps. First, regress $x_t = \delta' z_t + \varepsilon_t$ and define the fitted values $\hat{x}_t = \hat{\delta}' z_t$. Second, regress $y_t = \beta' \hat{x}_t + u_t$. This coincides with (7.22) when z_t has as many elements as x_t .

7.4.1 Distribution of IV Estimator

Use (7.19) to substitute for y_t in (7.22)

$$\hat{\beta}_{iv} = \beta + S_{zx}^{-1} (z_1 u_1 + z_2 u_2 + \dots z_T u_T), \text{ where } S_{zx} = \sum_{t=1}^T z_t x'_t. \quad (7.23)$$

Since we have strong beliefs that $\text{Cov}(z_t, u_t) = 0$, this expression shows that $\hat{\beta}_{iv}$ should be consistent. Under iid assumptions, the variance-covariance is

$$\text{Var}(\hat{\beta}) = S_{zx}^{-1} \left(\sum_{t=1}^T z_t z'_t \right) \sigma^2 S_{zx}^{-1}. \quad (7.24)$$

There are few results on small sample properties, but it is often noticed that IV is often imprecise and even biased.

7.5 Asymptotic Normality

Issue: what is the distribution of your estimator in large samples?

Remark 7.4 (*Central limit theorem, simple version*) A CLT says that $\sqrt{T}\bar{x} \sim N()$ when T becomes really large. This holds for most random variables. Notice that the distribution of \bar{x} converges to a spike as T increases (LLN), but the distribution of $\sqrt{T}\bar{x}$ converges to a normal distribution. See Figure 7.2.

To see how this works, subtract β from both sides of (7.1), and multiply both sides by \sqrt{T} to get

$$\sqrt{T}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{T} \sum_{t=1}^T x_t x'_t \right)^{-1}}_{\rightarrow \Sigma_{xx}^{-1}} \underbrace{\sqrt{T} \frac{1}{T} \sum_{t=1}^T x_t u_t}_{\sqrt{T} \times \text{sample average}} \quad (7.25)$$

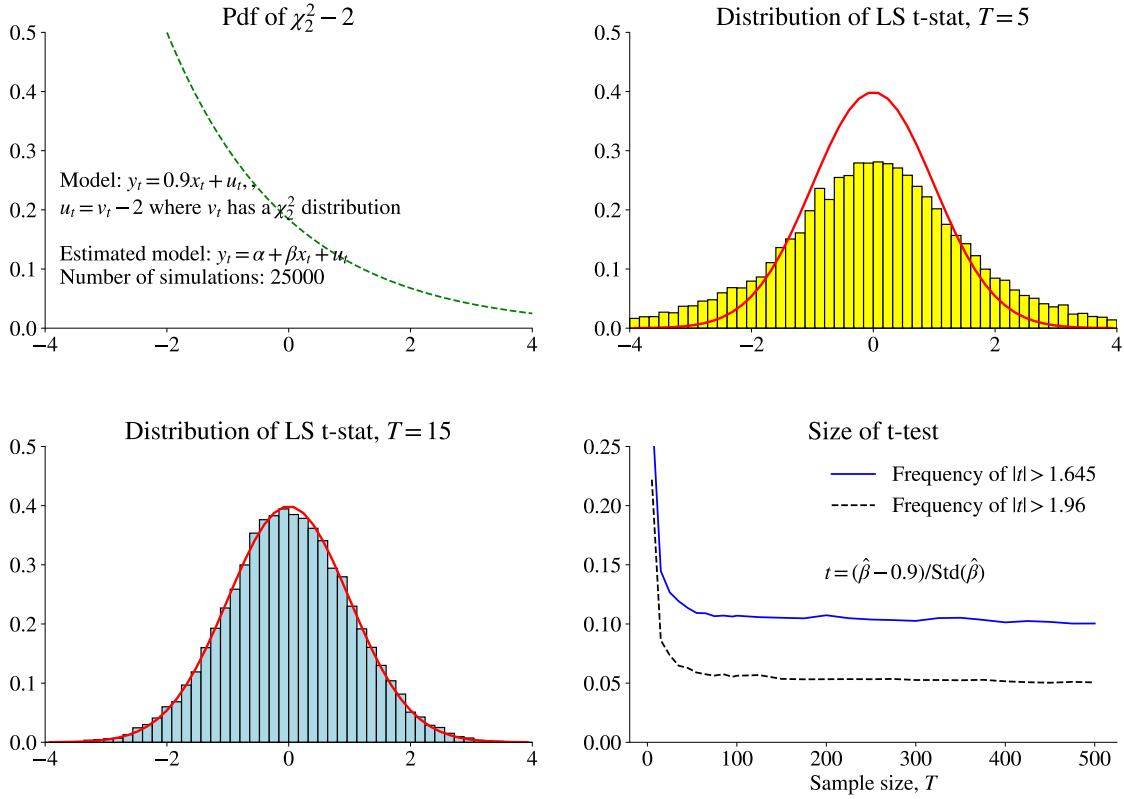


Figure 7.7: Results from a Monte Carlo experiment with thick-tailed errors.

The first term converges (by a LLN) to a constant matrix, while the second term is $\sqrt{T} \times$ sample average (of $x_t u_t$). 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 7.7 for an example (expressed in terms of a *t*-stat).

If an estimator is consistent and asymptotically normal, then use the results as an approximation in large samples

$$\sqrt{T}(\hat{\beta} - \beta) \rightarrow N(0, \sigma^2 \Sigma_{xx}^{-1}) \text{ or } \hat{\beta} \rightarrow N(\beta, \sigma^2 \Sigma_{xx}^{-1}/T) \quad (7.26)$$

Remark 7.5 Step 1: If $\text{Var}[\sqrt{T}(\hat{\beta} - \beta)] = \sigma^2 \Sigma_{xx}^{-1}$, then $\text{Var}[\sqrt{T}(\hat{\beta} - \beta)/\sqrt{T}] = \sigma^2 \Sigma_{xx}^{-1}/T$; step 2: if $E(\hat{\beta} - \beta) = 0$, then $E(\hat{\beta}) = \beta$.

Chapter 8

Time Series Analysis

Reference: Newbold (1995) 17 or Pindyck and Rubinfeld (1998) 13.5, 16.1–2, and 17.2
More advanced material is denoted by a star (*). It is not required reading.

8.1 Descriptive Statistics

The s th *autocovariance* of y_t is estimated by

$$\widehat{\text{Cov}}(y_t, y_{t-s}) = \sum_{t=1}^T (y_t - \bar{y})(y_{t-s} - \bar{y}) / T, \text{ where } \bar{y} = \sum_{t=1}^T y_t / T. \quad (8.1)$$

The convention in time series analysis is that we use the same estimated (using all data) mean in both places and that we divide by T .

The s th *autocorrelation* is estimated as

$$\hat{\rho}_s = \frac{\widehat{\text{Cov}}(y_t, y_{t-s})}{\widehat{\text{Std}}(y_t)^2}. \quad (8.2)$$

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 (not ρ_0 , of course), then for any i and j different from zero

$$\sqrt{T} \begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \quad (8.3)$$

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,

$$\sqrt{T} \hat{\rho}_s \xrightarrow{d} N(0, 1), \quad (8.4)$$

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). See Figure 8.1 for an illustration.

Example 8.1 (*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 $|\hat{\rho}_1| > 1.64/\sqrt{100} = 0.164$ for rejection, and with $T = 1000$ we need $|\hat{\rho}_1| > 1.64/\sqrt{1000} \approx 0.052$.

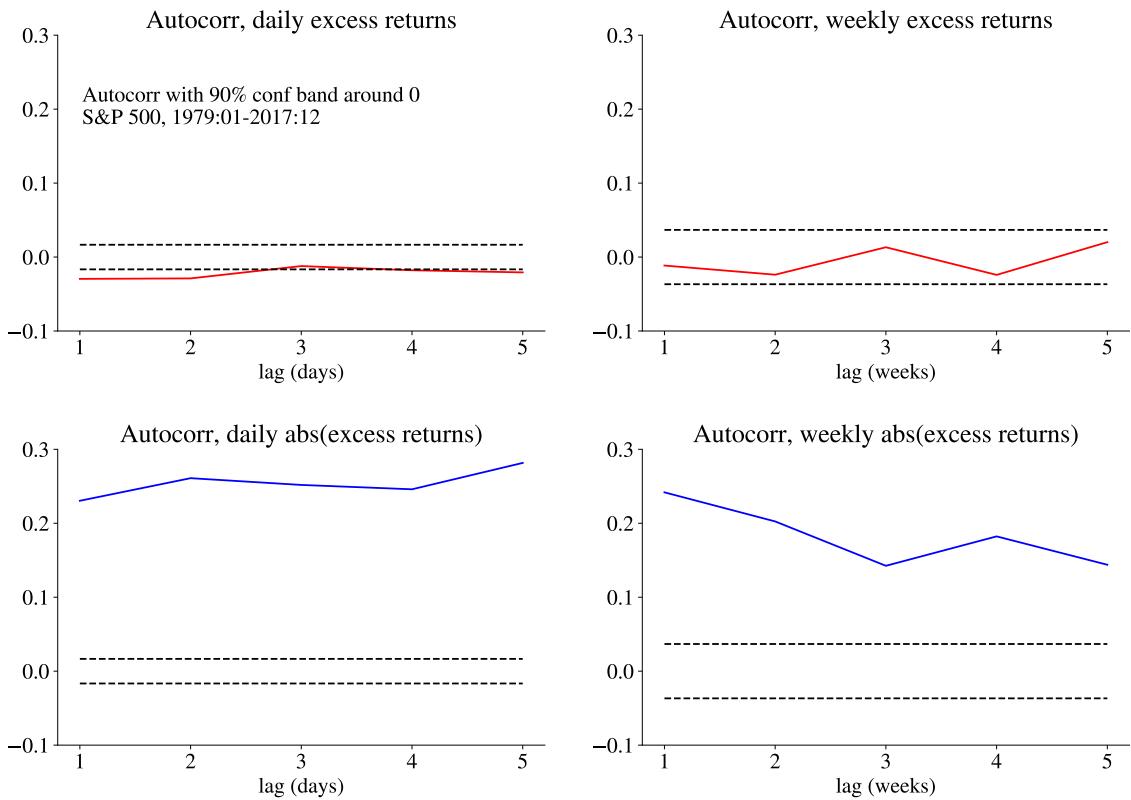


Figure 8.1: Predictability of US stock returns

The *Box-Pierce test* follows directly from the result in (8.3), since it shows that $\sqrt{T} \hat{\rho}_i$

and $\sqrt{T} \hat{\rho}_j$ are independent $N(0,1)$ variables. Therefore, the sum of the square of them is distributed as a χ^2 variable. The test statistic is

$$Q_L = T \sum_{s=1}^L \hat{\rho}_s^2 \xrightarrow{d} \chi_L^2. \quad (8.5)$$

However, you could also test $T(\rho_2^2 + \rho_5^2)$, and it would also have a χ_2^2 distribution.

Example 8.2 (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_1^2 distribution is 2.71, so the null hypothesis of no autocorrelation is rejected.

The choice of lag order in (8.5), 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 statistics is not affected much by increasing L , but the critical values increase).

Partial autocorrelations are discussed in Section 8.5.1.

Remark 8.3 (Runs test*) A “runs test” is a non-parametric test of randomness. Let x_t be an indicator variable

$$x_t = \begin{cases} 0 & \text{if } y_t \leq q \\ 1 & \text{if } y_t > q \end{cases}$$

where q typically (but not necessarily) is the mean of y_t . Let $T_1 = \sum_{t=1}^T x_t$, that is the number of occasions when $y_t > q$, and $T_2 = T - T_1$ (the number of occasions when $y_t \leq q$). Also define the numbers of runs, that is, the number of changes in the x_t series (where the first observation is counted as a change). That is define

$$r = 1 + \sum_{t=2}^T |x_t - x_{t-1}|.$$

It is straightforward (but tedious) to show that, under the null hypothesis of randomness,

$$\begin{aligned} \mathbb{E} r &= 2 \frac{T_1 T_2}{T} + 1 \text{ and} \\ \text{Var}(r) &= \frac{(\mathbb{E} r - 1)(\mathbb{E} r - 2)}{T - 1}. \end{aligned}$$

We can therefore test the null hypothesis of randomness by a t-stat

$$\frac{r - \mathbb{E} r}{\sqrt{\text{Var}(r)}} \xrightarrow{d} N(0, 1).$$

The basic intuition of the test is that a positive autocorrelation would lead to too few runs ($r < E r$): the y_t variable would stay on one side of the threshold q for long spells of time—and hence there would be few changes in x_t . Negative autocorrelation is just the opposite, since it tends to give a zigzag pattern around the mean. See Figure 8.2 for an example.

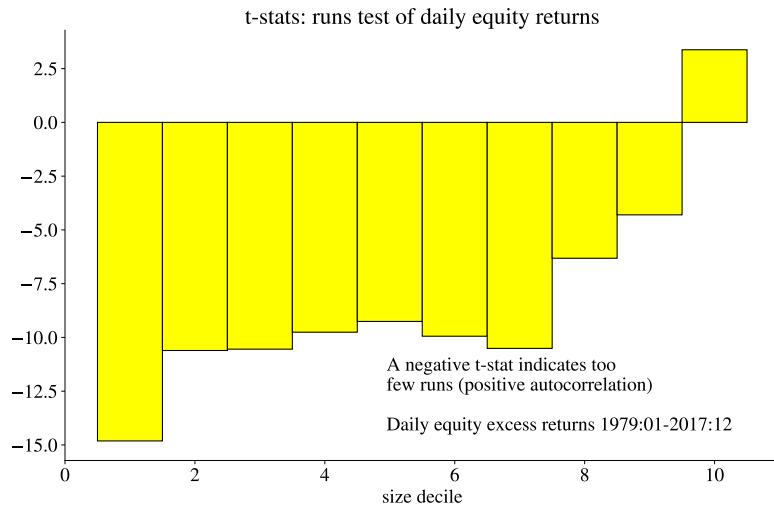


Figure 8.2: Runs test

8.2 Stationarity

The process y_t is (weakly) stationary if the mean, variance, and covariances are finite and constant across time

$$E y_t = \mu < \infty \quad (8.6)$$

$$\text{Var}(y_t) = \gamma_0 < \infty \quad (8.7)$$

$$\text{Cov}(y_t, y_{t-s}) = \gamma_s < \infty \quad (8.8)$$

The *autocorrelation function* is just the autocorrelation coefficient ρ_s as a function of s . Notice that

$$\lim_{|s| \rightarrow \infty} \rho_s = 0 \text{ for any stationary series.} \quad (8.9)$$

See Figure 8.3 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 . For any stationary series, the IRF converges to zero as the horizon (s) is increased. See Figure 8.4 for an example.

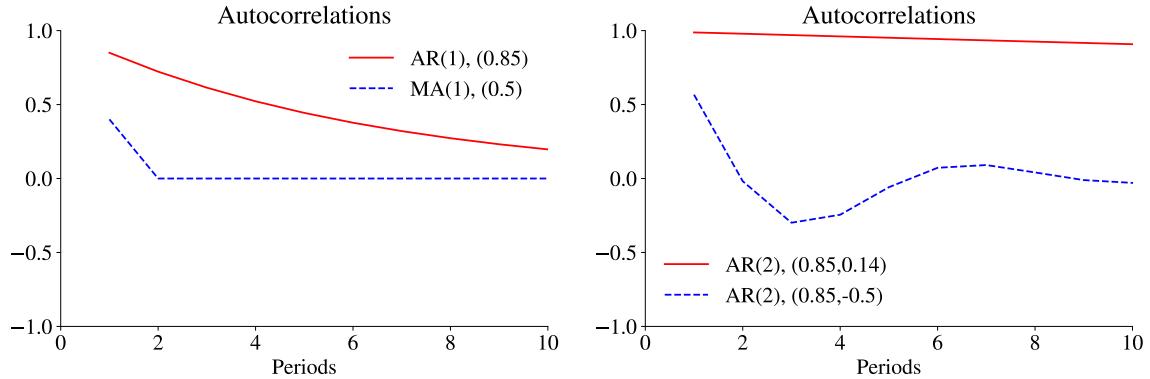


Figure 8.3: Example of autocorrelation functions

8.3 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

$$\begin{aligned} E \varepsilon_t &= 0 \\ \text{Var}(\varepsilon_t) &= \sigma^2, \text{ and} \\ \text{Cov}(\varepsilon_{t-s}, \varepsilon_t) &= 0 \text{ if } s \neq 0. \end{aligned} \tag{8.10}$$

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 variable that has a non-zero mean, we can form

$$y_t = \mu + \varepsilon_t, \tag{8.11}$$

where μ is a constant. This process is most easily estimated by the sample mean and variance or by OLS with a constant as the only regressor.

The *impulse response function* of a white noise process is just a blip.

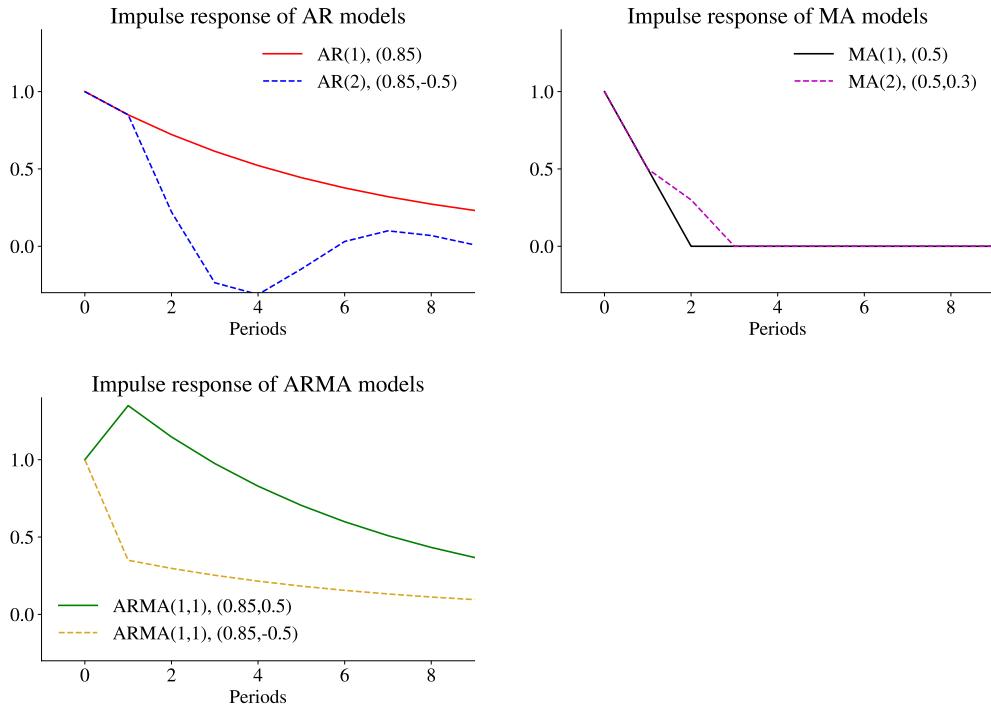


Figure 8.4: Impulse responses

8.4 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 autoregressive processes. The process is assumed to have a zero mean (or is demeaned: the original variable minus its mean, $y_t = x_t - \bar{x}$), but it is straightforward to put in any mean or trend.

An AR(1) is

$$y_t = ay_{t-1} + \varepsilon_t, \text{ with } \text{Var}(\varepsilon_t) = \sigma^2, \quad (8.12)$$

where ε_t is the white noise process in (8.10) which is uncorrelated with y_{t-1} . If $|a| < 1$, then the effect of a shock eventually dies out: y_t is stationary.

The *impulse response function* of an AR(1) with $a > 0$ shows an exponentially decaying response and with $a < 0$ it is zigzag response that decreases in amplitude. See Figure 8.4 for an illustration.

The AR(1) model can be *estimated with OLS* (since ε_t and y_{t-1} are uncorrelated) and the usual tools for testing significance of coefficients and estimating the variance of the residual all apply.

The basic properties of an AR(1) process are (provided $|a| < 1$)

$$\text{Var}(y_t) = \sigma^2 / (1 - a^2) \quad (8.13)$$

$$\text{Corr}(y_t, y_{t-s}) = a^s, \quad (8.14)$$

so the variance and autocorrelation are increasing in a (assuming $a > 0$). When $|a| \geq 1$, then the process is non-stationary, which means that the impulse response does not converge to zero as the horizon increases (see below for a discussion of such models: they are tricky).

See Figure 8.5 for an illustration.

Remark 8.4 (*Autocorrelation and autoregression*). Notice that the OLS estimate of a in (8.12) is essentially the same as the sample autocorrelation coefficient in (8.2). This follows from the fact that the slope coefficient is $\text{Cov}(y_t, y_{t-1}) / \text{Var}(y_{t-1})$. The denominator can be a bit different since a few data points are left out in the OLS estimation, but the difference is likely to be small.

Example 8.5 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 $a > 0$.

If $a = 1$ in (8.12), then we get a *random walk*. It is clear from the previous analysis that a random walk is non-stationary—that is, the effect of a shock never dies out. This implies that the variance is infinite and that the standard tools for testing coefficients are invalid. The solution is to study changes in y instead: $y_t - y_{t-1}$. In general, processes with the property that the effect of a shock never dies out are called non-stationary or unit root or integrated processes. Try to avoid them.

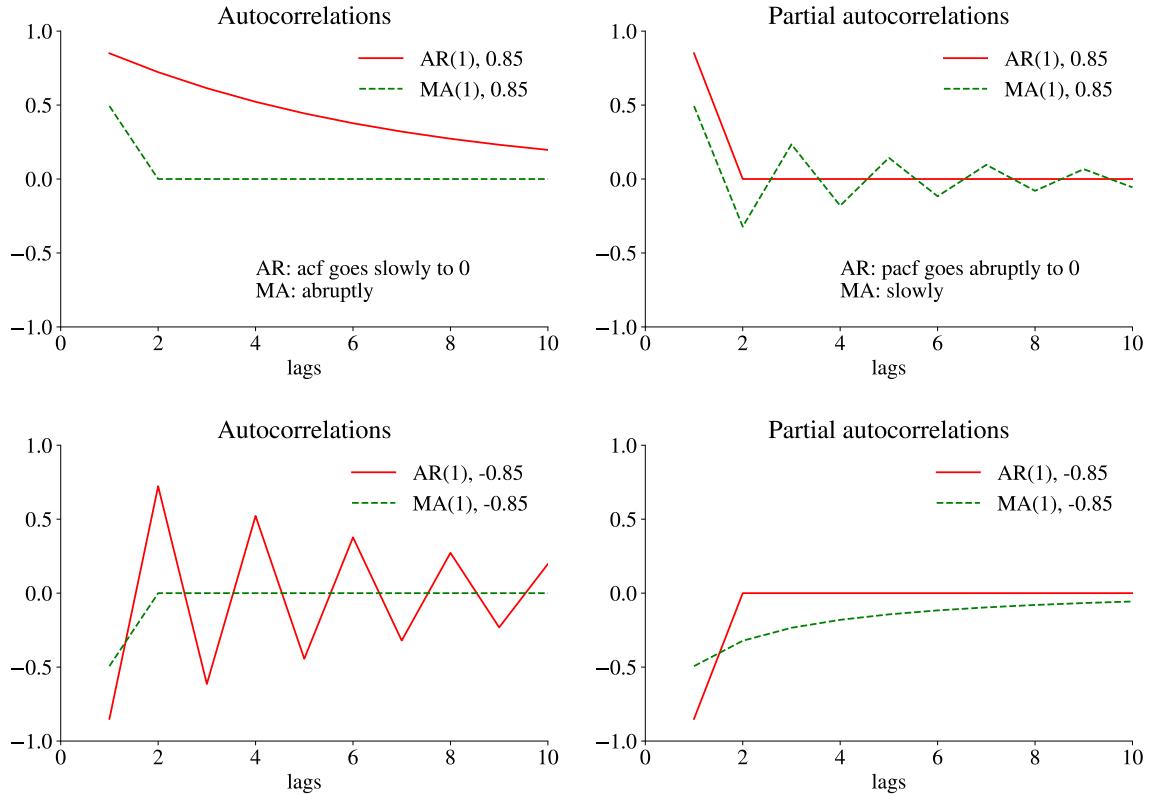


Figure 8.5: Autocorrelations and partial autocorrelations

8.4.1 More on the Properties of an AR(1) Process*

Solve (8.12) backwards by repeated substitution

$$y_t = a\underbrace{(ay_{t-2} + \varepsilon_{t-1})}_{y_{t-1}} + \varepsilon_t \quad (8.15)$$

$$= a^2 y_{t-2} + a\varepsilon_{t-1} + \varepsilon_t \quad (8.16)$$

$$\vdots \quad (8.17)$$

$$= a^{K+1} y_{t-K-1} + \sum_{s=0}^K a^s \varepsilon_{t-s}. \quad (8.18)$$

The factor $a^{K+1} y_{t-K-1}$ declines monotonically to zero if $0 < a < 1$ as K increases, and declines in an oscillating fashion if $-1 < a < 0$. In either case, the AR(1) process is

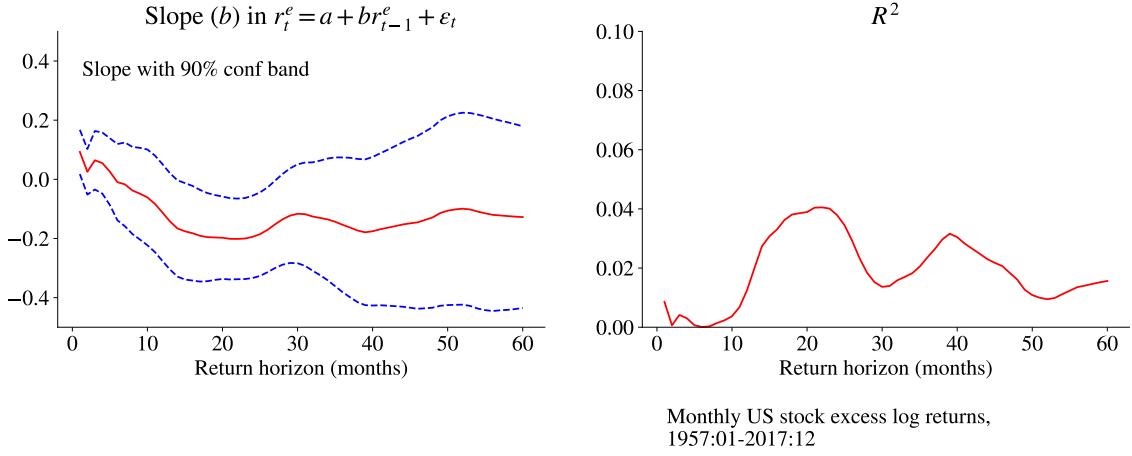


Figure 8.6: Predicting US stock returns (various investment horizons) with lagged returns.

covariance *stationary* and we can then take the limit as $K \rightarrow \infty$ to get

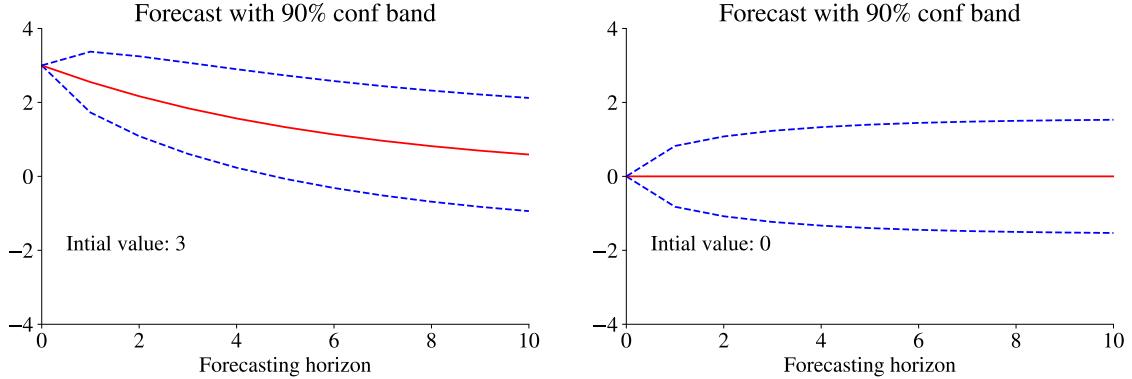
$$\begin{aligned} y_t &= \varepsilon_t + a\varepsilon_{t-1} + a^2\varepsilon_{t-2} + \dots \\ &= \sum_{s=0}^{\infty} a^s \varepsilon_{t-s}. \end{aligned} \tag{8.19}$$

Since ε_t is uncorrelated over time, y_{t-1} and ε_t are uncorrelated. We can therefore calculate the variance of y_t in (8.12) as the sum of the variances of the two components on the right hand side

$$\begin{aligned} \text{Var}(y_t) &= \text{Var}(ay_{t-1}) + \text{Var}(\varepsilon_t) \\ &= a^2 \text{Var}(y_{t-1}) + \text{Var}(\varepsilon_t) \\ &= \text{Var}(\varepsilon_t)/(1-a^2), \text{ since } \text{Var}(y_{t-1}) = \text{Var}(y_t). \end{aligned} \tag{8.20}$$

In this calculation, we use the fact that $\text{Var}(y_{t-1})$ and $\text{Var}(y_t)$ are equal. Formally, this follows from that they are both linear functions of current and past ε_s terms (see (8.19)), which have the same variance over time (ε_t is assumed to be white noise).

Note from (8.20) that the variance of y_t is increasing in the absolute value of a , which is illustrated in Figure 8.7. The intuition is that a large $|a|$ implies that a shock have effect over many time periods and thereby create movements (volatility) in y_t .



AR(1) model: $y_{t+1} = 0.85y_t + \varepsilon_{t+1}$, with $\sigma = 0.5$

Figure 8.7: Properties of AR(1) process

Similarly, the covariance of y_t and y_{t-1} is

$$\begin{aligned}\text{Cov}(y_t, y_{t-1}) &= \text{Cov}(ay_{t-1} + \varepsilon_t, y_{t-1}) \\ &= a \text{Cov}(y_{t-1}, y_{t-1}) \\ &= a \text{Var}(y_t).\end{aligned}\tag{8.21}$$

We can then calculate the first-order autocorrelation as

$$\begin{aligned}\text{Corr}(y_t, y_{t-1}) &= \frac{\text{Cov}(y_t, y_{t-1})}{\text{Std}(y_t) \text{Std}(y_{t-1})} \\ &= a.\end{aligned}\tag{8.22}$$

It is straightforward to show that

$$\text{Corr}(y_t, y_{t-s}) = \text{Corr}(y_{t+s}, y_t) = a^s.\tag{8.23}$$

8.4.2 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 (8.12) we get

$$y_{t+1} = ay_t + \varepsilon_{t+1}.\tag{8.24}$$

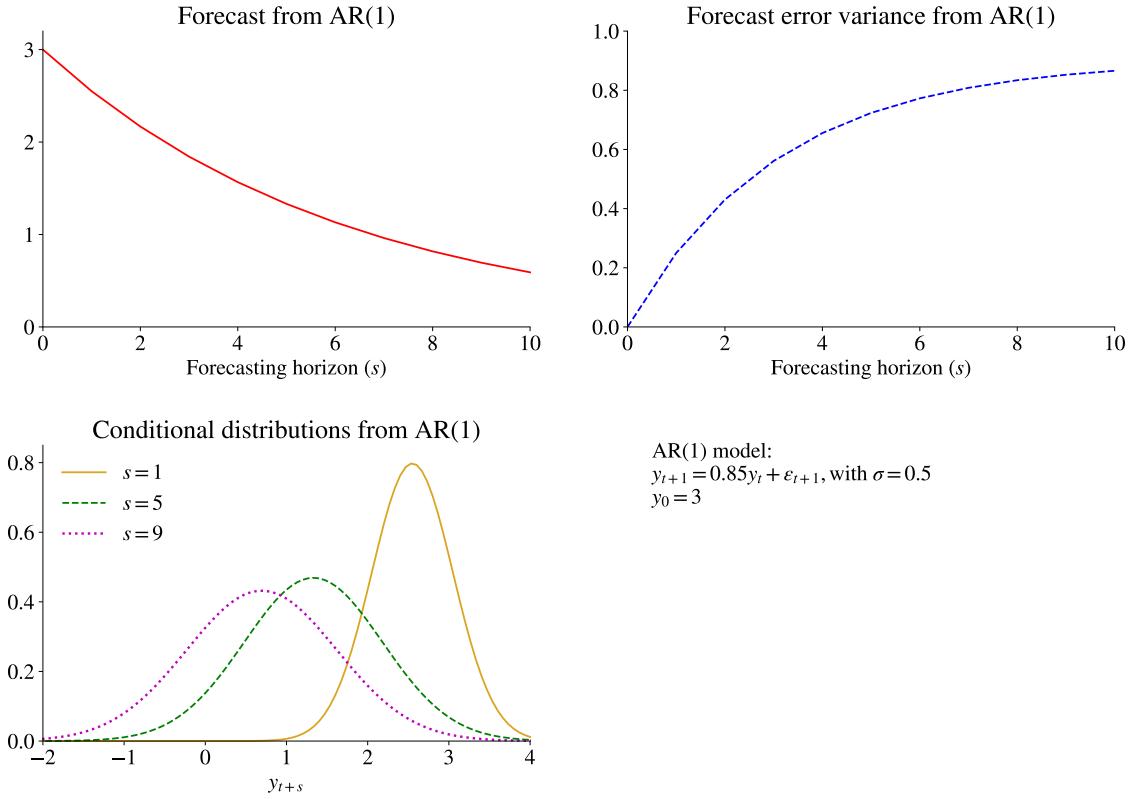


Figure 8.8: Properties of forecasts from AR(1) process

Since the best guess of ε_{t+1} is that it is zero, the best forecast and the associated forecast error are

$$E_t y_{t+1} = ay_t, \text{ and} \quad (8.25)$$

$$y_{t+1} - E_t y_{t+1} = \varepsilon_{t+1} \text{ with variance } \sigma^2. \quad (8.26)$$

We may also want to forecast y_{t+2} using the information in t . To do that note that (8.12) gives

$$\begin{aligned} y_{t+2} &= ay_{t+1} + \varepsilon_{t+2} \\ &= a(\underbrace{ay_t + \varepsilon_{t+1}}_{y_{t+1}}) + \varepsilon_{t+2} \\ &= a^2 y_t + a\varepsilon_{t+1} + \varepsilon_{t+2}. \end{aligned} \quad (8.27)$$

Since the $E_t \varepsilon_{t+1}$ and $E_t \varepsilon_{t+2}$ are both zero and $\text{Cov}(\varepsilon_{t+1}, \varepsilon_{t+2}) = 0$ we get that

$$E_t y_{t+2} = a^2 y_t, \text{ and} \quad (8.28)$$

$$y_{t+2} - E_t y_{t+2} = a\varepsilon_{t+1} + \varepsilon_{t+2} \text{ with variance } a^2\sigma^2 + \sigma^2. \quad (8.29)$$

More generally, we have

$$E_t y_{t+s} = a^s y_t, \quad (8.30)$$

$$\text{Var}(y_{t+s} - E_t y_{t+s}) = (1 + a^2 + a^4 + \dots + a^{2(s-1)})\sigma^2 \quad (8.31)$$

$$= \frac{a^{2s} - 1}{a^2 - 1}\sigma^2. \quad (8.32)$$

Notice that the point forecast converges towards zero and the variance of the forecast error variance to the unconditional variance (see Example 8.5).

Example 8.6 If $y_t = 3$, $a = 0.85$ and $\sigma = 0.5$, then the forecasts and the forecast error variances become

Horizon s	$E_t y_{t+s}$	$\text{Var}(y_{t+s} - E_t y_{t+s})$
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	$0.85^{25} \times 3 = 0.05$	$\frac{0.85^{50}-1}{0.85^2-1} \times 0.5^2 = 0.90$

To calculate the forecasts, you could alternatively apply a simple *recursive approach* like

$$E_t y_{t+1} = a y_t \quad (8.33)$$

$$E_t y_{t+2} = a E_t y_{t+1} \quad (8.34)$$

and so forth. We will later use the same principle for more complicated AR models.

If the shocks ε_t are normally distributed, then we can calculate 90% confidence intervals around the point forecasts in (8.25) and (8.28) as

$$90\% \text{ confidence band of } E_t y_{t+1} : a y_t \pm 1.64 \times \sigma \quad (8.35)$$

$$90\% \text{ confidence band of } E_t y_{t+2} : a^2 y_t \pm 1.64 \times \sqrt{a^2\sigma^2 + \sigma^2}. \quad (8.36)$$

(Recall that 90% of the probability mass is within the interval -1.64 to 1.64 in the $N(0,1)$ distribution). To get 95% confidence bands, replace 1.64 by 1.96. See Figures 8.7–8.8 for illustrations.

Example 8.7 Continuing Example 8.6, we get the following 90% confidence bands

<u>Horizon s</u>	
1	$2.55 \pm 1.64 \times \sqrt{0.25} \approx [1.7, 3.4]$
2	$2.17 \pm 1.64 \times \sqrt{0.43} \approx [1.1, 3.2]$
25	$0.05 \pm 1.64 \times \sqrt{0.90} \approx [-1.5, 1.6]$

Remark 8.8 (White noise as special case of AR(1).) When $a = 0$ in (8.12), then the AR(1) collapses to a white noise process. The forecast is then a constant (zero) for all forecasting horizons, see (8.30), and the forecast error variance is also the same for all horizons, see (8.32).

8.4.3 Adding a Constant to the AR(1)

The discussion of the AR(1) worked with a zero mean variable, but that was just for convenience (to make the equations shorter). One way to work with a variable x_t with a non-zero mean, is to first estimate its sample mean \bar{x} and then let the y_t in the AR(1) model (8.12) be a demeaned variable $y_t = x_t - \bar{x}$.

To include a mean, μ , in the theoretical expressions, we just need to substitute $x_t - \mu$ for y_t everywhere. For instance, in (8.12) we would get

$$\begin{aligned} x_t - \mu &= a(x_{t-1} - \mu) + \varepsilon_t \text{ or} \\ x_t &= (1 - a)\mu + ax_{t-1} + \varepsilon_t. \end{aligned} \tag{8.37}$$

Estimation by LS will therefore give an intercept that equals $(1 - a)\mu$ and a slope coefficient that equals a .

8.5 AR(p)

The p th-order autoregressive process, AR(p), is a straightforward extension of the AR(1)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \varepsilon_t. \tag{8.38}$$

The AR(p) can capture richer dynamics than the AR(1). See Figure 8.4 for an illustration of impulse response functions.

This process can also be estimated with OLS since ε_t is uncorrelated with lags of y_t . Adding a constant to the theoretical expressions is straightforward: substitute $x_t - \mu$ for y_t everywhere.

Remark 8.9 (*Stationarity of an AR(p) model**) To investigate if the AR(p) model is stationary (the impulse responses converge to zero), rewrite it on vector form (also called “companion form”) as

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Calculate the eigenvalues (λ_i for $i = 1, \dots, p$) of the (companion) matrix on the right hand side. If $|\lambda_i| < 1$ for all i , then the model is stationary. Basically, this means that the model is “stable” in the sense that the effect of a shock eventually dies out.

8.5.1 Partial Autocorrelations

The p th partial autocorrelation tries to measure 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,

$$\text{partial autocorrelation}(s) = \text{Corr}(y_t - \hat{y}_t, y_{t-s} - \hat{y}_{t-s}), \quad (8.39)$$

where \hat{y}_t and \hat{y}_{t-s} are the best (linear) estimates of y_t and y_{t-s} based on $(y_{t-1}, \dots, y_{t-s+1})$. Yes, the “regressors” are the same in both regressions.

Example 8.10 (*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 autocorrelation coefficient is then $\text{Corr}(\hat{\varepsilon}_t, \hat{u}_{t-3})$.

For instance, if y_t is generated by an AR(1) model, then 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 estimated by a in an AR(1)

$$y_t = \underline{a} y_{t-1} + \varepsilon_t. \quad (8.40)$$

This gives the same result as following the definition in (8.39). The second partial autocorrelation is estimated by the second slope coefficient (a_2) in an AR(2)

$$y_t = a_1 y_{t-1} + \underline{a_2} y_{t-2} + \varepsilon_t, \quad (8.41)$$

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 forth. See Figure 8.5 for an illustration.

To choose a model, study the ACF and PACF—and check that residual are close to white noise (or at least not strongly autocorrelated). To avoid overfitting, try “punishing” models with too many parameters. Akaike’s Information Criterion (AIC) and the Bayesian information criterion (BIC) are

$$AIC = \ln \hat{\sigma}^2 + 2 \frac{p+1}{T} \quad (8.42)$$

$$BIC = \ln \hat{\sigma}^2 + \frac{p+1}{T} \ln T, \quad (8.43)$$

where $\hat{\sigma}^2$ is the variance of the fitted residuals. Choose the model with the lowest AIC or BIC. (Notice, however, that AIC often exaggerates the lag length.) This provides a trade-off between fit (low $\hat{\sigma}^2$) and number of parameters ($p+1$). See Figure 8.9 for an empirical illustration.

8.5.2 Forecasting with an AR(p)

As an example, consider making a forecast of y_{t+1} based on the information in t by using an AR(2)

$$y_{t+1} = a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1}. \quad (8.44)$$

This immediately gives the one-period point forecast

$$\mathbb{E}_t y_{t+1} = a_1 y_t + a_2 y_{t-1}. \quad (8.45)$$

The two-period forecast is then calculated recursively as

$$\mathbb{E}_t y_{t+2} = a_1 \mathbb{E}_t y_{t+1} + a_2 y_t. \quad (8.46)$$

Figure 8.10 gives an empirical example.

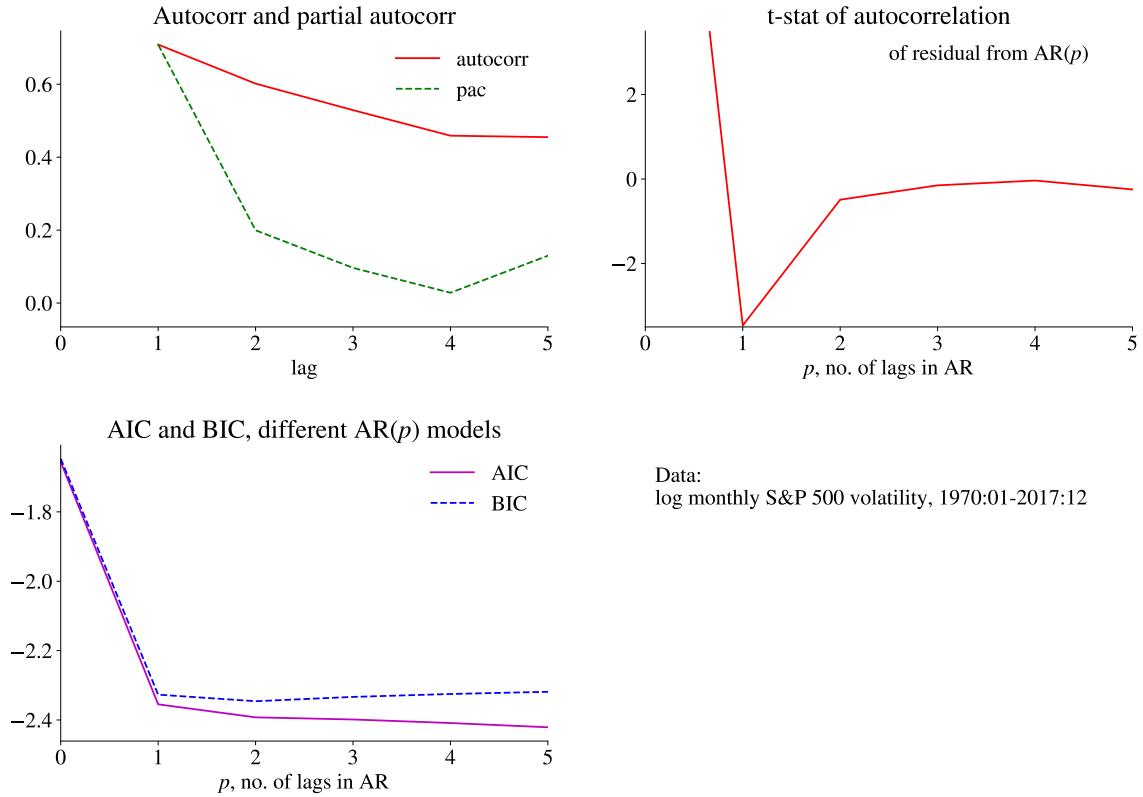


Figure 8.9: Example of choosing lag length in an AR model

Example 8.11 (Forecasts) With $(a_1, a_2) = (0.85, -0.5)$ and $(y_t, y_{t-1}) = (3, 1)$ we get

$$E_t y_{t+1} = 0.85 \times 3 - 0.5 \times 1 = 2.05,$$

$$E_t y_{t+2} = 0.85 \times 2.05 - 0.5 \times 3 = 0.24.$$

The calculations of the forecasting error variance is straightforward, but somewhat messy. However, both the one-period and two-period forecasts are linear functions of y_t and y_{t-1} . We could therefore estimate the following two equations with OLS

$$\begin{aligned} y_{t+1} &= a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1} \\ y_{t+2} &= b_1 y_t + b_2 y_{t-1} + v_{t+2}. \end{aligned}$$

This will give results that are very similar to (8.45)–(8.46). (Asymptotically, the results will be the same.) The variances of ε_{t+1} and v_{t+2} are also very similar to the theoretical results.

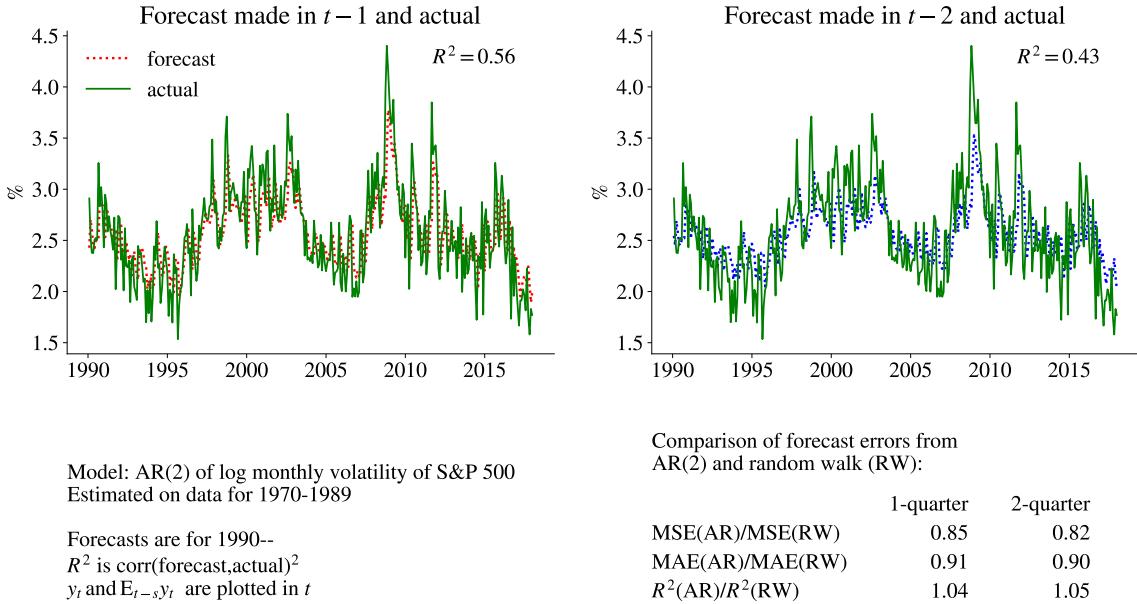


Figure 8.10: Forecasting with an AR(2)

8.5.3 Autoregressions versus Autocorrelations*

It is straightforward to see the relation between autocorrelations and the AR model. This relation is given by the *Yule-Walker equations*. For instance, for an AR(1), the autoregression coefficient is simply the first autocorrelation coefficient. For higher-order process, the transformation is non-linear, but testing if all the autocorrelations are zero is essentially the same as testing if all the autoregressive coefficients are zero.

To illustrate the Yule-Walker equations, consider an AR(2)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t. \quad (8.47)$$

Clearly, the covariance of y_t and y_{t-s} must be the same as the covariance between the right hand side of (8.47) and y_{t-s}

$$\begin{aligned} \text{Cov}(y_t, y_{t-s}) &= \text{Cov}(a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t, y_{t-s}) \\ &= a_1 \text{Cov}(y_{t-1}, y_{t-s}) + a_2 \text{Cov}(y_{t-2}, y_{t-s}) + \text{Cov}(\varepsilon_t, y_{t-s}). \end{aligned}$$

Applying this on $s = 1, 2$ (and replacing $\text{Cov}(y_t, y_{t-s})$ and $\text{Cov}(y_{t-s}, y_t)$ by γ_s etc) gives

$$\begin{aligned}\gamma_1 &= a_1\gamma_0 + a_2\gamma_1 \\ \gamma_2 &= a_1\gamma_1 + a_2\gamma_0,\end{aligned}\tag{8.48}$$

provided $\text{Cov}(\varepsilon_t, y_{t-1}) = 0$ and $\text{Cov}(\varepsilon_t, y_{t-2}) = 0$. This holds in (at least) two settings: when y_t is truly an AR(2) process *or* when we have estimated (8.47) by OLS (since it creates fitted residuals that uncorrelated with the regressors).

With information on the autocovariances ($\gamma_0, \gamma_1, \gamma_2$) we can solve these equations for the autoregression parameters (a_1, a_2). To see how, rewrite (8.48) as

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix},\tag{8.49}$$

which we can solve for (a_1, a_2) as

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}.\tag{8.50}$$

Notice that dividing both sides of (8.49) by γ_0 would not change the solution of (a_1, a_2) . This means that γ_i could effectively be either autocovariances or autocorrelations (in the latter case, $\gamma_0 = 1$).

With p autocovariances (instead of 2) we get

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix}.\tag{8.51}$$

This also provides a mechanism for calculating the partial autocorrelations (a_p for $p = 1, 2, 3, \dots$) from the autocovariances. First, set $p = 1$ in (8.51) to find the AR(1) coefficient (which will be $a_1 = \gamma_1/\gamma_0$). Second, set $p = 2$ and solve to find a_2 . Third, set $p = 3$ and solve to find a_3 and so forth.

8.6 Moving Average (MA)

A q^{th} -order moving average process MA(q) is

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad (8.52)$$

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) can show many different patterns, but they are all 0 at horizons beyond q . See Figure 8.4 for an illustration.

Estimation of an MA processes is typically done by *maximum likelihood*. LS does not work at all 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 8.12 (*MLE of an MA(1)*) The MA(1) is $y_t = \varepsilon_t + \theta \varepsilon_{t-1}$. Assume that $\varepsilon_0 = 0$, so you can calculate $\varepsilon_1 = y_1$ and $\varepsilon_t = y_t - \theta \varepsilon_{t-1}$ for $t \geq 2$. If $\varepsilon_t \sim N(0, \sigma^2)$, then the log-likelihood function for T observations is $-\ln(2\pi)T/2 - \ln(\sigma^2)T/2 - \sum_{t=1}^T \varepsilon_t^2/(2\sigma^2)$. This is maximized with respect to θ and σ^2 . Clearly, the ε_t series must be recalculated in each iteration (as θ changes).

Remark 8.13 (*MLE of an MA(2)**) The MA(2) is $y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$. Assume that $(\varepsilon_{-1}, \varepsilon_0) = (0, 0)$. Calculate $\varepsilon_1 = y_1$, $\varepsilon_2 = y_2 - \theta_1 \varepsilon_1$, and $\varepsilon_t = y_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$ for $t \geq 3$. It can be noticed that ε_t is an AR(2) where y_t is the “shock” and the autoregressive coefficients are $(-\theta_1, -\theta_2)$. Once we have a series of fitted residuals, use them in the likelihood function as for an MA(1). Consequently, this estimation approach only works if $(-\theta_1, -\theta_2)$ defines a stationary AR(2). Otherwise, more sophisticated methods are needed.

We can often rewrite (“invert”) an MA(q) as an AR(∞). For instance, if $|\theta| < 1$ then the MA(1) can be written

$$y_t = \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} + \dots + \varepsilon_t. \quad (8.53)$$

Proof. (*of (8.53)) Notice that $\varepsilon_t = y_t - \theta \varepsilon_{t-1}$. Substitute for ε_{t-1} by using $y_{t-1} - \theta \varepsilon_{t-2}$, to get $\varepsilon_t = y_t - \theta y_{t-1} + \theta^2 \varepsilon_{t-2}$. Keep substituting and notice that the $\theta^s \varepsilon_{t-s}$ goes to zero as s goes to infinity (provided $|\theta| < 1$). Rearrange as (8.53). ■

Equation (8.53) shows that when $\theta > 0$, then autoregression coefficients converge in a zigzag way to zero. In contrast, when $\theta < 0$, the convergence is monotonic. This pattern can help us to distinguish a true AR process from an MA process. 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 an 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 8.5 for an illustration.

Remark 8.14 (*MA form and the impulse response function**) First, notice that any stationary process can be rewritten on (“inverted to”) MA form. For instance, an AR(1) can be written as an MA(∞)

$$\begin{aligned} y_t &= ay_{t-1} + \varepsilon_t \\ &= a\underbrace{(ay_{t-2} + \varepsilon_{t-1})}_{y_{t-1}} + \varepsilon_t = a^2y_{t-2} + a\varepsilon_{t-1} + \varepsilon_t \\ &\vdots \\ &= \sum_{s=0}^{\infty} a^s \varepsilon_{t-s}, \end{aligned}$$

which is on MA form. Second, notice that the impulse response (the dynamic response to a shock in period t) is the same as the MA coefficients $(1, \theta_1, \theta_2, \dots)$. For instance, write out (8.52) 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 $(1, \theta_1, \theta_2, \dots)$.

8.7 ARMA(p,q)

When the autocorrelations and partial autocorrelations show mixed patterns, then a combination of AR and MA models might be appropriate.

Autoregressive-moving average (ARMA) models add a moving average structure to an AR model. For instance, an ARMA(2,1) is

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t + \theta \varepsilon_{t-1}, \quad (8.54)$$

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. ARMA models can generate complicated dynamics. See Figure 8.4 for impulse response functions.

ARMA are harder to estimate than the autoregressive model, and we typically use *MLE*. The appropriate specification of the model (number of lags of y_t and ε_t) is often unknown. The Box-Jenkins methodology is a set of guidelines for arriving at the correct specification by starting with some model, study the autocorrelation structure of the fitted residuals and then changing the model.

Remark 8.15 (*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 likelihood function $-\ln(2\pi)T/2 - \ln(\sigma^2)T/2 - \sum_{t=1}^T \varepsilon_t^2/(2\sigma^2)$ and maximize with respect to $(a_1, a_2, \theta, \sigma^2)$. As usual, the ε_t series must be recalculated for every new guess of the parameter vector.

Most ARMA models can be well approximated by an AR model—provided we add some extra lags. Since AR models are so simple to estimate, this approximation approach is often used.

8.8 VAR(p)

Let y_t be an $n \times 1$ vector of variables. The VAR(p) is

$$y_t = \mu + A_1 y_{t-1} + \dots + A_p y_{t-p} + \varepsilon_t, \quad (8.55)$$

where A_1, \dots, A_p are $n \times n$ matrices and ε_t is an $n \times 1$ vector of shocks. The vector autoregression is a multivariate version of an AR process: we can think of y_t and ε_t in (8.38) as vectors and the a_i as matrices.

To gauge the dynamics we can calculate the *impulse response function* of (all the n elements of) y_t to a shock to the 1st element of the $n \times 1$ vector ε_0 (set element 1 of ε_0 equal to one, but all other elements equal to zero). Then, we redo the same exercise, but now with respect to the 2nd element of ε_0 , and so on. In this way we get the response of each element of y_t with respect to each element of ε_0 . See Figure 8.11 for an illustration. It suggests that a low-order VAR model (here a VAR(1)) can create more complicated dynamics than a low-order AR model.

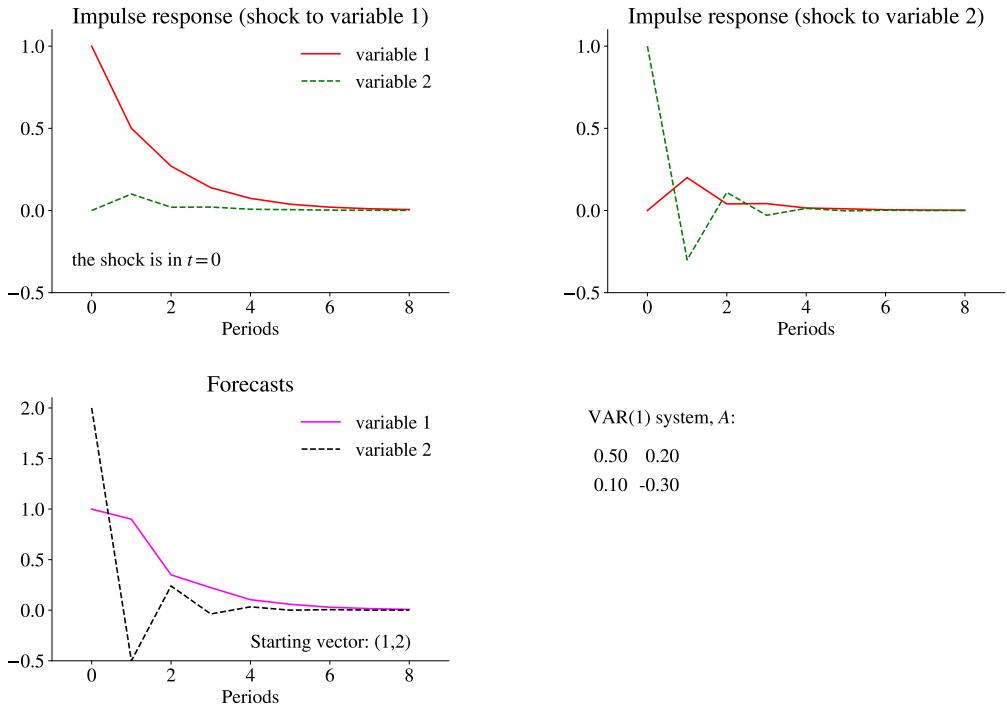


Figure 8.11: Properties of a VAR(1) model

For instance the VAR(1) of two variables (x_t and z_t) is (in matrix form)

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{bmatrix}, \quad (8.56)$$

or equivalently

$$x_t = A_{11}x_{t-1} + A_{12}z_{t-1} + \varepsilon_{xt}, \text{ and} \quad (8.57)$$

$$z_t = A_{21}x_{t-1} + A_{22}z_{t-1} + \varepsilon_{zt}. \quad (8.58)$$

Both (8.57) and (8.58) are regression equations, which can be *estimated with OLS* (since ε_{xt} and ε_{zt} are uncorrelated with x_{t-1} and z_{t-1}).

Example 8.16 (VAR(1) of 2 variables). Let the A matrix in (8.56) be

$$A = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix}.$$

Remark 8.17 (Stationarity*) The VAR(1) is stationary if the A matrix in (8.56) has $|\lambda_i| <$

1 where λ_i are the eigenvalues. For instance, for the 2-variable VAR(1) in Example 8.16 we get $(|\lambda_1|, |\lambda_2|) \approx (0.52, 0.32)$.

With the information available in t , that is, information about x_t and z_t , (8.57) and (8.58) can be used to forecast one- and two-step ahead as

$$\begin{bmatrix} E_t x_{t+1} \\ E_t z_{t+1} \end{bmatrix} = A \begin{bmatrix} x_t \\ z_t \end{bmatrix} \text{ and} \quad (8.59)$$

$$\begin{bmatrix} E_t x_{t+2} \\ E_t z_{t+2} \end{bmatrix} = A \begin{bmatrix} E_t x_{t+1} \\ E_t z_{t+1} \end{bmatrix} = AA \begin{bmatrix} x_t \\ z_t \end{bmatrix}, \quad (8.60)$$

where 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. Note that all we need to make the forecasts (for both $t + 1$ and $t + 2$) are the values in period t (x_t and z_t). This follows from that (8.56) is a first-order system where the values of x_t and z_t summarize all relevant information about the future that is available in t .

Example 8.18 (*Forecasts from a VAR(1)*) With the 2-variable VAR(1) in Example 8.16 and the initial values $(x_0, z_0) = (1, 2)$ we get

$$\begin{bmatrix} E_0 x_1 \\ E_0 z_1 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.9 \\ -0.5 \end{bmatrix} \text{ and}$$

$$\begin{bmatrix} E_0 x_2 \\ E_0 z_2 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix} \begin{bmatrix} 0.9 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0.35 \\ 0.24 \end{bmatrix}.$$

The forecast uncertainty about the one-period forecast is simple: the forecast error $x_{t+1} - E_t x_{t+1} = \varepsilon_{xt+1}$. The two-period forecast error, $x_{t+2} - E_t x_{t+2}$, is a linear combination of ε_{xt+1} , ε_{zt+1} , and ε_{xt+2} . The calculations of the forecasting error variance (as well as for the forecasts themselves) quickly get messy. This is even more true when the VAR system is of a higher order.

As for the AR(p) model, a practical way to get around the problem with complicated calculations is to estimate a separate model for each forecasting horizon. In a large sample, the difference between the two ways is trivial. For instance, suppose the correct model is the VAR(1) in (8.56) and that we want to forecast x one and two periods ahead.

From (8.59) and (8.60) we see that the regression equations should be of the form

$$x_{t+1} = \delta_1 x_t + \delta_2 z_t + u_{t+1}, \text{ and} \quad (8.61)$$

$$x_{t+2} = \gamma_1 x_t + \gamma_2 z_t + w_{t+s}. \quad (8.62)$$

With estimated coefficients (OLS can be used), it is straightforward to calculate forecasts and forecast error variances.

Example 8.19 (VAR(2) of 2×1 vector.) Let $y_t = (x_t, y_t)$. Then a VAR(2) is

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} A_{1,11} & A_{1,12} \\ A_{1,21} & A_{1,22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} A_{2,11} & A_{2,12} \\ A_{2,21} & A_{2,22} \end{bmatrix} \begin{bmatrix} x_{t-2} \\ z_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{bmatrix}.$$

Remark 8.20 (Stationarity of a VAR(p) model*) Construct the companion matrix as for the AR(p), but recall that a_i are matrices and replace the ones (1) with I_n where n indicates the number of variables in the VAR system. Then, the model is stationary if $|\lambda_i| < 1$ where λ_i are the eigenvalues of the companion matrix. For instance, for 2-variable VAR(2) in Remark 8.19 we get) Continuing on the previous example, we get

$$\begin{bmatrix} A_{1,11} & A_{1,11} & A_{2,11} & A_{2,12} \\ A_{1,21} & A_{1,22} & A_{2,21} & A_{2,22} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

8.8.1 Granger Causality

If z_t can help predict future x , over and above what lags of x itself can, then z is said to *Granger Cause* x . This is a statistical notion of causality, and may not necessarily have much to do with true causality (for instance, Christmas cards may Granger cause Christmas). In (8.61) z does Granger cause x if $\delta_2 \neq 0$, which is easily tested. More generally, there may be more lags of both x and z in the equation, so we need to test if all coefficients on different lags of z are zero.

8.9 Non-stationary Processes

8.9.1 Introduction

A *trend-stationary process* has a (deterministic) trend. The simplest example is

$$y_t = \mu + \beta t + \varepsilon_t \quad (8.63)$$

where ε_t is white noise. It can be made stationary by subtracting the linear trend

$$y_t - \beta t = \mu + \varepsilon_t. \quad (8.64)$$

We can typically apply all standard econometric methods on this de-trended data. This can be done in two steps: first estimate the trend and then use $y_t - \hat{\beta}t$ in the subsequent analysis. Alternatively, we can work also with the y_t series directly, provided we 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

$$y_t = \mu + y_{t-1} + \varepsilon_t, \quad (8.65)$$

where ε_t is white noise. The name “unit root process” comes from the fact that the largest eigenvalues 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 differences

$$y_t - y_{t-1} = \mu + \varepsilon_t. \quad (8.66)$$

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 8.21 (*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 and is therefore 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 $x_t = y_t - y_{t-1}$ 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

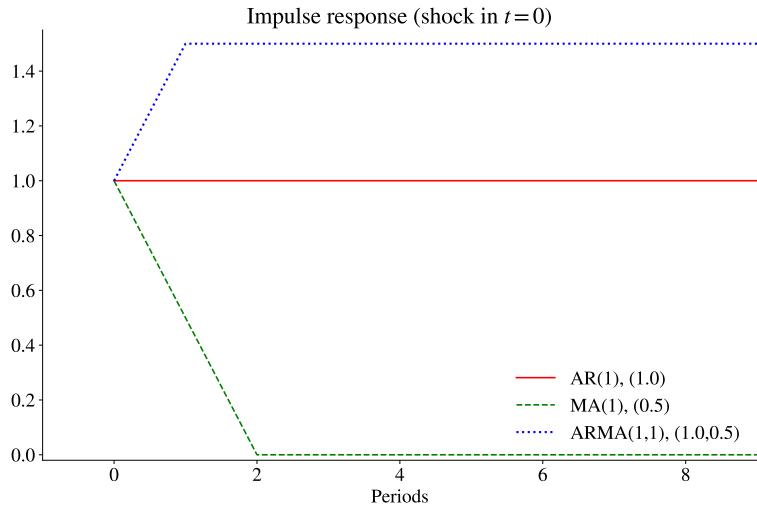


Figure 8.12: Impulse responses

easily seen for the random walk. Substitute repeatedly in (8.65) to get

$$\begin{aligned}
 y_t &= \mu + (\mu + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
 &\vdots \\
 &= t\mu + y_0 + \sum_{s=1}^t \varepsilon_s. \tag{8.67}
 \end{aligned}$$

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 . This process is clearly non-stationary. See Figure 8.12 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 random walk with drift, (8.67), the distribution conditional on the information in $t = 0$ is $N(y_0 + t\mu, s\sigma^2)$ if the innovations are normally distributed. This means that the expected change is $t\mu$ and that the conditional variance grows linearly with the forecasting horizon. The unconditional variance is therefore infinite and the standard results on inference are not applicable. See Figure 8.13.

A process could have two unit roots (integrated of order 2: I(2)). In this case, we need to difference twice to make it stationary. Alternatively, 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.

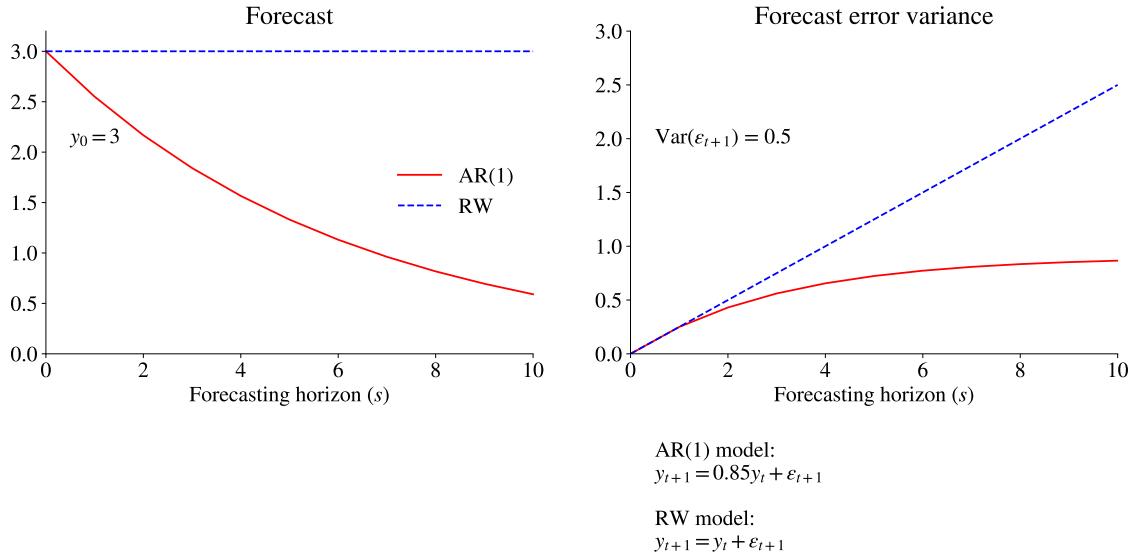


Figure 8.13: Properties of forecasts from random walk process

Example 8.22 (*Two unit roots**) Suppose y_t in Example (8.21) 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 (integrated of order 2: I(2) and needs to be differenced twice to become stationary).

Example 8.23 (*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 8.24 (*Lag operator**) A common and convenient way of dealing with leads and lags is the lag operator, L . It is such that

$$L^s y_t = y_{t-s}$$

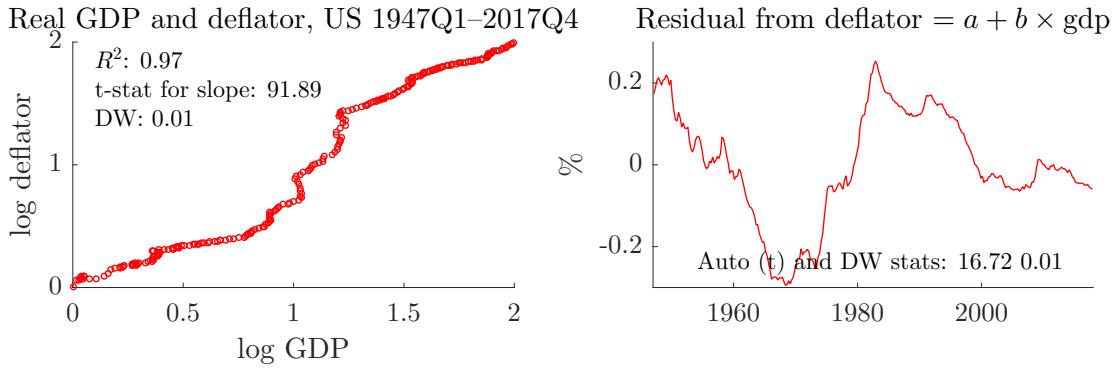


Figure 8.14: Example of a spurious regression

For instance, the AR(1) model

$$y_t = \theta \underbrace{y_{t-1}}_{Ly_t} + \varepsilon_t, \text{ or}$$

$$(1 - \theta L) y_t = \varepsilon_t, \text{ or}$$

$$\theta(L) y_t = \varepsilon_t,$$

where $\theta(L) = (1 - \theta L)$ is a lag polynomial. Similarly, an ARMA(2,1) can be written

$$y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} = \varepsilon_t + \alpha_1 \varepsilon_{t-1}$$

$$(1 - \theta_1 L - \theta_2 L^2) y_t = (1 + \alpha_1 L) \varepsilon_t.$$

8.9.2 Spurious Regressions

Strong trends often causes problems in econometric models where y_t is regressed on x_t . In essence, if no trend is included in the regression, then x_t will appear to be significant, just because it is a proxy for that trend. The same holds for unit root processes, even if they have no deterministic trends. The reason is that the innovations accumulate and the series therefore tend to be trending in small samples. A warning sign of a spurious regression is when $R^2 > DW$ statistics.

See Figure 8.14 for an empirical example and Figures 8.15–8.17 for a Monte Carlo simulation.

For trend-stationary data, this problem is easily solved by detrending with a linear trend (before estimating or just adding a trend to the regression).

However, this is usually a poor method for a unit root processes. What is needed is a

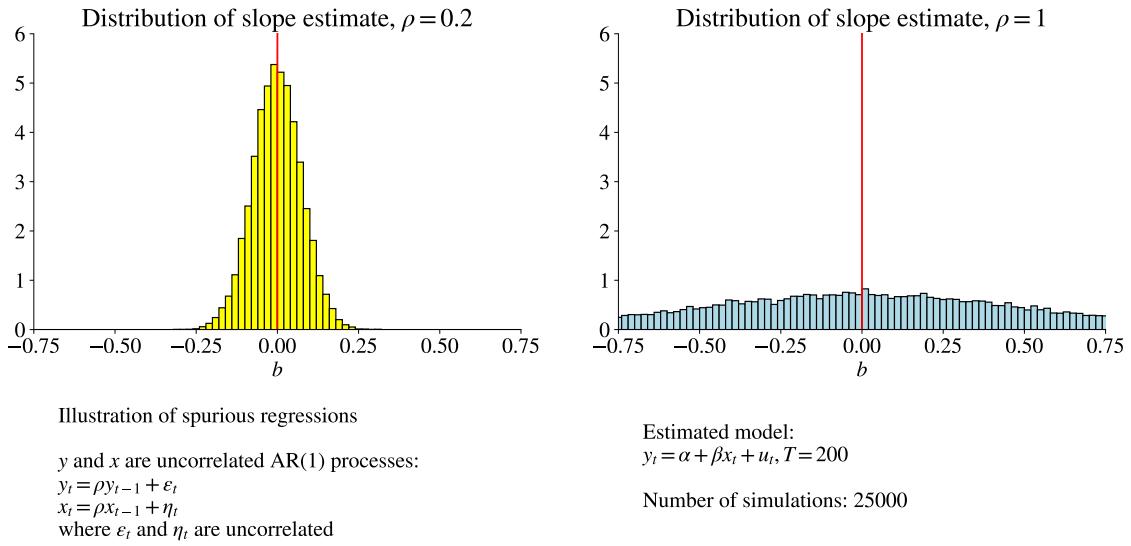


Figure 8.15: Distribution of slope coefficient when y_t and x_t are independent AR(1) processes

first difference. For instance, a first difference of the random walk with drift is

$$\begin{aligned}\Delta y_t &= y_t - y_{t-1} \\ &= \mu + \varepsilon_t,\end{aligned}\tag{8.68}$$

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 .

One may then be tempted to try first-differencing all non-stationary series, since it may be hard to tell if they are unit root process or just trend-stationary. For instance, a first difference of the trend stationary process, (8.63), gives

$$y_t - y_{t-1} = \beta + \varepsilon_t - \varepsilon_{t-1}.\tag{8.69}$$

Its unclear if this is an improvement: the trend is gone, but the errors are now of MA(1) type, which means that they are autocorrelated.

8.9.3 Testing for a Unit Root

Suppose we run an OLS regression of

$$y_t = a y_{t-1} + \varepsilon_t,\tag{8.70}$$

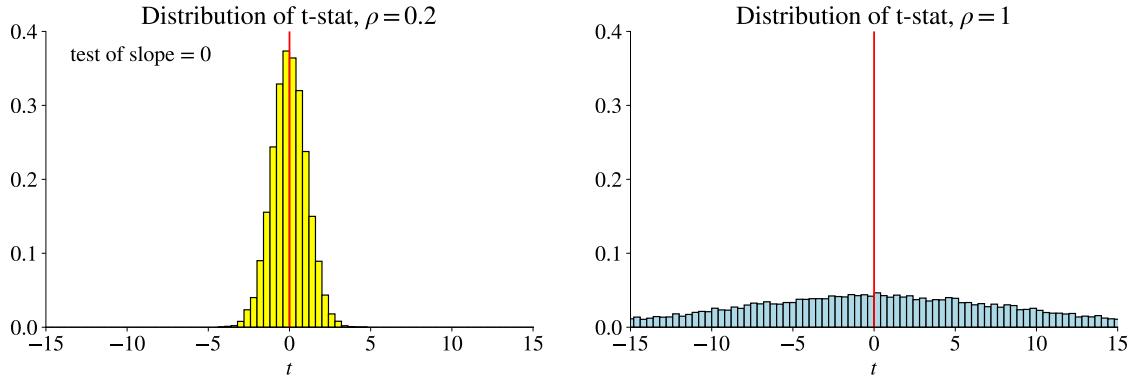


Figure 8.16: Distribution of t-statistics when y_t and x_t are independent AR(1) processes. See Figure 8.15.

where the true value of $|a| < 1$. The asymptotic distribution of the LS estimator is

$$\sqrt{T}(\hat{a} - a) \sim N(0, 1 - a^2). \quad (8.71)$$

(The variance follows from the standard OLS formula where the variance of the estimator is $\sigma^2(X'X/T)^{-1}$. Here $\text{plim } X'X/T = \text{Var}(y_t)$ 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 (8.71). 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 (8.70).

In practice, the approach is to run the regression (8.70) with a constant (and perhaps even a time trend), calculate the test statistic

$$DF = \frac{\hat{a} - 1}{\text{Std}(\hat{a})}, \quad (8.72)$$

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 (8.70)), do an *aug-*

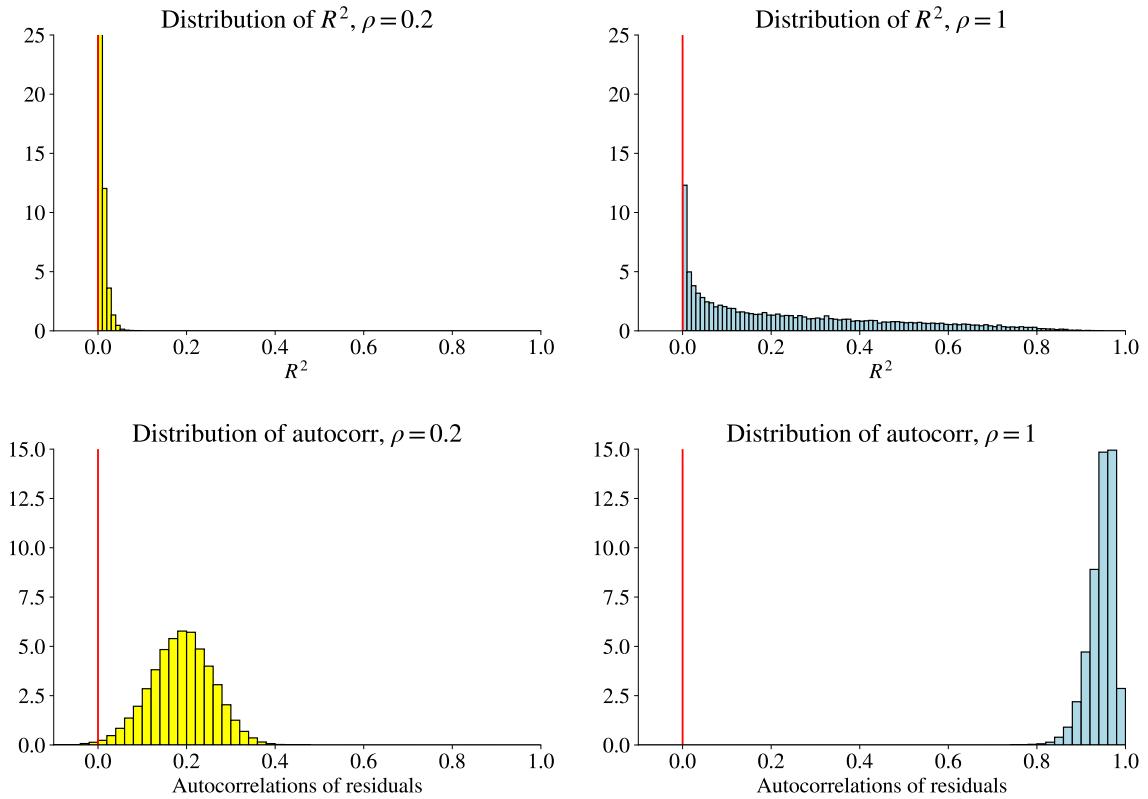


Figure 8.17: Distribution of R^2 and autocorrelation of residuals. See Figure 8.15.

mented Dickey-Fuller test (ADF)

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_{2t}, \text{ or}$$

$$\Delta y_t = \delta + (\theta_1 + \theta_2 - 1) y_{t-1} - \theta_2 \Delta y_{t-1} + \varepsilon_{2t}, \quad (8.73)$$

and test the coefficient on y_{t-1} in (8.73) (which should equal $\theta_1 + \theta_2 - 1$) against the alternative that it is less than zero. The critical values are as for the DF test. If ε_{2t} is autocorrelated, then add further lags.

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

$$y_t = a + \varepsilon_t. \quad (8.74)$$

Second, define

$$S_t = \sum_{s=1}^t \hat{\epsilon}_s \text{ for } t = 1, \dots, T \text{ and} \quad (8.75)$$

$$\hat{\sigma}^2 = \text{Var}(\hat{\epsilon}_t). \quad (8.76)$$

Third, the test statistic is

$$KPSS = \frac{1}{T^2} \sum_{t=1}^T S_t^2 / \hat{\sigma}^2 \quad (8.77)$$

Reject stationarity if $KPSS > 0.463$ (a 5% critical value). We could also include a linear trend in (8.74). The 5% critical value is then 0.146.

In principle, 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.

8.9.4 Cointegration*

An exception to the “spurious regression” result: Y_t and X_t are I(1) but share a common stochastic trend such that

$$y_t - \alpha - \beta x_t \text{ is I}(0). \quad (8.78)$$

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 are I(1) and therefore have high sample variance: OLS will pick $\hat{\beta}$ close to β .

In (8.78), we call $(1, -\beta)$ the *cointegrating vector*, since

$$\begin{bmatrix} 1 & -\beta \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} \text{ is I}(0) \quad (8.79)$$

Example 8.25 Y_t is GDP, x_t is private consumption. Suppose both are driven by the non-stationary productivity of the economy, A_t , plus other stationary stuff (z_t, w_t)

$$\begin{aligned} y_t &= \gamma A_t + z_t \\ x_t &= \delta A_t + w_t \end{aligned}$$

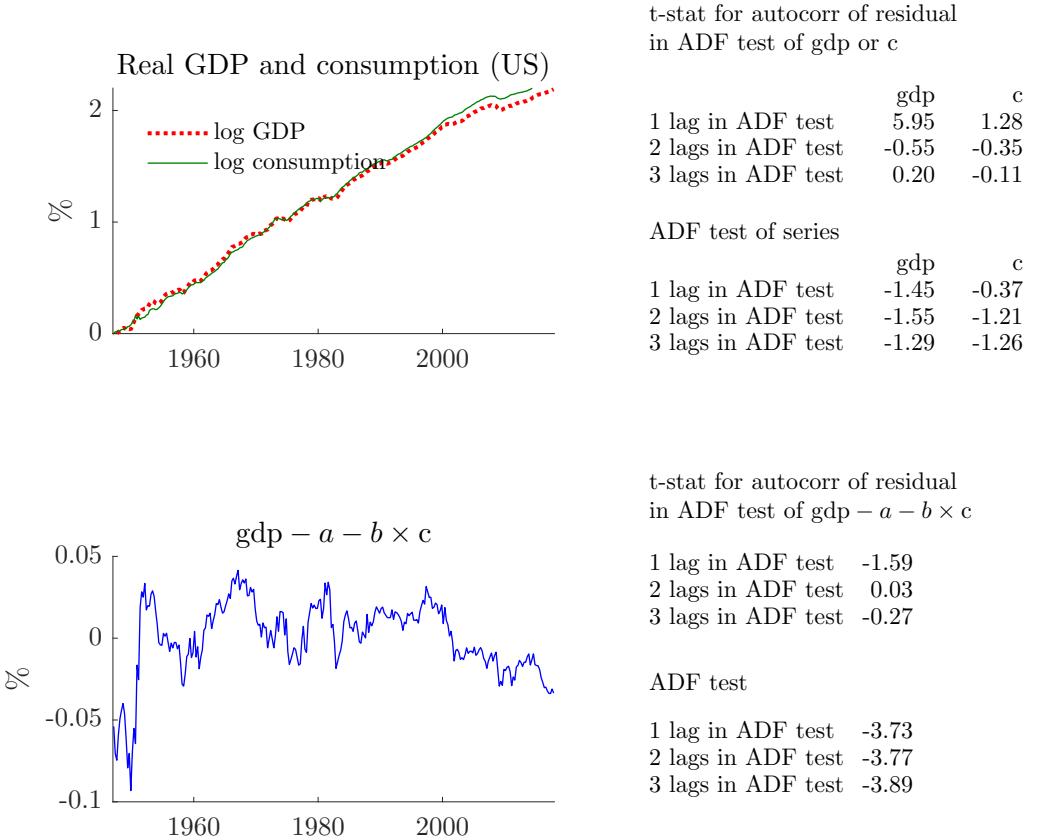


Figure 8.18: Unit root tests, US quarterly macro data

From the second equation $A_t = (x_t - w_t)/\delta$, use in first equation

$$\underbrace{y_t}_{I(1)} = \frac{\gamma}{\delta} \underbrace{x_t}_{I(1)} + z_t - \underbrace{\frac{\gamma}{\delta} w_t}_{I(0)}$$

To test if y_t and x_t are cointegrated, we need to study three things. First, does it make sense? Look at data, and consider the (economic) theory. Second, are both x_t and y_t I(I)? Do Dickey-Fuller tests, etc. Third, are (\hat{a}, \hat{b}) in from the regression

$$y_t = a + bx_t + \varepsilon_t \quad (8.80)$$

such that $\hat{\varepsilon}_t$ is I(0)? To determine the latter, do an ADF test on $\hat{\varepsilon}_t$, but use special critical values— H_0 : no cointegration (so ε_t is I(1)). 5% critical values: -3.34 (if x_t is a scalar).

One way to incorporate the cointegration in a model of the short-run dynamics is to

use a *Error-Correction Model*, for instance,

$$\begin{aligned}\Delta y_t &= \delta + \phi_1 \Delta x_{t-1} - \gamma (y_{t-1} - \beta x_{t-1}) + \varepsilon_t \text{ or perhaps} \\ &= \delta + \phi_1 \Delta x_{t-1} + \theta_1 \Delta y_{t-1} - \gamma (y_{t-1} - \beta x_{t-1}) + \varepsilon_t\end{aligned}\quad (8.81)$$

Recall: (y_t, x_t) are I(1), but $y_{t-1} - \beta x_{t-1}$ is I(0), so all terms in (8.81) are I(0). We typically do not put the intercept into the cointegrating relation (as there is already another intercept in the equation).

If $\gamma > 0$, then the system is driven back to a stationary path for $y - \beta x$: the “error correction mechanism.” Can have more lags of both Δy and Δx .

Estimation is fairly straightforward (Engle-Granger’s 2-step method). First, estimate the cointegrating vector. Second, use it in (8.81) and estimate the rest of the parameters. Standard tests can be applied to them.

	Δgdp
Coint res _{t-1}	-0.07 (-2.19)
Δgdp_{t-1}	0.14 (1.56)
Δc_{t-1}	0.33 (3.15)
Δgdp_{t-2}	0.02 (0.21)
Δc_{t-2}	0.21 (2.39)
const	0.00 (2.24)
R^2	0.23
obs	281.00

Table 8.1: Error-correction model for log real US GDP growth, 1947Q1–2017Q4. Numbers in parentheses are t-stats. The Coint res is the residual from regressing the log GDP level on the log consumption level.

Chapter 9

Predicting Asset Returns

Reference (medium): Elton, Gruber, Brown, and Goetzmann (2010) 17 (efficient markets) and 26 (earnings estimation)

Additional references: Campbell, Lo, and MacKinlay (1997) 2 and 7; Cochrane (2001) 20.1

More advanced material is denoted by a star (*). It is not required reading.

9.1 Autocorrelations and Autoregressions

9.1.1 Autocorrelation Coefficients

When the true autocorrelations are all zero (not ρ_0 , of course), then for any i and j different from zero

$$\sqrt{T} \begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \quad (9.1)$$

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). See Figure 9.2.

9.1.2 Autoregressions

An alternative way of testing autocorrelations is to estimate an AR model

$$y_t = c + a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \varepsilon_t, \quad (9.2)$$

and then test if all slope coefficients (a_1, a_2, \dots, a_p) are zero with a χ^2 or F test. This approach is somewhat less general than the Box-Pierce test, but most stationary time

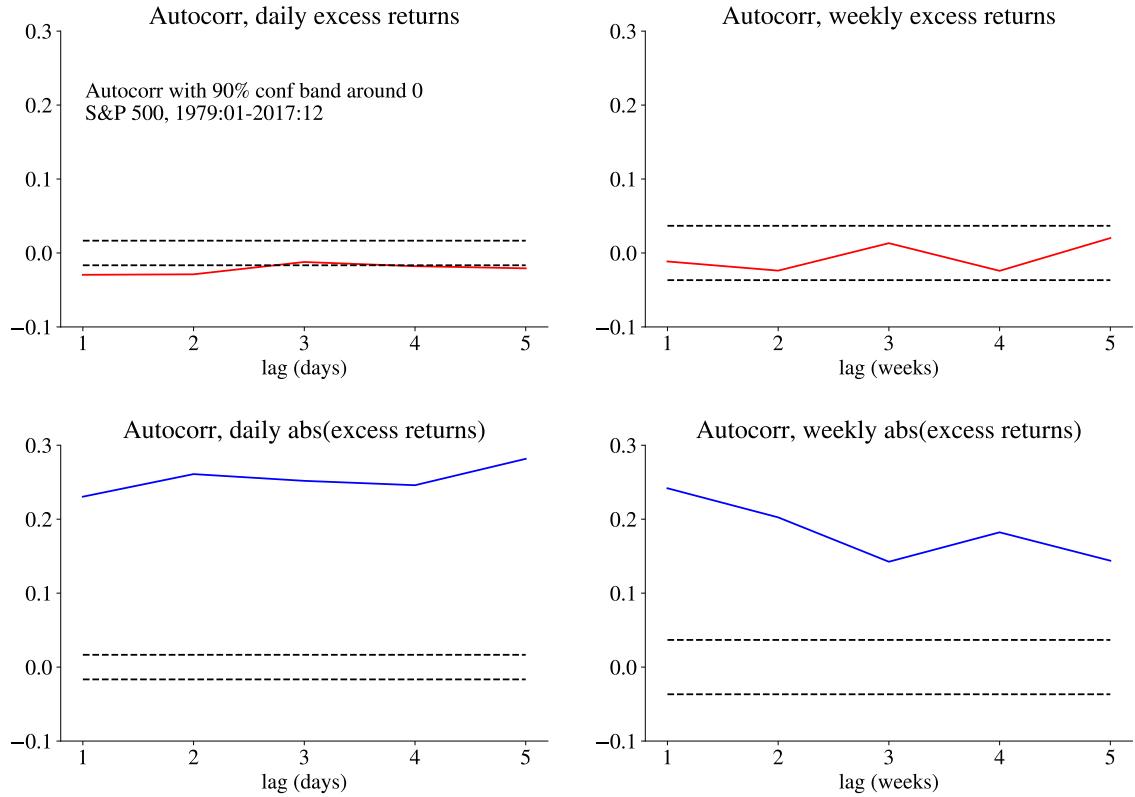


Figure 9.1: Predictability of daily US stock returns

series processes can be well approximated by an AR of relatively low order. See Figure 9.3 for an illustration.

The autoregression can also allow for the coefficients to depend on the market situation. For instance, consider an AR(1), but where the autoregression coefficient may be different depending on the sign of last period's return

$$y_t = \alpha + \beta Q_{t-1} y_{t-1} + \gamma(1 - Q_{t-1}) y_{t-1} + \varepsilon_t, \text{ where} \quad (9.3)$$

$$Q_{t-1} = \begin{cases} 1 & \text{if } y_{t-1} < 0 \\ 0 & \text{else.} \end{cases}$$

See Figure 9.4 for an illustration.

Inference of the slope coefficient in autoregressions on returns for longer data horizons than the data frequency (for instance, analysis of weekly returns in a data set consisting of daily observations) must be done with care. If only non-overlapping returns are used (for instance, the weekly returns recorded on Wednesdays only), the standard LS expression

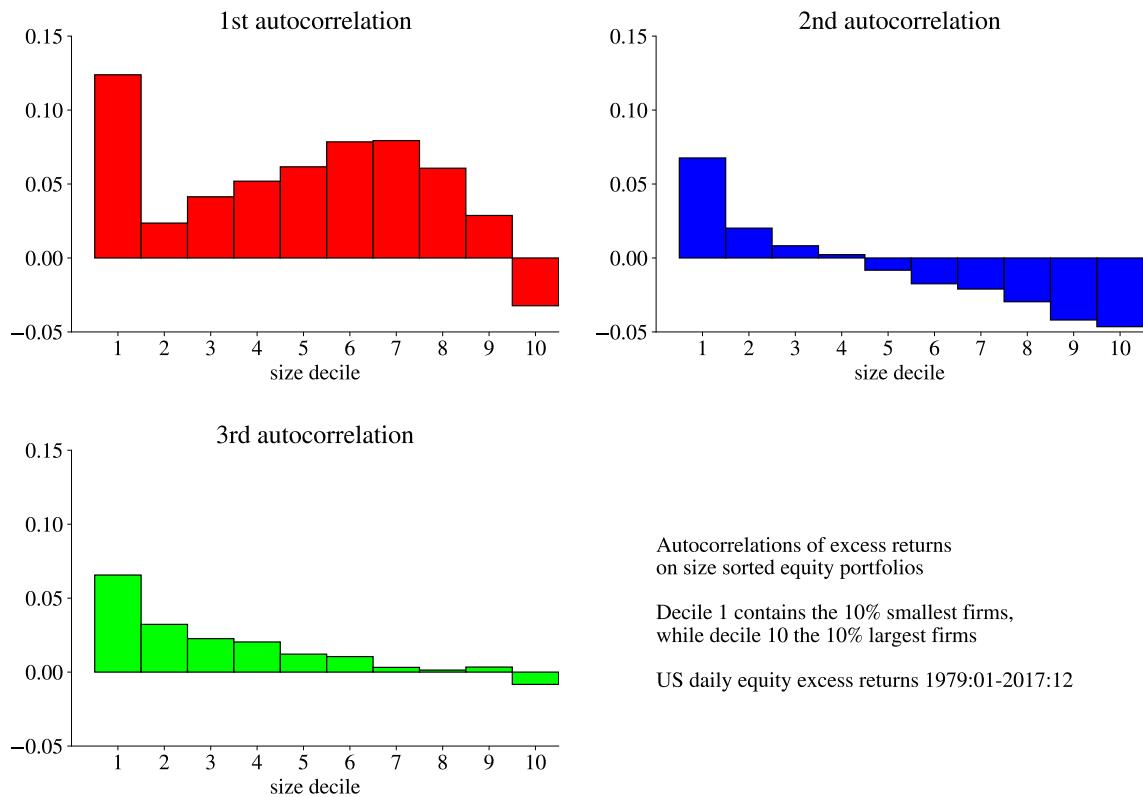


Figure 9.2: Predictability of daily US stock returns, size deciles

for the standard deviation of the autoregressive parameter is likely to be reasonable. This is not the case if overlapping returns (all daily data on weekly returns) are used.

Remark 9.1 (*Overlapping returns**) Consider an AR(1) for the two-period return, $y_{t-1} + y_t$

$$y_{t+1} + y_{t+2} = a + b_2(y_{t-1} + y_t) + \varepsilon_{t+2}.$$

Two successive observations with non-overlapping returns are then

$$y_{t+1} + y_{t+2} = a + b_2(y_{t-1} + y_t) + \varepsilon_{t+2}$$

$$y_{t+3} + y_{t+4} = a + b_2(y_{t+1} + y_{t+2}) + \varepsilon_{t+4}.$$

Suppose that y_t is not autocorrelated, so the slope coefficient $b_2 = 0$. We can then write

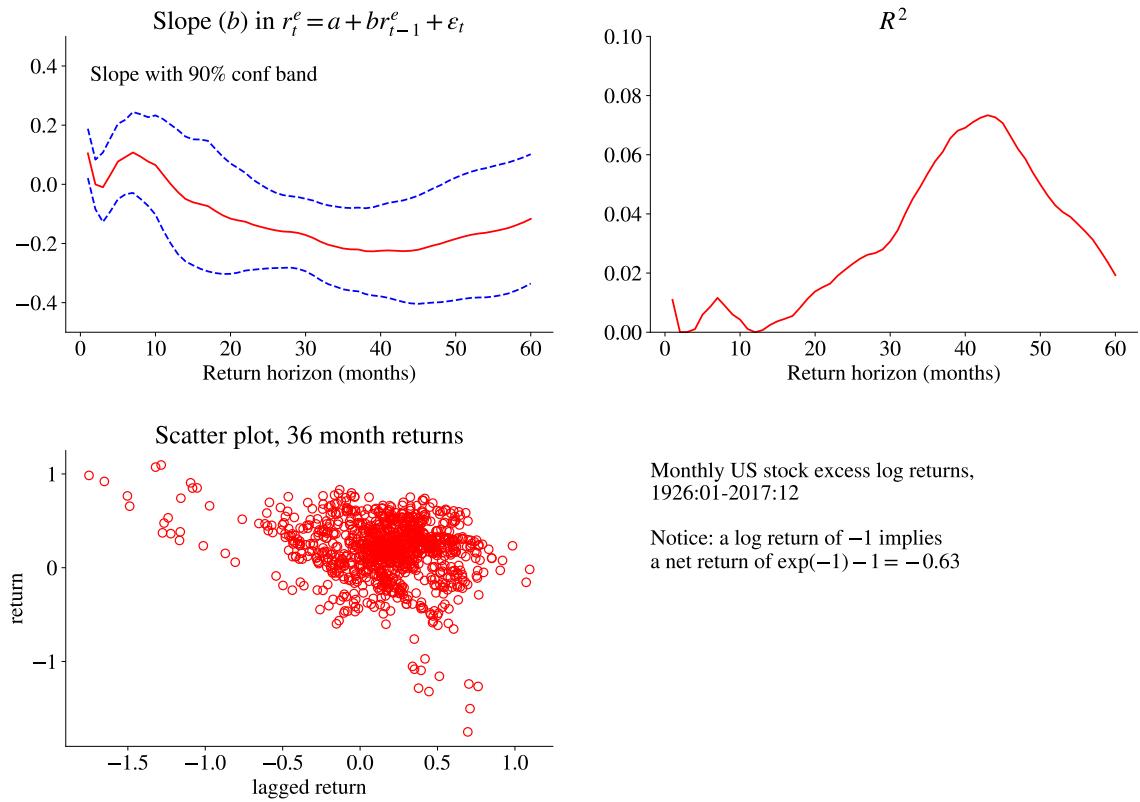


Figure 9.3: Predictability of monthly US stock returns

the residuals as

$$\varepsilon_{t+2} = -a + y_{t+1} + y_{t+2}$$

$$\varepsilon_{t+4} = -a + y_{t+3} + y_{t+4},$$

which are uncorrelated. Compare this to the case where we use overlapping data. Two successive observations are then

$$y_{t+1} + y_{t+2} = a + b_2(y_{t-1} + y_t) + \varepsilon_{t+2}$$

$$y_{t+2} + y_{t+3} = a + b_2(y_t + y_{t+1}) + \varepsilon_{t+3}.$$

As before, $b_2 = 0$ if y_t has no autocorrelation, so the residuals become

$$\varepsilon_{t+2} = -a + y_{t+1} + y_{t+2}$$

$$\varepsilon_{t+3} = -a + y_{t+2} + y_{t+3},$$

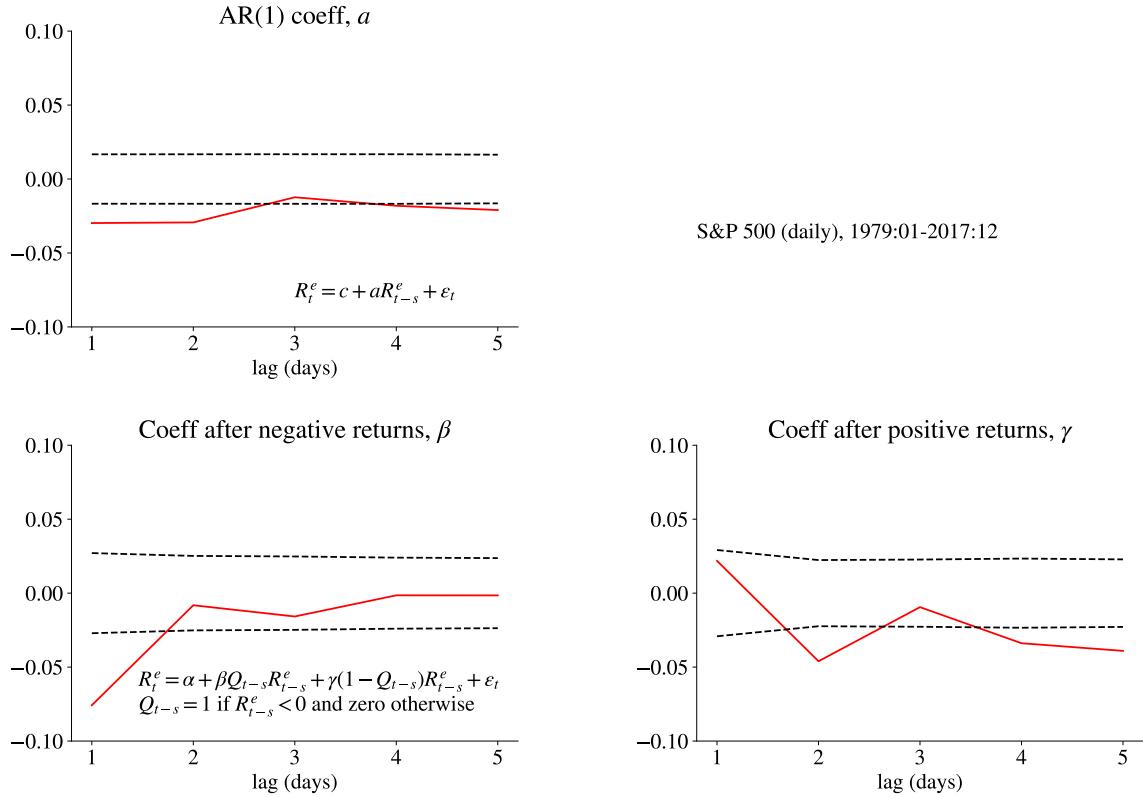


Figure 9.4: Predictability of daily US stock returns, asymmetric model

which are correlated since y_{t+2} shows up in both. This demonstrates that overlapping return data introduces autocorrelation of the residuals—which has to be handled in order to make correct inference.

9.1.3 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

$$VR_q = \frac{\text{Var}\left(\sum_{s=0}^{q-1} y_{t-s}\right)}{q \text{Var}(y_t)}. \quad (9.4)$$

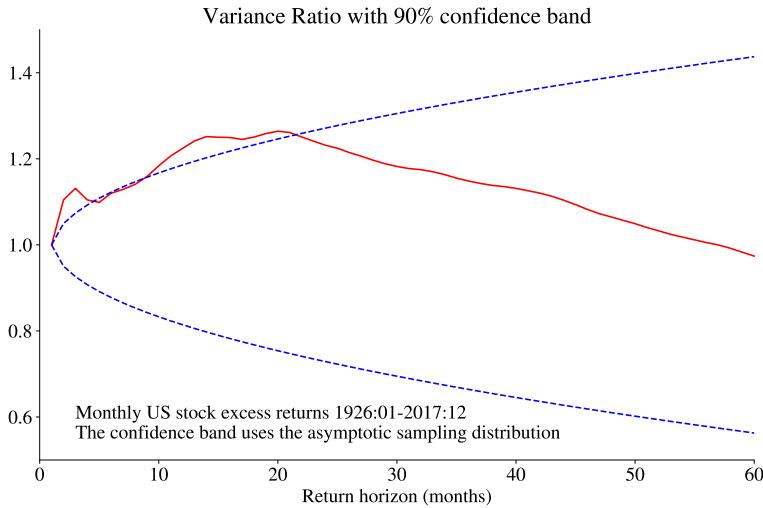


Figure 9.5: Variance ratios, monthly US excess stock returns

To see that this is related to predictability, consider the 2-period variance ratio.

$$\begin{aligned}
 VR_2 &= \frac{\text{Var}(y_t + y_{t-1})}{2 \text{Var}(y_t)} & (9.5) \\
 &= \frac{\text{Var}(y_t) + \text{Var}(y_{t-1}) + 2 \text{Cov}(y_t, y_{t-1})}{2 \text{Var}(y_t)} \\
 &= 1 + \frac{\text{Cov}(y_t, y_{t-1})}{\text{Var}(y_t)} \\
 &= 1 + \rho_1. & (9.6)
 \end{aligned}$$

It is clear from (9.6) 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.

The estimation of VR_q is typically *not* done by replacing the population variances in (9.4) with the sample variances, since this would require using non-overlapping long returns—which wastes a lot of data points. For instance, if we have 24 years of data and we want to study the variance ratio for the 5-year horizon, then 4 years of data are wasted.

Instead, we typically rely on a transformation of (9.4)

$$\begin{aligned}
VR_q &= \frac{\text{Var}\left(\sum_{s=0}^{q-1} y_{t-s}\right)}{q \text{Var}(y_t)} \\
&= \sum_{s=-(q-1)}^{q-1} \left(1 - \frac{|s|}{q}\right) \rho_s \text{ or} \\
&= 1 + 2 \sum_{s=1}^{q-1} \left(1 - \frac{s}{q}\right) \rho_s.
\end{aligned} \tag{9.7}$$

To estimate VR_q , we first estimate the autocorrelation coefficients (using all available data points for each estimation) and then calculate (9.7).

See Figure 9.5 for an illustration.

Remark 9.2 (*Sampling distribution of \widehat{VR}_q) Under the null hypothesis that there is no autocorrelation, (9.1) and (9.7) give

$$\sqrt{T} (\widehat{VR}_q - 1) \xrightarrow{d} N \left[0, \sum_{s=1}^{q-1} 4 \left(1 - \frac{s}{q}\right)^2 \right].$$

Example 9.3 (Sampling distributions of \widehat{VR}_2 and \widehat{VR}_3)

$$\begin{aligned}
\sqrt{T} (\widehat{VR}_2 - 1) &\xrightarrow{d} N(0, 1) \text{ or } \widehat{VR}_2 \xrightarrow{d} N(1, 1/T) \\
\text{and } \sqrt{T} (\widehat{VR}_3 - 1) &\xrightarrow{d} N(0, 20/9) \text{ or } \widehat{VR}_3 \xrightarrow{d} N[1, (20/9)/T].
\end{aligned}$$

Remark 9.4 The results in CLM Table 2.5 and 2.6 (weekly CRSP stock index returns, early 1960s to mid 1990s) show variance ratios above one and increasing with the number of lags, q . The results for individual stocks in CLM Table 2.7 show variance ratios close to, or even below, unity. Cochrane Tables 20.5–6 report weak evidence for more mean reversion in multi-year returns (annual NYSE stock index, 1926 to mid 1990s).

9.2 Other Predictors and Methods

There are many other possible predictors of future stock returns. For instance, both the dividend-price ratio and nominal interest rates have been used to predict long-run returns, and lagged short-run returns on other assets have been used to predict short-run returns.

9.2.1 Lead-Lags

Stock indices have more positive autocorrelation than (most) individual stocks: there should therefore be fairly strong cross-autocorrelations across individual stocks. (See Campbell, Lo, and MacKinlay (1997) Tables 2.7 and 2.8.) Indeed, this is also what is found in US data where weekly returns of large size stocks forecast weekly returns of small size stocks. See Figures 9.6–9.7 for illustrations.

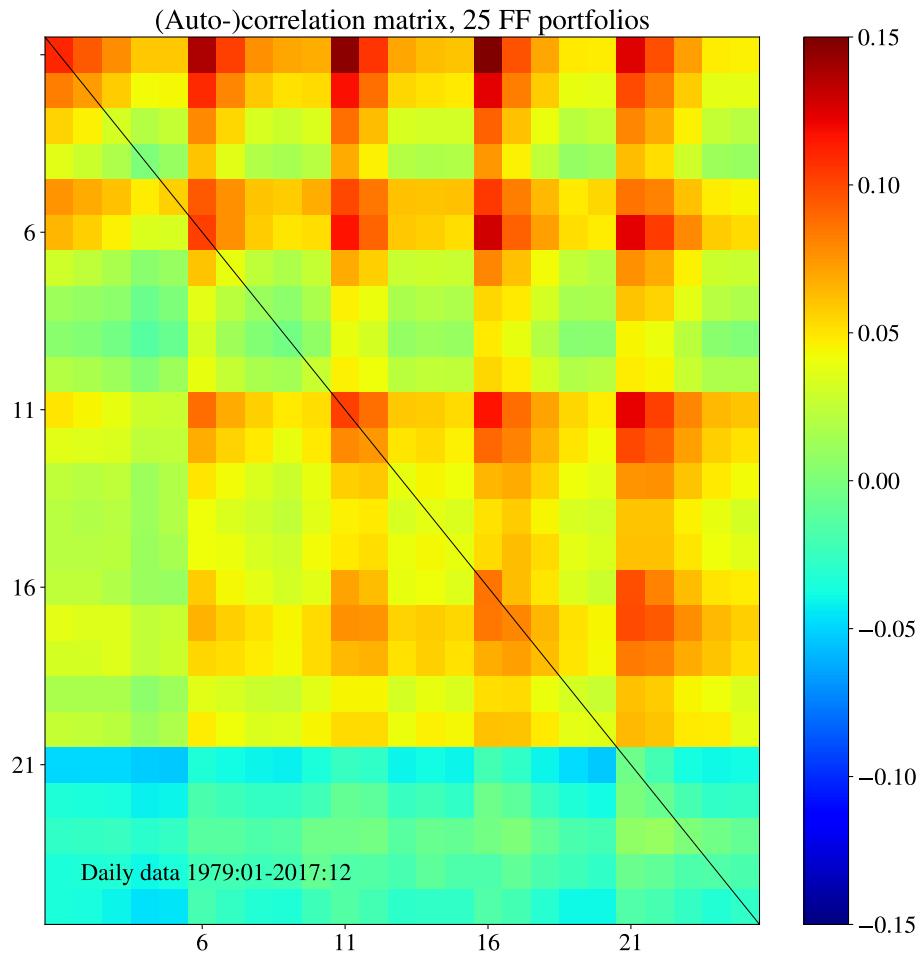


Figure 9.6: Illustration of the cross-autocorrelations, $\text{Corr}(R_t, R_{t-k})$, daily FF data. Dark colours indicate high correlations, light colours indicate low correlations.

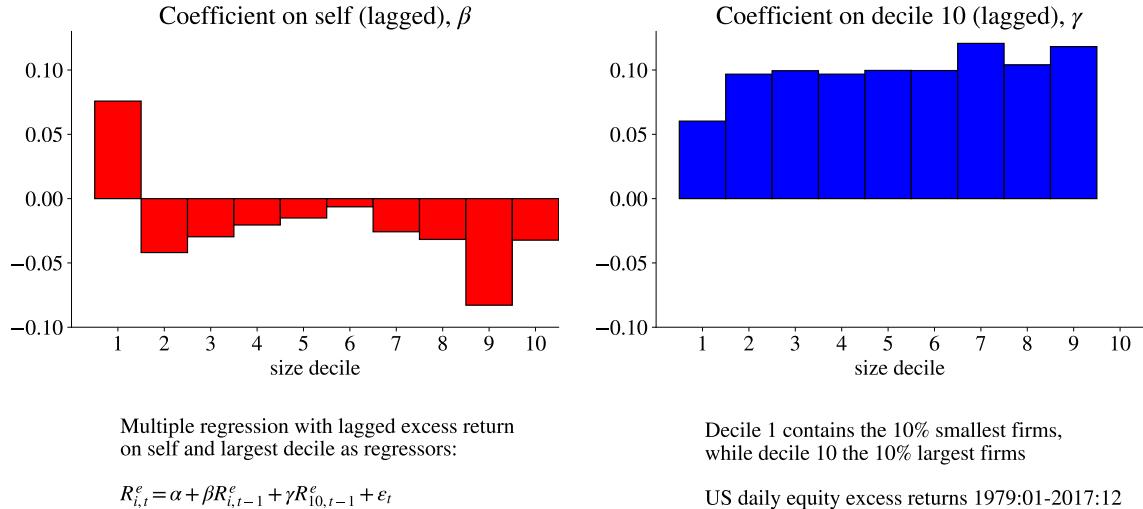


Figure 9.7: Spillover effects, daily data

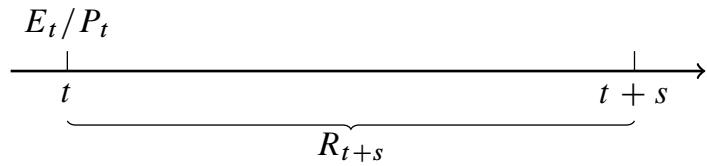


Figure 9.8: Using E/P or D/P to predict returns

9.2.2 Dividend-Price Ratio as a Predictor

One of the most successful attempts to forecast long-run returns is a regression of future returns on the current earnings-price ratio (here in logs)

$$R_{t+s} = \alpha + \beta_s \ln(E_t/P_t) + \varepsilon_{t+s}, \quad (9.8)$$

where R_{t+s} is the return from t to $t+s$. This is illustrated in Figure 9.8.

For instance, CLM Table 7.1, report R^2 values from this regression which are close to zero for monthly returns, but they increase to 0.4 for 4-year returns (US, value weighted index, mid 1920s to mid 1990s). See Figure 9.9 for an illustration.

9.2.3 Predictability but No Autocorrelation

The evidence for US stock returns is that long-run returns may perhaps be predicted by the dividend-price ratio or interest rates, but that the long-run autocorrelations are weak (long-

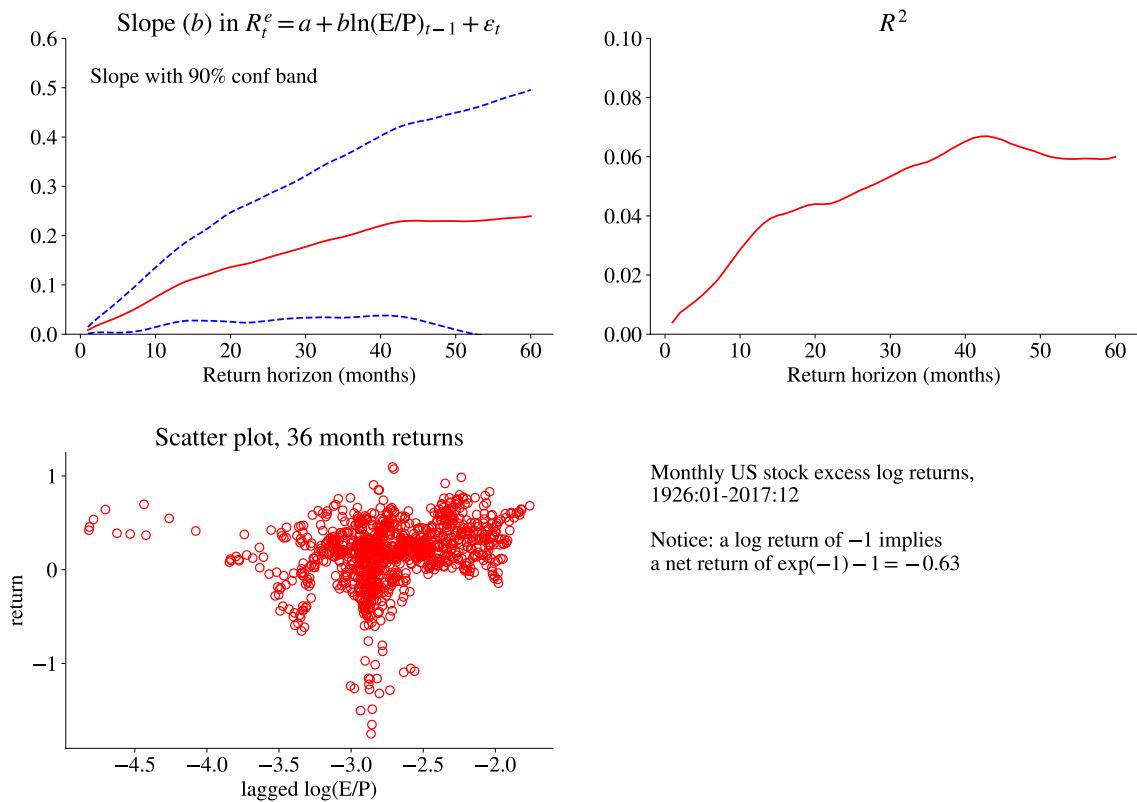


Figure 9.9: Forecast regressions, monthly US stock returns

run US stock returns appear to be “weak-form efficient” but not “semi-strong efficient”). This should remind us of the fact that predictability and autocorrelation need not be the same thing: although autocorrelation implies predictability, we can have predictability without autocorrelation.

9.3 Out-of-Sample Forecasting Performance

9.3.1 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 estimate the prediction equation using

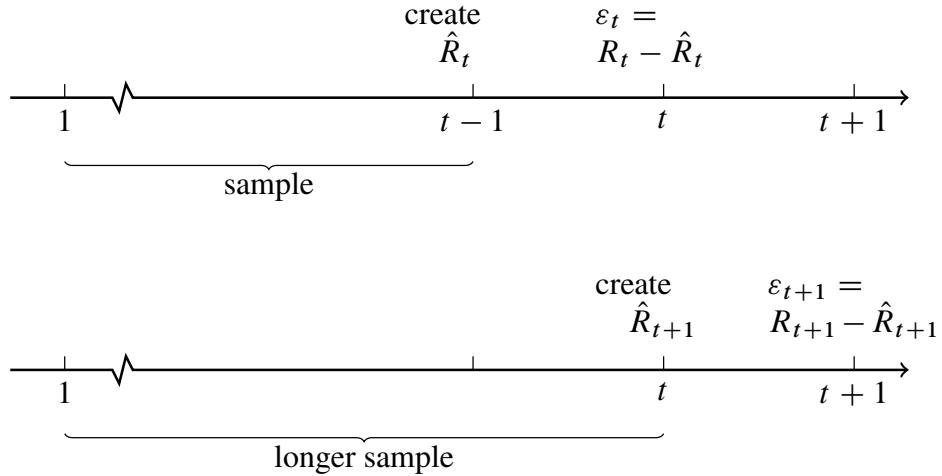


Figure 9.10: Out-of-sample forecasting

data up to and including $t-1$, and then make a forecast for period t . See Figure 9.10 for an illustration.

The forecasting performance of the equation is then compared with a benchmark prediction model (for instance, using the historical average as the predictor). Notice that this benchmark model is also estimated on data up to and including $t-1$, so it changes over time. Effectively, they are comparing the forecast performance of two models estimated in a recursive way (long and longer sample). The comparison can be done in terms of the RMSE and an “out-of-sample R^2 ”

$$R_{OS}^2 = 1 - \frac{1}{T} \sum_{t=s}^T (y_t - \hat{y}_t)^2 / \frac{1}{T} \sum_{t=s}^T (y_t - \tilde{y}_t)^2, \quad (9.9)$$

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 $t-1$). In practice, the paper uses the historical average as the benchmark prediction model.

See Figures 9.11–9.12 for an illustrations. The evidence suggests that the in-sample long-run predictability vanishes out-of-sample. It also suggests that there is still some short-run predictability for small firm stocks.

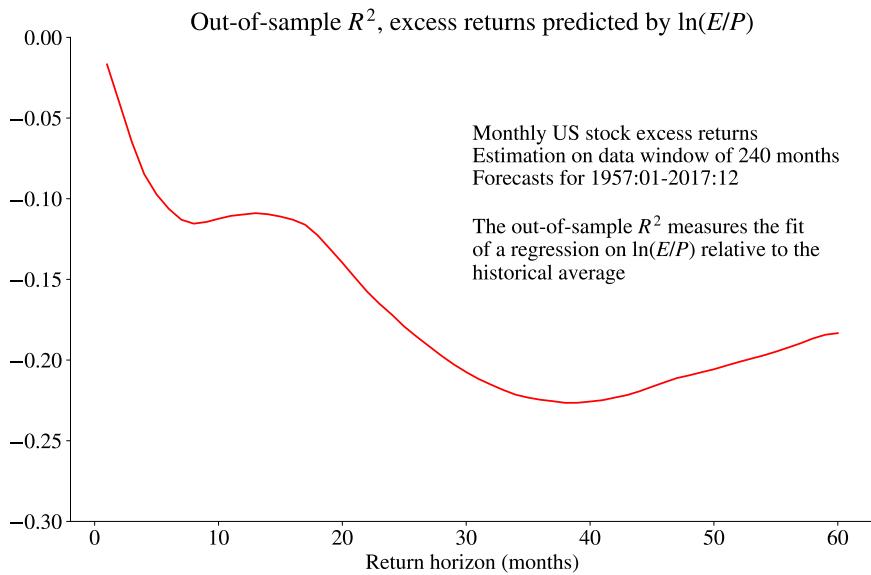


Figure 9.11: Predictability of monthly US stock returns, out-of-sample

9.3.2 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” of this strategy in relation to some benchmark portfolios. The trading strategy should, of course, be based on the variable that is supposed to forecast returns.

A common way (since Jensen, updated in Huberman and Kandel (1987)) is to study the performance of a portfolio by running the following regression

$$R_{pt}^e = \alpha + \beta R_{mt}^e + \varepsilon_t, \text{ with} \quad (9.10)$$

$$\mathbb{E} \varepsilon_t = 0 \text{ and } \text{Cov}(R_{mt}^e, \varepsilon_t) = 0,$$

where R_{pt}^e is the excess return on the portfolio being studied and R_{mt}^e the excess returns of a vector of benchmark portfolios (for instance, only the market portfolio if we want to rely on CAPM. Neutral performance requires $\alpha = 0$, which can be tested with a t test.

See Figure 9.13 for an illustration.

9.3.3 More Evidence on Trading Strategies

Figure 9.14 illustrates the *out-of-sample performance on daily returns*. Figure 9.14 shows that extreme S&P 500 returns are followed by mean-reverting movements the following

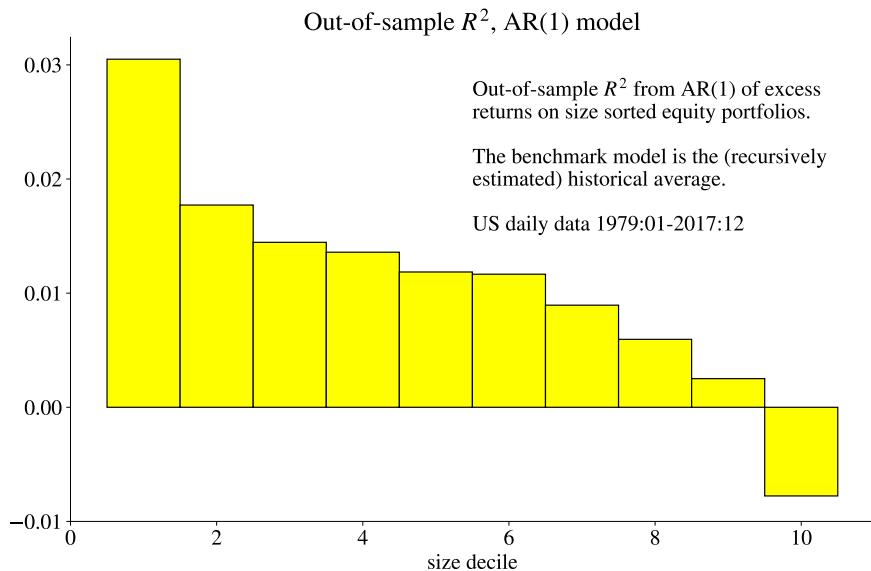


Figure 9.12: Predictability of daily US stock returns, out-of-sample

day (negative autocorrelation)—which suggests that a trading strategy should sell after a high return and buy after a low return. Compare with Figure 9.2.

Figure 9.15 compares the results for different size categories —and illustrates that there is more predictability (indicating positive autocorrelation) for small stocks. Once again, compare with Figure 9.2.

Figure 9.16 illustrates the *out-of-sample performance on long-run returns*. Figure 9.16 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 9.9 (in-sample, but there we used e/p , not p/e) but also Figure 9.11 (out-of-sample).

9.3.4 Technical Analysis

Main reference: Bodie, Kane, and Marcus (2002) 12.2; Reilly and Brown (2012) 16; Neely (1997) (overview, foreign exchange market)

Further reading: Murphy (1999) (practical, a believer's view); The Economist (1993) (overview, the perspective of the early 1990s); Brock, Lakonishok, and LeBaron (1992) (empirical, stock market); Lo, Mamaysky, and Wang (2000) (academic article on return distributions for “technical portfolios”)

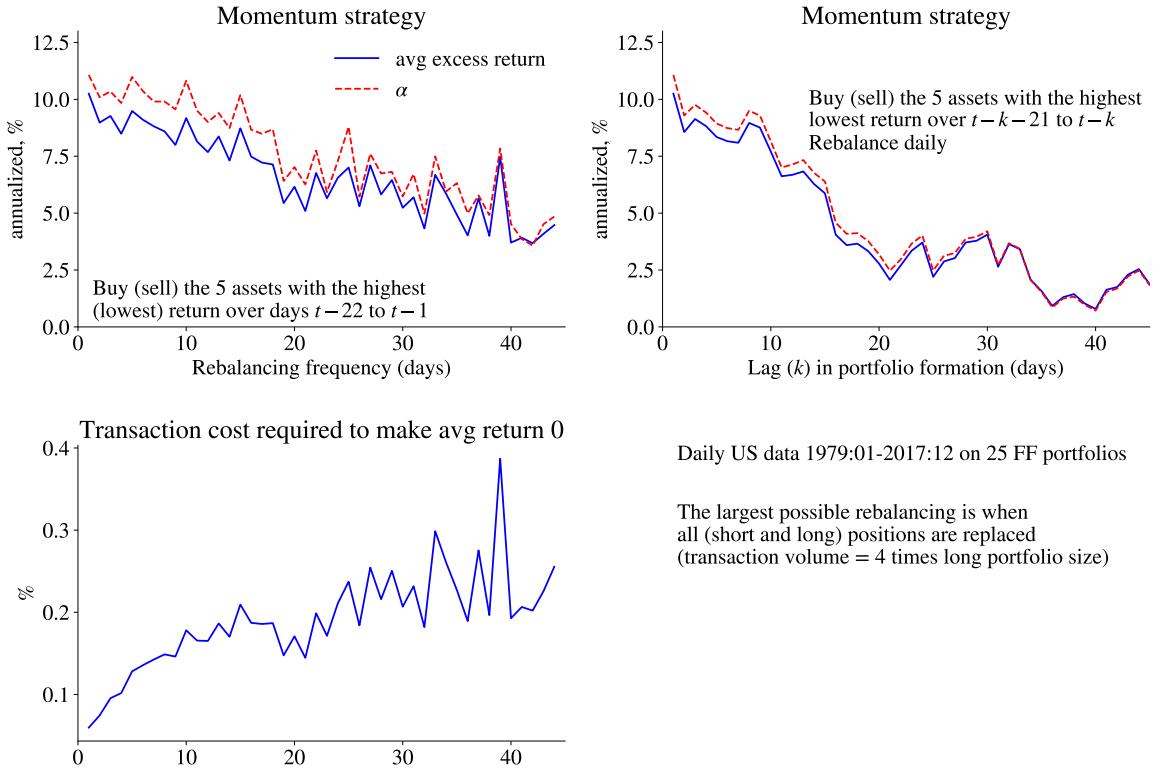


Figure 9.13: Predictability of daily US stock returns, momentum strategy

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. The logic is that an observed price move must be due to some news (exactly which one is not very important) and that old patterns can tell us where the price will move in the near future. This is an attempt to gather more detailed information than that used by the market as a whole. In practice, much of technical analysis amounts to plotting different transformations (for instance, a moving average) of prices—and to spot systematic patterns. It is also common to incorporate information from other markets (for instance, the CBOE put/call ratio is often interpreted as a bearish attitude), trading volume or measures of market wide trends (the “breadth” of the market compares the number of assets with price increase/decrease).

This section summarizes some simple trading rules.

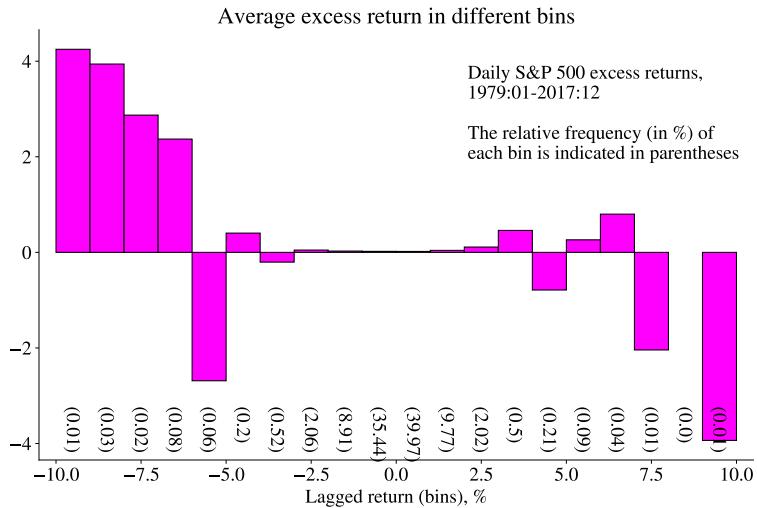


Figure 9.14: Predictability of daily US stock returns, out-of-sample

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.¹

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, with $S < L$ and let b be a bandwidth (perhaps 0.01). Then, a MA rule for period t could be

$$\begin{cases} \text{buy in } t \text{ if } MA_{t-1}(S) > MA_{t-1}(L)(1+b) \\ \text{sell in } t \text{ if } MA_{t-1}(S) < MA_{t-1}(L)(1-b) \\ \text{no change} \quad \quad \quad \text{otherwise} \end{cases}, \text{ where} \quad (9.11)$$

$$MA_{t-1}(S) = (p_{t-1} + \dots + p_{t-S})/S.$$

The difference between the two moving averages is called an *oscillator*

$$\text{oscillator}_t = MA_t(S) - MA_t(L), \quad (9.12)$$

(or sometimes, moving average convergence divergence, MACD) and the sign is taken as a trading signal (this is the same as a moving average crossing, MAC).² A version of

¹In physics, momentum equals the mass times speed.

²Yes, the rumour is true: the tribe of chartists is on the verge of developing their very own language.

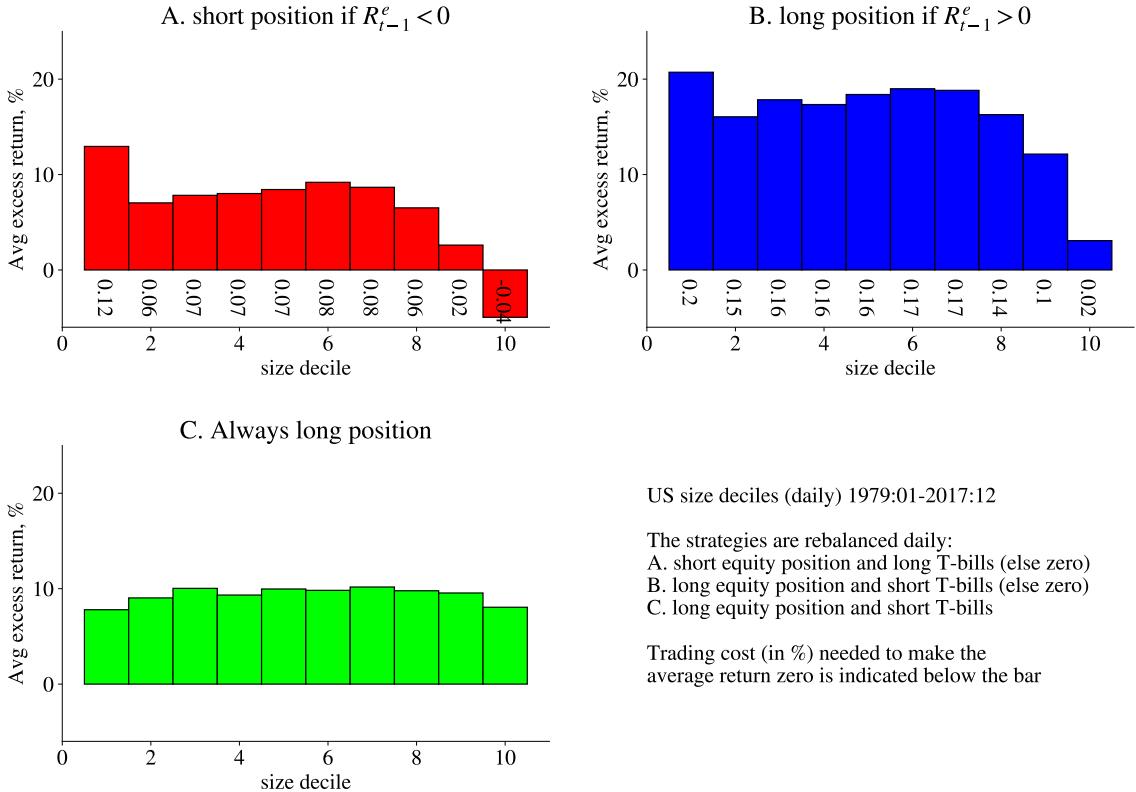


Figure 9.15: Predictability of daily US stock returns, out-of-sample

the moving average oscillator is the *relative strength index*³, 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. Yet another version is to compare the oscillator _{t} to an moving average of the oscillator (also called a signal line).

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 (perhaps because they believe that prices cannot pass this level; clear risk of circular reasoning or self-fulfilling prophecies). Round numbers often play the role as resistance levels. Once this artificial resistance level has been broken, the price can possibly rise substantially. 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 proxied by minimum and maximum values over a data

³Not 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).

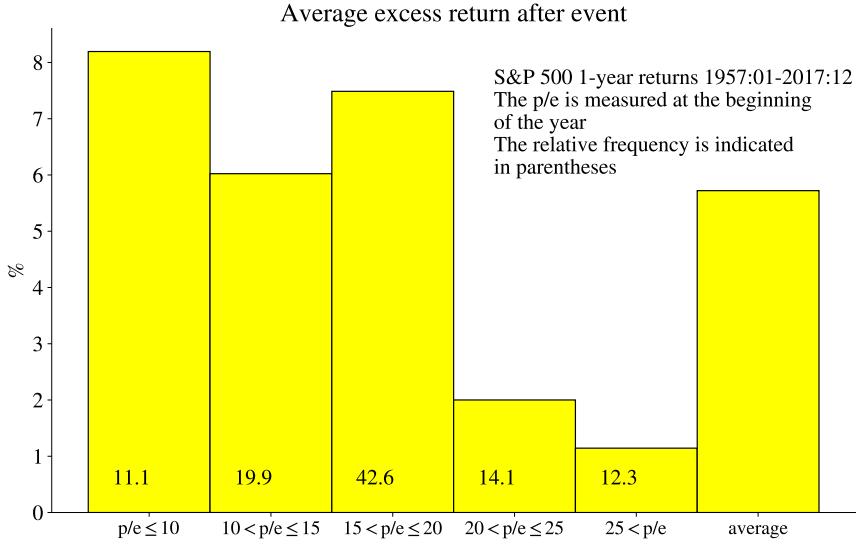


Figure 9.16: Predictability of annual US stock returns, out-of-sample

window of length L . With a bandwidth b (perhaps 0.01), the rule for period t could be

$$\begin{bmatrix} \text{buy in } t \text{ if } P_t > M_{t-1}(1 + b) \\ \text{sell in } t \text{ if } P_t < m_{t-1}(1 - b) \\ \text{no change otherwise} \end{bmatrix}, \text{ where} \quad (9.13)$$

$$M_{t-1} = \max(p_{t-1}, \dots, p_{t-S})$$

$$m_{t-1} = \min(p_{t-1}, \dots, p_{t-S}).$$

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 a thousand 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 basic problem with this interpretation is that there is a buyer for every seller, so we could equally well interpret the situations as if there is a buying pressure.)

Technical Analysis and Mean Reversion

If we instead believe in mean reversion of the prices, then we can essentially reverse the previous trading rules: we would typically sell when the price is high. See Figure 9.17 and Table 9.1.

Some investors argue that markets show periods of mean reversion and then periods with trends—and that both can be exploited. Clearly, the concept of support and resistance levels (or more generally, a channel) is based on mean reversion between these points. A new trend is then supposed to be initiated when the price breaks out of this band.

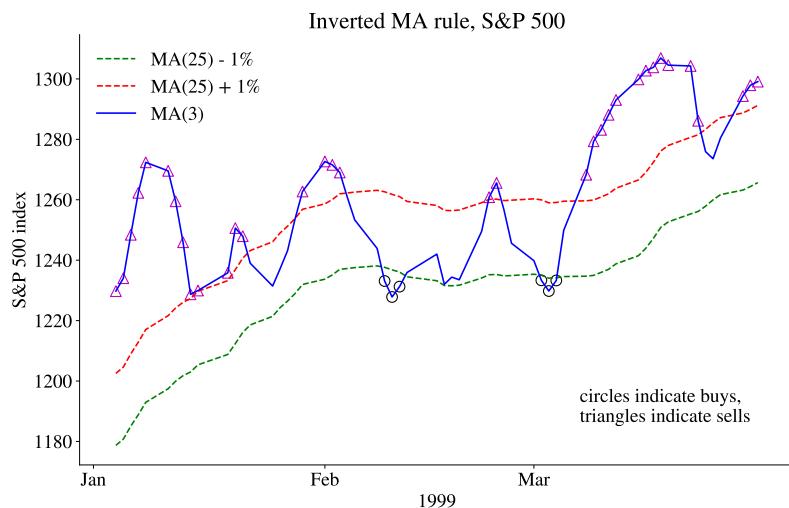


Figure 9.17: Example of a daily trading rule

9.4 Forecast Averaging

Reference: Elliot and Timmermann (2016) 14

	Mean	Std
All days	6.2	17.5
After buy signal	16.6	26.4
After neutral signal	5.3	14.3
After sell signal	1.4	13.6

Table 9.1: Excess returns (annualized, in %) from technical trading rule (Inverted MA rule). Daily S&P 500 data 1990:01-2017:12

Averaging across forecasts have often proved to be a good way of producing a superior forecast.

There are two main cases: (1) when we have access to the forecasts and also the data/model that produced them and (2) when we have access to the forecasts only. We discuss them in reverse order.

Suppose we have access to K different forecasts (\hat{R}_t^i for $i = 1$ to K) of the return R_t . All these forecasts are made in period $t - h$ (with $h \geq 1$). We form a weighted average as

$$R_t^* = \sum_{i=1}^K w_i \hat{R}_t^i, \text{ with } \sum_{i=1}^K w_i = 1. \quad (9.14)$$

For instance, w 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 various forecasts. It can then be argued that the proper way to proceed is to pool all the data and apply the model selection techniques. However, the unweighted average across forecasts often perform reasonably well.

9.5 Evaluating Forecasting Performance

Further reading: Diebold (2001) 11; Stekler (1991); Diebold and Mariano (1995); Clark and West (2007)

To do a solid evaluation of the forecast performance (of some forecaster/forecast method/forecast institute), we need a sample (history) of the forecasts and the resulting forecast errors. The reason is that the forecasting performance for a single period is likely to be dominated by luck, so we can only expect to find systematic patterns by looking at

results for several periods.

To set up tests of the forecasting performance, let ε_t be the forecast error in period t

$$\varepsilon_t = R_t - \hat{R}_t, \quad (9.15)$$

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.)

Quite often, we compare a forecast method (or forecasting institute) with a benchmark forecast like a “no change,” 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 should be done based on the true 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

$$R_t = \beta_0 + \beta_1 x_{t-h} + \varepsilon_t \quad (9.16)$$

to minimize the sum of squared residuals. This will, among other things, give a zero mean of the fitted residuals and also a zero correlation between the fitted residual and 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 to the forecast method, irrespective of whether a LS regression has been used or not. In practice, this means studying (i) whether the forecast error, e_t , has a zero mean; (ii) the mean squared (or absolute value) of the forecast error ; (iii) the fraction of times the squared (or absolute value) of the forecast error is lower than some threshold; (iv) the profit from investing by following a forecasting model; (v) if the forecast errors are autocorrelated or correlated with past information.

To perform formal tests of forecasting performance a Diebold and Mariano (1995) test is typically performed. To implement it, consider two different forecasts. For instance,

the first forecast could come from a naive forecasting model (for instance, no change) that you hope to beat (forecast errors e_t) and the other is your estimated model (forecast errors ε_t). To test the different aspects discussed before, let $\delta(x)$ be an indicator function that is one if x is true and zero otherwise, and let R_t^e and R_t^ε denote the returns from following trading strategies based on the different forecasts. Then, we could consider, for instance, the following moment conditions

$$g_t = e_t - \varepsilon_t, \text{ or} \quad (9.17)$$

$$g_t = e_t^2 - \varepsilon_t^2 \text{ or } g_t = |e_t| - |\varepsilon_t|, \text{ or} \quad (9.18)$$

$$g_t = \delta[\text{sign}(\tilde{R}_t) \neq \text{sign}(R_t)] - \delta[\text{sign}(\hat{R}_t) \neq \text{sign}(R_t)], \quad (9.19)$$

where $\delta(x) = 1$ if x is true and zero otherwise.

The different moment conditions correspond to the different aspects of the forecasts discussed above. For instance, (9.17) is for testing if the two methods have the same average forecast error, while (9.18) tests the MSE, which is an application of the Mariano-Diebold approach. In contrast, (9.19) tests if the e model forecasts the wrong sign of the return more often than the ε model does.

From the usual properties of a sample average and the assumption that g_t is not auto-correlated, we typically have that

$$\bar{g} \xrightarrow{d} N(0, \text{Var}(g_t)/T), \quad (9.20)$$

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 can be used to construct a t-test. The tests are typically two-sided (are the forecast errors different?), but it sometimes makes sense to use a one-sided test, in particular, when testing nested models (are the errors from the larger model “smaller” than the errors from the smaller model?).

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 \(2001\)](#)), which typically means that instead of studying $g_t = e_t^2 - \varepsilon_t^2$, we use $g_t = 2e_t(e_t - \varepsilon_t)$.

Remark 9.5 (*Empirical results on predicting annual S&P 500 returns*) [Table 9.2](#) summarizes results. The combined model seems to do slightly better than the two individual models.

	AR(1)		E/P		Combination	
	mean	t-stat	mean	t-stat	mean	t-stat
MSE in-sample	291.70		275.99			
R^2_{oos}	-0.05		-0.05		-0.02	
$e - \varepsilon$	0.25	1.95	-1.27	-1.16	-0.51	-0.91
$e^2 - \varepsilon^2$	-14.49	-1.60	-15.59	-0.62	-6.40	-0.53
$ e - \varepsilon $	-0.23	-1.52	-0.68	-0.90	-0.25	-0.66
wrong sign	0.00		-0.04	-1.48		
$2e(e - \varepsilon)$	-13.50	-1.55	17.59	0.69	2.05	0.17

Table 9.2: Mariano-Diebold (and Clark-West) tests of forecasting 1-year S&P returns with different models. The total sample is 1946–2016, but the forecasts as made for 1971–2016. The e forecasts are the historical average returns while the ε forecasts are out-of-sample and based on the different regressions. The ‘wrong sign’ indicates the wrong sign of the forecast and takes the difference between the e model and the ε model. Estimation is done on an expanding data window. The std use a NW approach with 1 lag (year).

9.6 Security Analysts

Reference: Makridakis, Wheelwright, and Hyndman (1998) 10.1 and Elton, Gruber, Brown, and Goetzmann (2010) 26

9.6.1 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. (The number of litigations in the US after the technology boom/bust should serve as a strong reminder of this.)

9.6.2 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.

9.6.3 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 gives an annual 4% abnormal return 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.”

9.6.4 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).

9.6.5 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.

9.6.6 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.

9.6.7 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.

9.6.8 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.

Chapter 10

Maximum Likelihood Estimation

Reference: Verbeek (2012) 6

More advanced material is denoted by a star (*). It is not required reading.

10.1 Maximum Likelihood

Maximum likelihood estimation (MLE) has two main steps: (1) specify the likelihood function; (2) maximize it by choosing the parameter values. This estimator (estimation method) has very nice properties, provided the basic distributional assumptions are correct, that is, if we maximize the right likelihood function. In that case, MLE is typically the most efficient/precise estimators (at least in very large samples). ML also provides a coherent framework for testing hypotheses (including the Wald, LM, and LR tests). In addition, MLE is a way to construct a new estimator when there is no existing method that is tailor made for your model.

To understand the principle of maximum likelihood estimation, consider the following examples.

10.1.1 Example: Estimating the Mean with ML

Suppose we know $x_t \sim N(\mu, \sigma^2)$, but we don't know the value of μ (for now, assume we know the variance). Since x_t is a random variable, there is a probability of every observation and the density function of x_t is

$$L_t = \text{pdf}(x_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(x_t - \mu)^2}{\sigma^2}\right], \quad (10.1)$$

where L stands for “likelihood” and L_t denotes that this is the contribution of observation t . The basic idea of maximum likelihood estimation (MLE) is to pick model parameters to make the observed data have the highest possible probability. In case we had only observation t in our sample, we would get $\hat{\mu} = x_t$. This is the maximum likelihood estimator for this trivial sample.

What if there are T observations, x_1, x_2, \dots, x_T ? In the simplest case where x_i and x_j are independent, then the joint pdf is just the product of the individual pdfs (for instance, $\text{pdf}(x_i, x_j) = \text{pdf}(x_i) \text{pdf}(x_j)$) so

$$L = \text{pdf}(x_1) \times \text{pdf}(x_2) \times \dots \times \text{pdf}(x_T) \quad (10.2)$$

$$= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2}\left(\frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2} + \dots + \frac{(x_T - \mu)^2}{\sigma^2}\right)\right] \quad (10.3)$$

We want to maximize L with respect to μ . We clearly get the same result if we instead maximize $\ln L$ (since the logarithmic function is an increasing monotone function), which is sometimes easier. Therefore, take logs (to get the *log likelihood function*)

$$\ln L = \sum_{t=1}^T \ln L_t, \text{ where} \quad (10.4)$$

$$\ln L_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (x_t - \mu)^2. \quad (10.5)$$

See Figure 10.1 for an illustration. For instance, the left hand figure shows that a distribution centered on $\mu = 1.4$ gives a higher pdf for a data point $x_t = 1.4$ than a distribution with $\mu = 1.6$. This is summarized in the right hand figure which shows the log likelihood value for $x_t = 1.4$ (as a function of μ). However, for another x_t value, another μ might be better. For a sample (here with 2 data points), the optimal μ is a compromise.

The derivative with respect to μ is

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} [(x_1 - \mu) + (x_2 - \mu) + \dots + (x_T - \mu)]. \quad (10.6)$$

To maximize the likelihood function, find the value of $\hat{\mu}$ that makes $\partial \ln L / \partial \mu = 0$. Clearly, that is the usual sample average

$$\hat{\mu} = (x_1 + x_2 + \dots + x_T) / T. \quad (10.7)$$

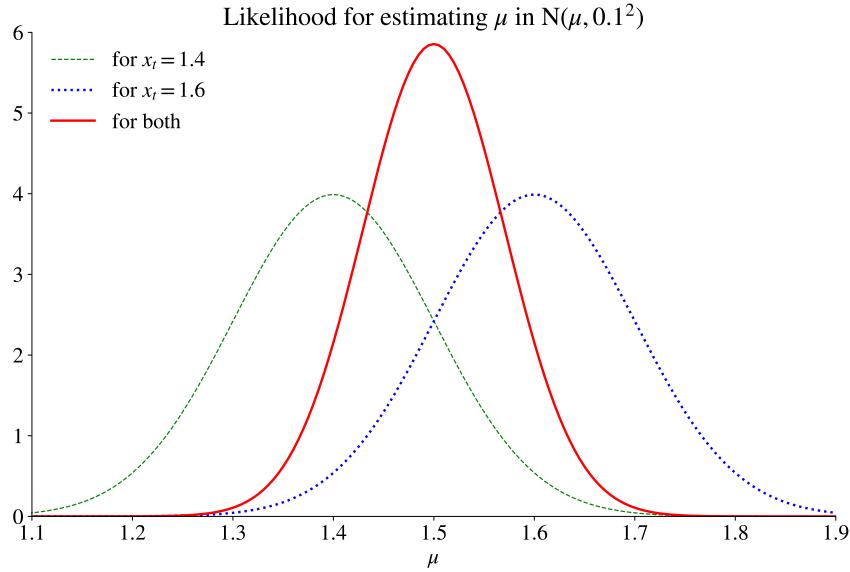


Figure 10.1: Loglikelihood function for estimating a sample mean

10.1.2 Example: Estimating the Variance with ML

To instead estimate the variance (now assuming you know the mean μ), use (10.4) and find the value σ^2 that makes $\partial \ln L / \partial \sigma^2 = 0$

$$\begin{aligned} 0 &= \frac{\partial \ln L}{\partial \sigma^2} \\ &= -\frac{T}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} [(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_T - \mu)^2], \end{aligned} \quad (10.8)$$

so

$$\hat{\sigma}^2 = \frac{1}{T} [(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_T - \mu)^2]. \quad (10.9)$$

This is clearly the usual formula for a sample variance, although we divide by T , not by $T - 1$ (so $\hat{\sigma}^2$ is biased, but the bias disappears as $T \rightarrow \infty$).

10.1.3 MLE of a Regression

To apply this idea to a (multiple) regression model

$$y_t = \beta' x_t + u_t, \quad (10.10)$$

we assume that u_t is iid $N(0, \sigma^2)$. The probability density function of u_t is then

$$\text{pdf}(u_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}u_t^2/\sigma^2). \quad (10.11)$$

Since the errors are independent, we get the joint pdf of the u_1, u_2, \dots, u_T by multiplying the marginal pdfs of each of the errors

$$\begin{aligned} L &= \text{pdf}(u_1) \times \text{pdf}(u_2) \times \dots \times \text{pdf}(u_T) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2}\left(\frac{u_1^2}{\sigma^2} + \frac{u_2^2}{\sigma^2} + \dots + \frac{u_T^2}{\sigma^2}\right)\right]. \end{aligned} \quad (10.12)$$

We divide by the same σ^2 in all terms, since we assumed iid residuals—which implies (among other things) that the variance is the same for all observations.

Substitute $y_t - \beta'x_t$ for u_t and take logs to get the log likelihood function of the sample

$$\ln L = \sum_{t=1}^T \ln L_t, \text{ where} \quad (10.13)$$

$$\ln L_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_t - \beta'x_t)^2. \quad (10.14)$$

Suppose (for simplicity) that we happen to know the value of σ^2 . It is then clear that this likelihood function is maximized by minimizing the last term, which is proportional to the sum of squared errors: LS is ML when the errors are iid normally distributed (but only then). (This holds also when we do not know the value of σ^2 —just slightly messier to show it.) See Figure 10.2.

Example 10.1 Consider the regression model $y_t = \beta_1 x_t + u_t$, where we (happen to) know that $u_t \sim N(0, 1)$. 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

With $\beta_1 = 1.6$ we get

<u>t</u>	<u>u_t</u>	<u>u_1^2</u>	<u>ln pdf(u_1)</u>
1	$-1.5 - \mathbf{1.6} \times (-1) = 0.1$	0.01	-0.924
2	$-0.6 - \mathbf{1.6} \times 0 = -0.6$	0.36	-1.099
3	$2.1 - \mathbf{1.6} \times 1 = 0.5$	0.25	-1.044
<i>sum</i>	0	0.62	-3.067

With $\beta = 1.8$ and $\beta = 2.0$ we instead get

<u>t</u>	<u>u_t</u>	<u>u_1^2</u>	<u>ln pdf(u_1)</u>
1	$-1.5 - \mathbf{1.8} \times (-1) = 0.3$	0.09	-0.964
2	$-0.6 - \mathbf{1.8} \times 0 = -0.6$	0.36	-1.099
3	$2.1 - \mathbf{1.8} \times 1 = 0.3$	0.09	-0.964
<i>sum</i>	0	0.54	-3.027

<u>t</u>	<u>u_t</u>	<u>u_1^2</u>	<u>ln pdf(u_1)</u>
1	$-1.5 - \mathbf{2.0} \times (-1) = 0.5$	0.25	-1.044
2	$-0.6 - \mathbf{2.0} \times 0 = -0.6$	0.36	-1.099
3	$2.1 - \mathbf{2.0} \times 1 = 0.1$	0.01	-0.924
<i>sum</i>	0	0.62	-3.067

Among these alternatives, $\beta = 1.8$ has the highest log likelihood value (it is actually the optimum). See Figure 10.2.

10.1.4 MLE of a Regression with GARCH(1,1) Errors

Consider a regression model where the residuals are uncorrelated across time, but have time-varying volatility

$$y_t = b'x_t + u_t, \text{ where } u_t \text{ is } N(0, \sigma_t^2). \quad (10.15)$$

The variance follows the GARCH(1,1) process

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2. \quad (10.16)$$

(It is assumed that $\omega > 0$; $\alpha, \beta \geq 0$; and $\alpha + \beta < 1$.)

To estimate this model (that is, the parameters in $(b, \omega, \alpha, \beta)$, we could use a numerical

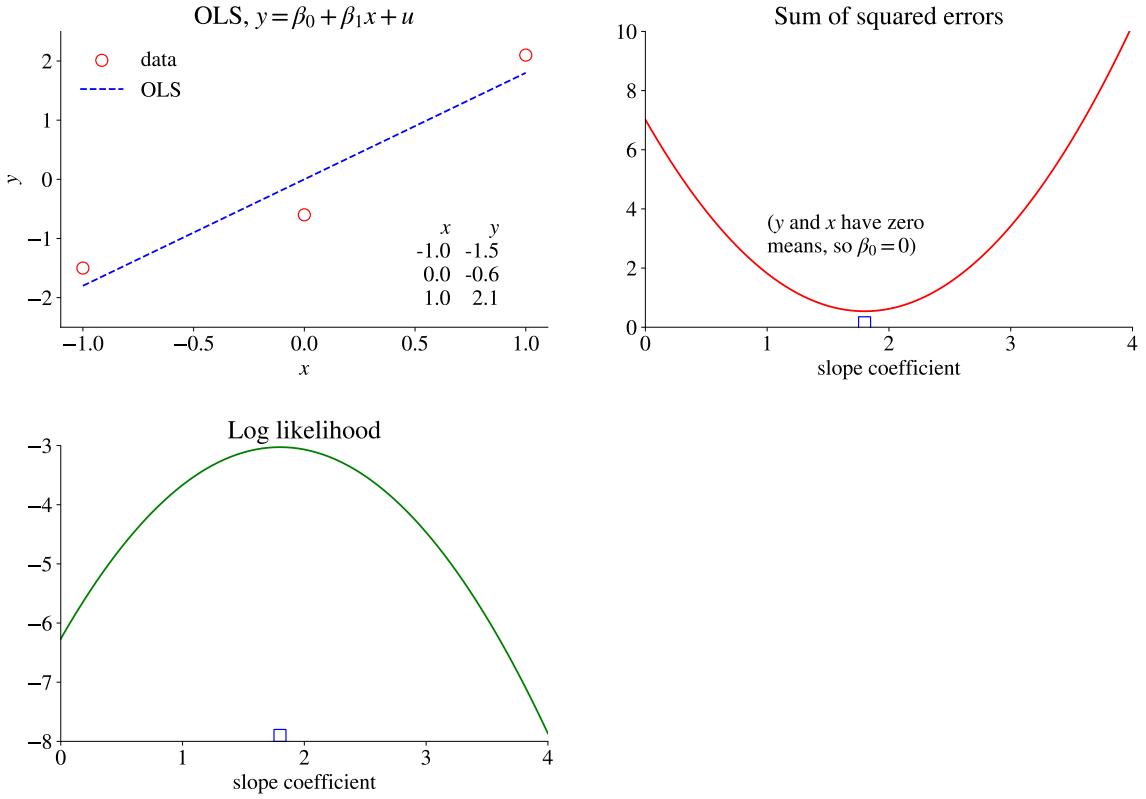


Figure 10.2: Example of OLS and ML estimation

optimization routine to maximize the log likelihood function

$$\ln L = \sum_{t=1}^T L_t, \text{ where } L_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma_t^2 - \frac{1}{2\sigma_t^2} u_t^2, \quad (10.17)$$

where u_t is calculated from (10.15) and σ_t^2 from (10.16). This is very similar to the linear regression (10.13)–(10.14), except that there is a time-subscript (and model for) σ_t^2 .

The optimization routine searches for the values of $(b, \omega, \alpha, \beta)$ that makes the value of the log likelihood function as large as possible. (However, it may be necessary to impose the parameter restrictions mentioned above.)

10.2 Key Properties of MLE

There are no general results on small-sample properties of MLE: it can be biased or not.

In contrast, MLE has very nice asymptotic (large-sample) properties, provided we

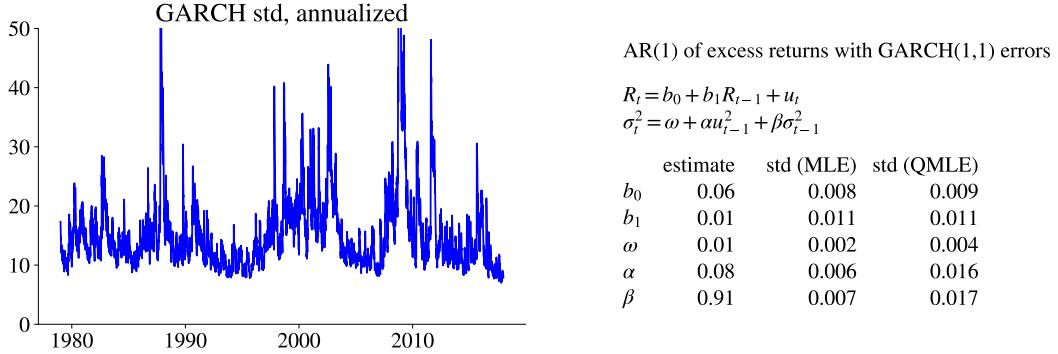


Figure 10.3: GARCH estimates

maximize the right likelihood function. If so, then:

1. MLE is consistent ($\Pr(|\hat{\beta} - \beta| > a^2) \rightarrow 0$) gets very small as T gets large, for any choice of a)
2. MLE is the most efficient/precise estimator, at least asymptotically (efficient = smallest variance)
3. MLE estimates ($\hat{\theta}$) are normally distributed,

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V), \quad (10.18)$$

$$V = I(\theta)^{-1} \text{ with } I(\theta) = -E \frac{\partial^2 \ln L}{\partial \theta \partial \theta} / T. \quad (10.19)$$

($I(\theta)$ is called the “information matrix”). The information matrix can also be written $I(\theta) = -E \frac{\partial^2 \log L_t}{\partial \theta \partial \theta}$, where $\log L_t$ is the log likelihood contribution of observation t .

4. ML also provides a coherent framework for testing hypotheses (including the Wald, LM, and LR tests).

See Figure 10.3 for an illustration.

10.2.1 Example of the Information Matrix

Differentiate (10.6), assuming we know σ^2 , to get

$$\frac{\partial^2 \ln L}{\partial \mu \partial \mu} = -\frac{T}{\sigma^2}. \quad (10.20)$$

The information matrix is

$$I(\theta) = -E \frac{\partial^2 \ln L}{\partial \theta \partial \theta} / T = \frac{1}{\sigma^2}, \quad (10.21)$$

which we combine with (10.18)–(10.19) to get

$$\sqrt{T}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2) \text{ or } \hat{\mu} \xrightarrow{d} N(\mu, \sigma^2 / T). \quad (10.22)$$

This is the standard expression for the distribution of a sample average. In practice, we replace σ^2 by the sample variance.

10.3 Three Test Principles

Wald test: estimate θ with MLE, check if $\hat{\theta} - \theta_{H_0}$ is too large. Example: t-test and F-test.

Likelihood ratio test: estimate θ with MLE as usual, estimate again by imposing the H_0 restrictions, test if $\ln L(\hat{\theta}) - \ln L(\text{"}\hat{\theta}\text{ with } H_0\text{ restrictions"}) = 0$. Example: compare the R^2 from a model without and with a restriction on some coefficient (for instance, that it is zero).

Lagrange multiplier test. Estimate θ under the H_0 restrictions, check if $\partial \ln L / \partial \theta = 0$ for unconstrained model is true when evaluated at “ $\hat{\theta}$ with H_0 restrictions.”

10.4 QMLE*

A MLE based on the wrong likelihood function (distribution) may still be useful.

Suppose we use the likelihood function \tilde{L} and get estimates $\hat{\theta}$ by

$$\frac{\partial \ln \tilde{L}}{\partial \theta} = \mathbf{0} \quad (10.23)$$

If \tilde{L} is wrong (therefore, it carries a tilde), then we are maximizing the wrong thing. For instance, we might have constructed \tilde{L} by assuming that the regression residuals are normally distributed, while they are, in fact, following a t -distribution. With some luck, we still get reasonable (consistent) estimates.

Example 10.2 (LS and QMLE) In a linear regression, $y_t = x'_t \beta + \varepsilon_t$, the first order condition for MLE based on the assumption that $\varepsilon_t \sim N(0, \sigma^2)$ is $\sum_{t=1}^T (y_t - x'_t \hat{\beta}) x_t = \mathbf{0}$. This has an expected value of zero (at the true parameters), even if the shocks have a, say, t_{22} distribution (which would define the correct likelihood function).

The example suggests that if

$$E \frac{\partial \ln \tilde{L}}{\partial \theta} = \mathbf{0}, \quad (10.24)$$

then the estimates are still consistent. We are doing *quasi-MLE* (or pseudo-MLE).

With QMLE, $\sqrt{T}(\hat{\theta} - \theta) \rightarrow^d N(0, V)$, but with the “sandwich” variance-covariance matrix

$$V = I(\theta)^{-1} E \left[\frac{\partial \ln \tilde{L}_t}{\partial \theta} \left(\frac{\partial \ln \tilde{L}_t}{\partial \theta} \right)' \right] I(\theta)^{-1}. \quad (10.25)$$

See Figure 10.3 for an illustration.

The practical implication: this is perhaps a “safer” way of constructing tests—since it is less restrictive than assuming that we have the exactly correct likelihood function.

Chapter 11

ARCH and GARCH

Reference: Bodie, Kane, and Marcus (2005) 13.4

Reference (advanced): Taylor (2005) 8–9; Verbeek (2012) 8; Campbell, Lo, and MacKinlay (1997) 12; Franses and van Dijk (2000)

11.1 Heteroskedasticity

11.1.1 Descriptive Statistics of Heteroskedasticity

Time-variation in volatility (heteroskedasticity) is a common feature of macroeconomic and 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

$$\sigma_t^2 = (u_{t-1}^2 + u_{t-2}^2 + \dots + u_{t-q}^2)/q. \quad (11.1)$$

Notice that σ_t^2 depends on lagged information, and could therefore be thought of as 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 *realised volatility*. Alternatively, this method can also be used to gauge seasonality in volatility by estimating the variance for each “season,” for instance, Mondays.

See Figures 11.1 and 11.2 for examples.

Unfortunately, the approach in (11.1) can produce quite abrupt changes in the estimate. An alternative is to apply an exponentially weighted moving average (EWMA) estimator of volatility, 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

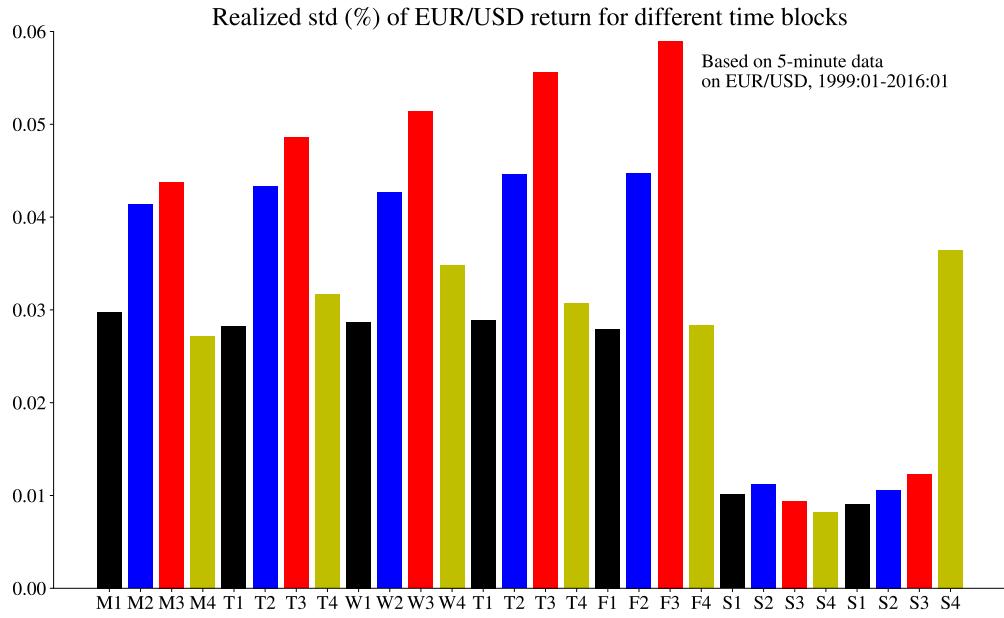


Figure 11.1: Standard deviation

$0 < \lambda < 1$, so

$$\sigma_t^2 = (1 - \lambda)(u_{t-1}^2 + \lambda u_{t-2}^2 + \lambda^2 u_{t-3}^2 + \dots). \quad (11.2)$$

See Figure 11.3 for an illustration of the weights. (The weights clearly sum to one.) This can also be calculated in a recursive way as

$$\sigma_t^2 = (1 - \lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2. \quad (11.3)$$

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 RISK Metrics uses this method with $\lambda = 0.94$ for use on daily data. Alternatively, λ can be chosen to minimize some criterion function like $\sum_{t=1}^T (u_t^2 - \sigma_t^2)^2$.

11.1.2 Predicting Realized Volatility

Volatility is often predictable, at least for horizons up to a couple of months. See Tables 11.1–11.2 for examples of very simple prediction equations.

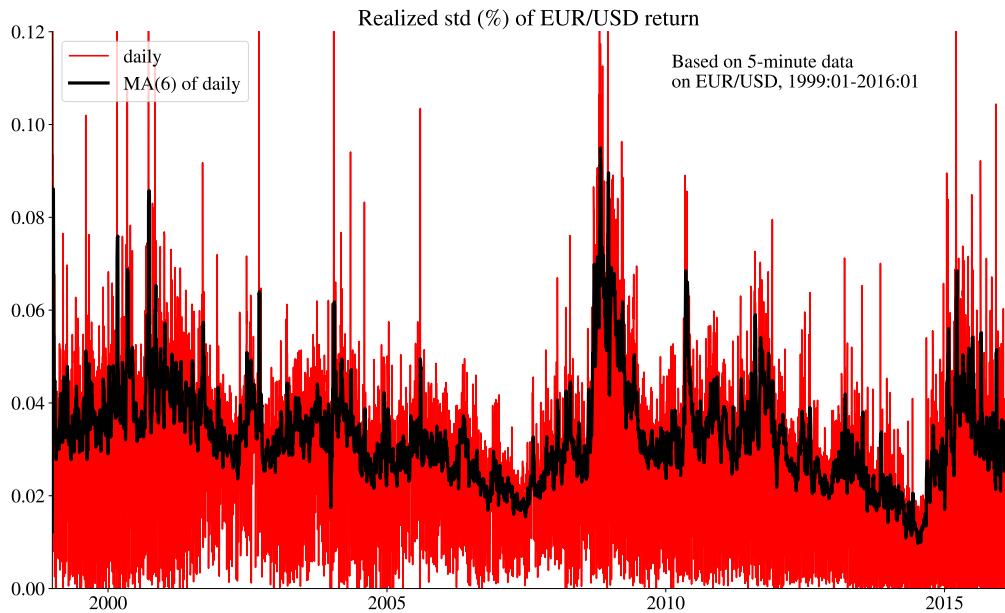


Figure 11.2: Standard deviation for EUR/USD exchange rate changes

11.1.3 Heteroskedastic Residuals in a Regression

Suppose we have a regression model

$$y_t = x_t' \beta + u_t, \text{ where} \quad (11.4)$$

$$\mathbb{E} u_t = 0 \text{ and } \text{Cov}(x_{it}, u_t) = 0.$$

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 actually are heteroskedastic. It is consistent (we get the correct values as the sample becomes really large)—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 typically not correct. To be precise, it is not heteroskedasticity in itself that invalidates the standard OLS standard errors—only the type of heteroskedasticity that is related to the (squares of) the regressors. In fact, this is exactly what White's test of heteroskedasticity is aimed at.

There are two ways to handle this problem. First, we could use MLE to incorporate the structure of the heteroskedasticity. For instance, we could combine the regression

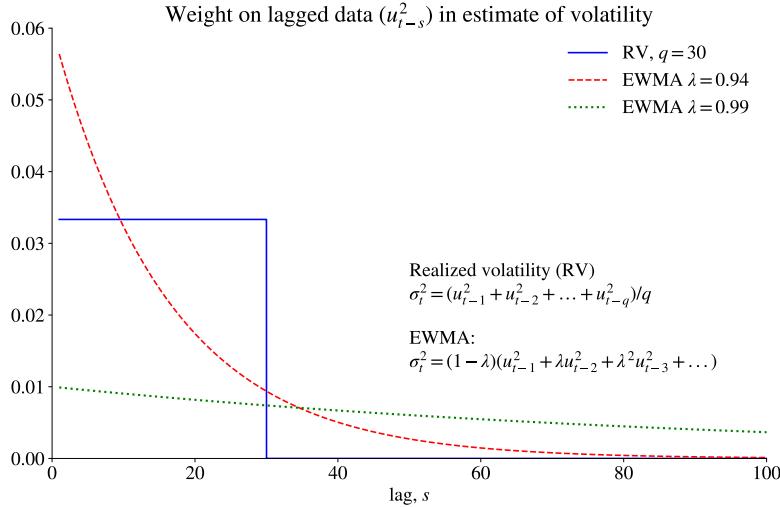


Figure 11.3: Weights on old data in the EMA approach to estimate volatility

model (11.4) with an ARCH structure of the residual. As a by-product we get the correct standard errors—provided, of course, the assumed distribution is correct. Second, we could stick to OLS, but use another expression for the variance of the coefficients: a “heteroskedasticity consistent covariance matrix,” among which “White’s covariance matrix” is the most common.

11.1.4 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 test of ARCH* is perhaps the most straightforward. Let u_t be a zero-mean variable, for instance, the residuals from OLS. Then, estimate an AR(q) model for the squares

$$u_t^2 = \omega + a_1 u_{t-1}^2 + \dots + a_q u_{t-q}^2 + v_t. \quad (11.5)$$

Under the null hypothesis of no ARCH effects, all slope coefficients are zero and the R^2 of the regression is zero. (This can be tested by noting that, under the null hypothesis, $TR^2/(1 - R^2) \sim \chi_q^2$.) Notice, however, that it is not evident that ARCH effects make the standard expression for the LS covariance matrix invalid (use White's test for that).

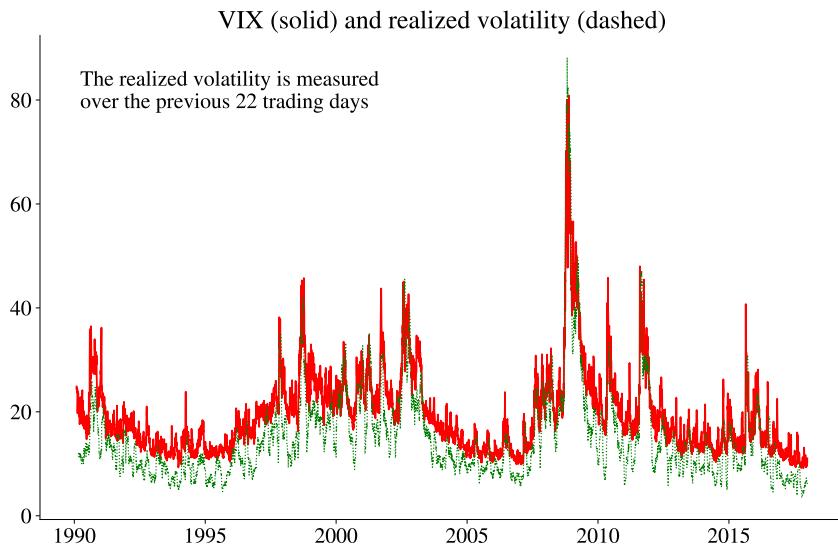


Figure 11.4: VIX and realized volatility (variance)

11.2 ARCH Models

This section discusses the Autoregressive Conditional Heteroskedasticity (ARCH) model. It is a model of how volatility depends on past volatility.

There are two basic reasons for being interested in an ARCH model. First, if residuals of the regression model (11.4) have ARCH features, then an ARCH model (that is, a specification of exactly how the ARCH features are generated) can help us estimate the regression model by maximum likelihood. Second, we may be interested in understanding the ARCH features more carefully, for instance, as an input in a portfolio choice process or for option pricing.

11.2.1 Properties of ARCH(1)

In the ARCH(1) model the residual in the regression equation (11.4), or some other zero-mean variable, can be written

$$u_t \sim N(0, \sigma_t^2), \text{ with} \quad (11.6)$$

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2, \text{ with } \omega > 0 \text{ and } 0 \leq \alpha < 1. \quad (11.7)$$

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

	(1)	(2)	(3)
log RV ($t - 22$)	0.72 (21.07)		0.12 (2.06)
log VIX ($t - 22$)		1.09 (26.39)	0.95 (12.97)
constant	0.71 (8.16)	-0.56 (-4.74)	-0.46 (-3.92)
R^2	0.52	0.62	0.63
obs	6980.00	7001.00	6980.00

Table 11.1: Regression of 22-day log realized S&P return volatility 1990:02-2017:12. 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.

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 already in $t - 1$. Therefore, the one-step ahead distribution is normal—which can be used for estimating the model with ML.

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 than a normal distribution with the same variance (excess kurtosis)—which is a common feature of financial data.

It is straightforward to show that the ARCH(1) model implies that we in period t can forecast the future conditional variance in $t + s$ as

$$E_t \sigma_{t+s}^2 = \bar{\sigma}^2 + \alpha^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1-\alpha}, \quad (11.8)$$

where $\bar{\sigma}^2$ is the unconditional variance and we recall that since σ_{t+1}^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.

See Figure 11.5 for an illustration of the fitted volatilities.

	log RV(EUR)	log RV(GBP)	log RV(CHF)	log RV(JPY)
lagged log RV	0.77 (64.58)	0.68 (30.61)	0.77 (49.16)	0.72 (49.78)
constant	-0.56 (-26.34)	-0.62 (-22.15)	-0.57 (-20.98)	-0.59 (-24.58)
D(Tue)	0.41 (18.02)	0.42 (16.02)	0.37 (15.53)	0.36 (13.11)
D(Wed)	0.34 (15.62)	0.34 (13.87)	0.32 (14.73)	0.32 (12.35)
D(Thu)	0.35 (15.55)	0.38 (15.06)	0.27 (12.59)	0.29 (11.26)
D(Fri)	0.30 (11.43)	0.34 (12.25)	0.27 (10.71)	0.30 (9.98)
R^2	0.60	0.47	0.60	0.53
obs	4454.00	4454.00	4454.00	4454.00

Table 11.2: Regression of daily log realized variance 1999:01-2016:01. All exchange rates are against the USD. The daily variances are calculated from 5 minute data. Numbers in parentheses are t-stats, based on Newey-West with 1 lag.

11.2.2 Estimation of the ARCH(1) Model

The most common way to estimate the model is to assume that $u_t \sim \text{iid } N(0, \sigma_t^2)$, as we did in (11.6), and to set up the likelihood function. The log likelihood is easily found, since the model is conditionally Gaussian. It is

$$\ln L = \sum_{t=1}^T L_t, \text{ where } L_t = -\frac{1}{2} \ln (2\pi) - \frac{1}{2} \ln \sigma_t^2 - \frac{1}{2} \frac{u_t^2}{\sigma_t^2}. \quad (11.9)$$

The estimates are found by maximizing the likelihood function (by choosing the parameters). This has to be done by a numerical optimization routine (which should preferably impose the constraints in (11.7)).

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 (11.7). Instead, if u_t is a residual from a regression equation (11.4), then we instead need to choose both the regression coefficients (β) in (11.4) and the parameters (ω and α) in (11.7).

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

$\alpha :$	$\underline{\gamma = 0}$		$\underline{\gamma = 1}$	
	0	1	0	1
Simulated	7.1	19.2	13.5	24.9
OLS formula	7.1	13.3	13.4	19.3
White's	7.0	18.5	13.3	24.3
Bootstrap	7.1	18.5	13.4	24.4
Bootstrap 2	7.0	18.5	13.3	24.3

Table 11.3: Standard error of OLS slope (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. Number of simulations: 25000. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

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 the information matrix. There is no need to use any adjusted (“White”) values.

Remark 11.1 (*Regression with ARCH(1) residuals*) To estimate the full model (11.4) and (11.7) by MLE, we can do as follows.

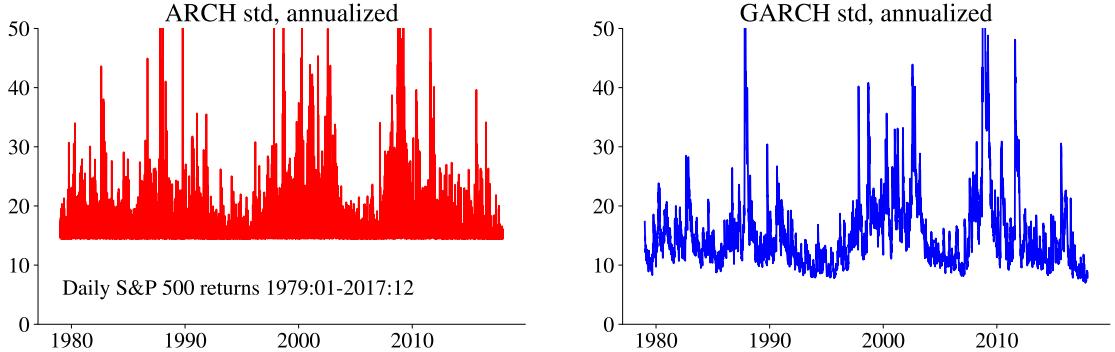
First, guess values of the parameters β (a vector if x_t is), and ω , and α . The guess of β can be taken from an LS estimation of (11.4), and the guess of ω and α from an LS estimation of $\hat{u}_t^2 = \omega + \alpha \hat{u}_{t-1}^2 + \varepsilon_t$ where \hat{u}_t are the fitted residuals from the LS estimation of (11.4).

Second, loop over the sample (first $t = 1$, then $t = 2$, etc.) and calculate u_t from (11.4) and σ_t^2 from (11.7). Plug in these numbers in (11.9) 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 11.2 (*Imposing parameter constraints on ARCH(1)*) To impose the restrictions in (11.7), iterate over values of $(\beta, \tilde{\omega}, \tilde{\alpha})$ and let $\omega = \tilde{\omega}^2$ and $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha})]$.

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 (11.9). Estimation using other likelihood functions, for instance, for a t-distribution can then be used. Or the estimation can be interpreted as a quasi-ML (is typically consistent, but requires different calculation of the covariance matrix of the parameters).



AR(1) of excess returns with ARCH(1) errors

	$R_t = b_0 + b_1 R_{t-1} + u_t$	$\sigma_t^2 = \omega + \alpha u_{t-1}^2$
b_0	0.05	0.012
b_1	-0.07	0.109
ω	0.83	0.074
α	0.30	0.106

AR(1) of excess returns with GARCH(1,1) errors

	$R_t = b_0 + b_1 R_{t-1} + u_t$	$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2$
b_0	0.06	0.009
b_1	0.01	0.011
ω	0.01	0.004
α	0.08	0.016
β	0.91	0.017

Figure 11.5: ARCH and GARCH estimates

It is straightforward to add more lags to (11.7). For instance, an ARCH(p) would be

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \dots + \alpha_p u_{t-p}^2. \quad (11.10)$$

The form of the likelihood function is the same except that we now need p starting values and that the upper boundary constraint should now be $\sum_{j=1}^p \alpha_j \leq 1$.

11.3 GARCH Models

Instead of specifying an ARCH model with many lags, it is typically more convenient to specify a low-order GARCH (Generalized ARCH) model. The GARCH(1,1) is a simple and surprisingly general model, where the volatility follows

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2, \text{with} \\ \omega > 0; \alpha, \beta &\geq 0; \text{ and } \alpha + \beta < 1. \end{aligned} \quad (11.11)$$

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

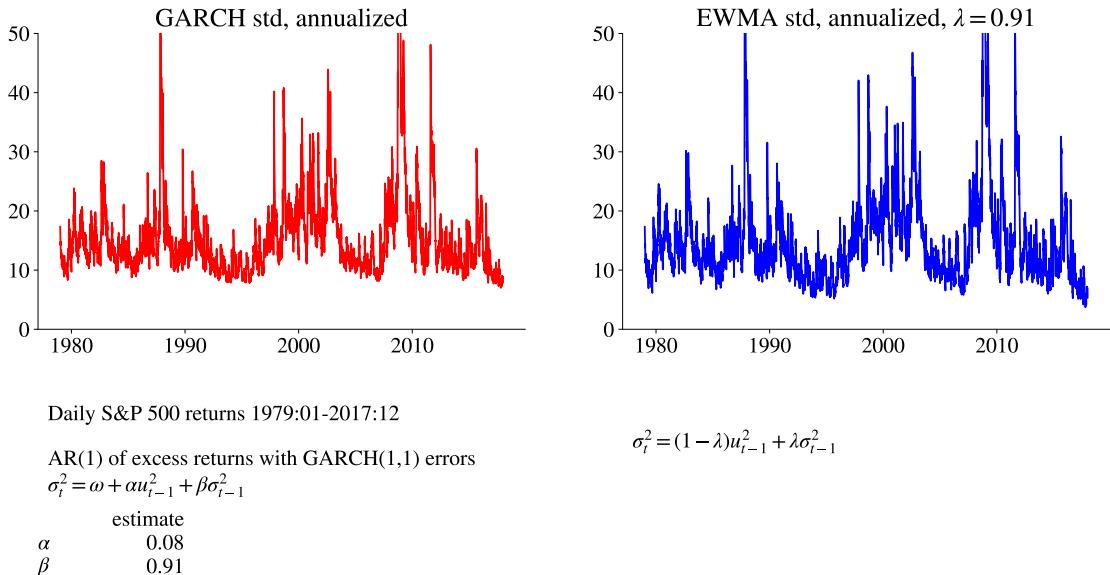


Figure 11.6: Conditional standard deviation, estimated by GARCH(1,1) model

therefore the unconditional variance finite.

Remark 11.3 *The GARCH(1,1) has many similarities with the exponential moving average estimator of volatility (11.3). The main differences are that the exponential moving average does not have a constant and volatility is non-stationary (the coefficients sum to unity).*

See Figure 11.6 for an example.

The GARCH(1,1) corresponds to an ARCH(∞) with geometrically declining weights, which is seen by solving (11.11) recursively by substituting for σ_{t-1}^2 (and then σ_{t-2}^2 , σ_{t-3}^2 , ...)

$$\sigma_t^2 = \frac{\omega}{1 - \beta} + \alpha \sum_{j=0}^{\infty} \beta^j u_{t-1-j}^2. \quad (11.12)$$

This suggests that a GARCH(1,1) might be a reasonable approximation of a high-order ARCH.

Also, the GARCH(1,1) model implies that we in period t can forecast the future conditional variance (σ_{t+s}^2) as

$$E_t \sigma_{t+s}^2 = \bar{\sigma}^2 + (\alpha + \beta)^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1 - \alpha - \beta}, \quad (11.13)$$

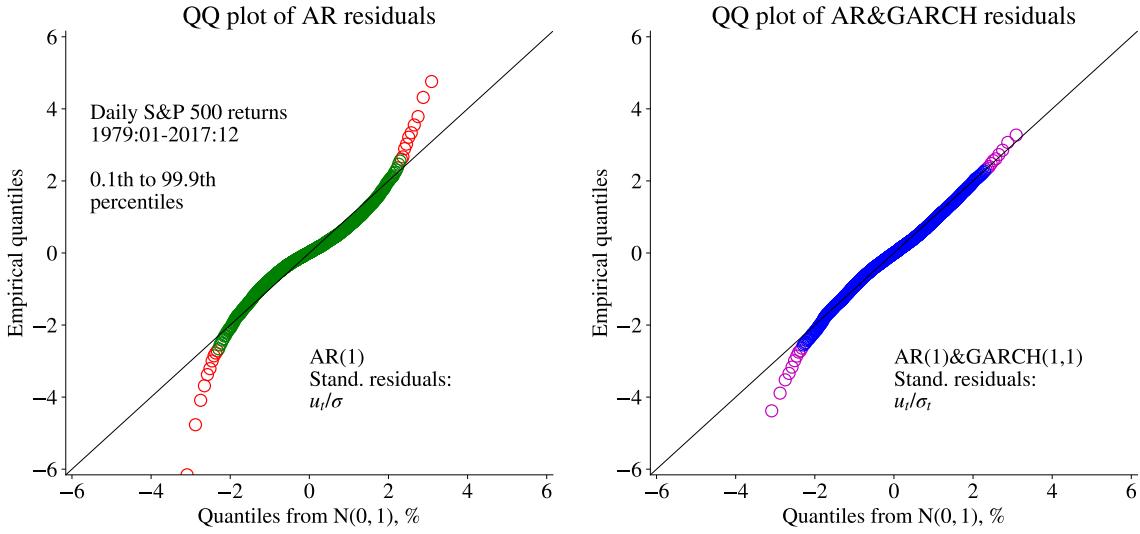


Figure 11.7: QQ-plot of residuals

which is of the same form as for the ARCH model (11.8), but where the sum of α and β is like an AR(1) parameter.

To estimate the model consisting of (11.4) and (11.11) we can still use the likelihood function (11.9) 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 11.4 (*Imposing parameter constraints on GARCH(1,1)*) To impose the restrictions in (11.11), iterate over values of $(b, \tilde{\omega}, \tilde{\alpha}, \tilde{\beta})$ and let $\omega = \omega^2$, $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$, and $\beta = \exp(\tilde{\beta})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$.

See Figure 11.7 for evidence of how the residuals become more normally distributed once the heteroskedasticity is handled.

Remark 11.5 (*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, $\text{VaR}_{0.95} = 0.08$ says that there is only an 5% chance that the loss will be greater than 8% of the investment. See Figure 11.8 for an illustration. When the return has an $N(\mu, \sigma^2)$ distribution, then $\text{VaR}_{95\%} = -(\mu - 1.64\sigma)$. See Figure 11.9 for an example of time-varying VaR, based on a GARCH model.

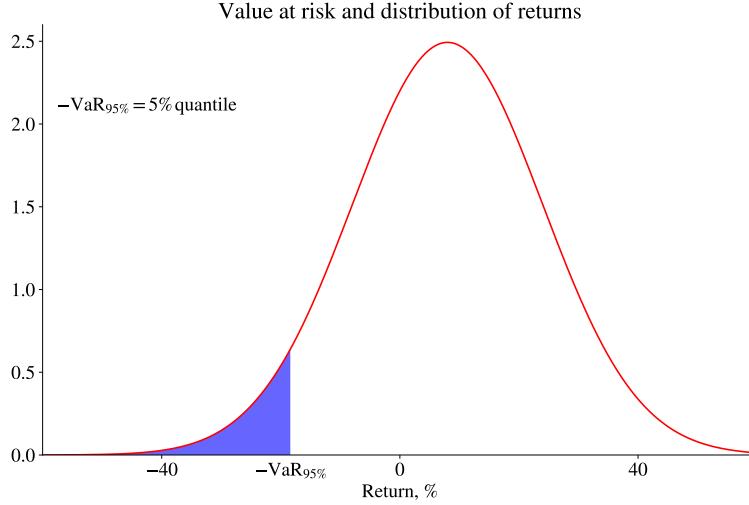


Figure 11.8: Value at risk

11.4 Non-Linear Extensions

A very large number of extensions have been suggested. I summarize a few of them, which can be estimated by MLE using the likelihood function (11.9).

An asymmetric GARCH (Glosten, Jagannathan, and Runkle (1993)) can be constructed as

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \delta(u_{t-1} > 0) u_{t-1}^2, \text{ where} \quad (11.14)$$

$$\delta(q) = \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases}$$

This means that the effect of the shock u_{t-1}^2 is α if the shock was negative and $\alpha + \gamma$ 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 u_{t-1} .

The EGARCH (exponential GARCH, Nelson (1991)) sets

$$\ln \sigma_t^2 = \omega + \alpha \frac{|u_{t-1}|}{\sigma_{t-1}} + \beta \ln \sigma_{t-1}^2 + \gamma \frac{u_{t-1}}{\sigma_{t-1}} \quad (11.15)$$

Apart from being written in terms of the log (which is a smart trick to make $\sigma_t^2 > 0$ hold 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,

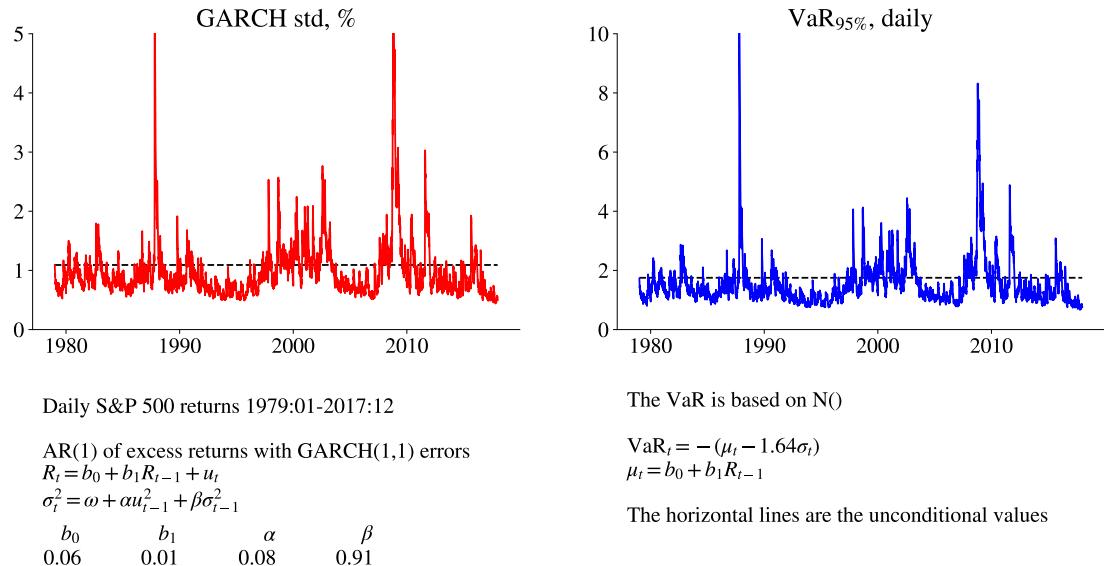


Figure 11.9: Conditional volatility and VaR

if $\gamma < 0$, then the volatility increases more in response to a negative u_{t-1} (“bad news”) than to a positive u_{t-1} . This model is stationary if $|\beta| < 1$.

Hentschel (1995) estimates several models of this type, as well as a very general formulation on daily stock index data for 1926 to 1990 (some 17,000 observations). Most standard models are rejected in favour of a model where σ_t depends on σ_{t-1} and $|u_{t-1} - b|^{3/2}$.

11.5 (G)ARCH-M*

It can make sense to let the conditional volatility enter the regression equation—for instance, as a proxy for risk which may influence the expected return.

We modify the regression (“mean”) equation (11.4) to include the conditional variance σ_t^2 (taken from any of the models for heteroskedasticity) as a regressor

$$y_t = x'_t b + \varphi \sigma_t^2 + u_t. \quad (11.16)$$

Note that σ_t^2 is predetermined, since it is a function of information in $t - 1$. This model can be estimated by using the likelihood function (11.9) to do MLE.

Remark 11.6 (*Coding of (G)ARCH-M*) We can use the same approach as in Remark

[11.1](#), except that we use (11.16) instead of (11.4) to calculate the residuals (and that we obviously also need a guess of φ).

Example 11.7 (*Theoretical motivation of GARCH-M*) An implication from mean-variance portfolio choice and CAPM is that $E(R_m - R_f) = k\sigma_m^2$, which says that the expected excess market return is increasing in both the market volatility (σ_m^2) and risk aversion (k).

11.6 Multivariate (G)ARCH

A first approach to estimate a time varying covariance of two series (u_{it} and u_{jt}) is the EWMA

$$\sigma_{ij,t} = (1 - \lambda)u_{i,t-1}u_{j,t-1} + \lambda\sigma_{ij,t-1}. \quad (11.17)$$

Combining this with similar estimates of the time varying volatility of each series ($\sigma_{i,t}^2$ and $\sigma_{j,t}^2$) allows us to calculate a time varying correlation as

$$\rho_{ij,t} = \sigma_{ij,t}/(\sigma_{i,t}\sigma_{j,t}). \quad (11.18)$$

See Figure 11.10 for an illustration.

11.6.1 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

$$\Sigma_t = E_{t-1} u_t u_t'. \quad (11.19)$$

Remark 11.8 (*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, $vech \begin{bmatrix} \mathbf{a}_{11} & a_{12} \\ \mathbf{a}_{21} & \mathbf{a}_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}$.

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.

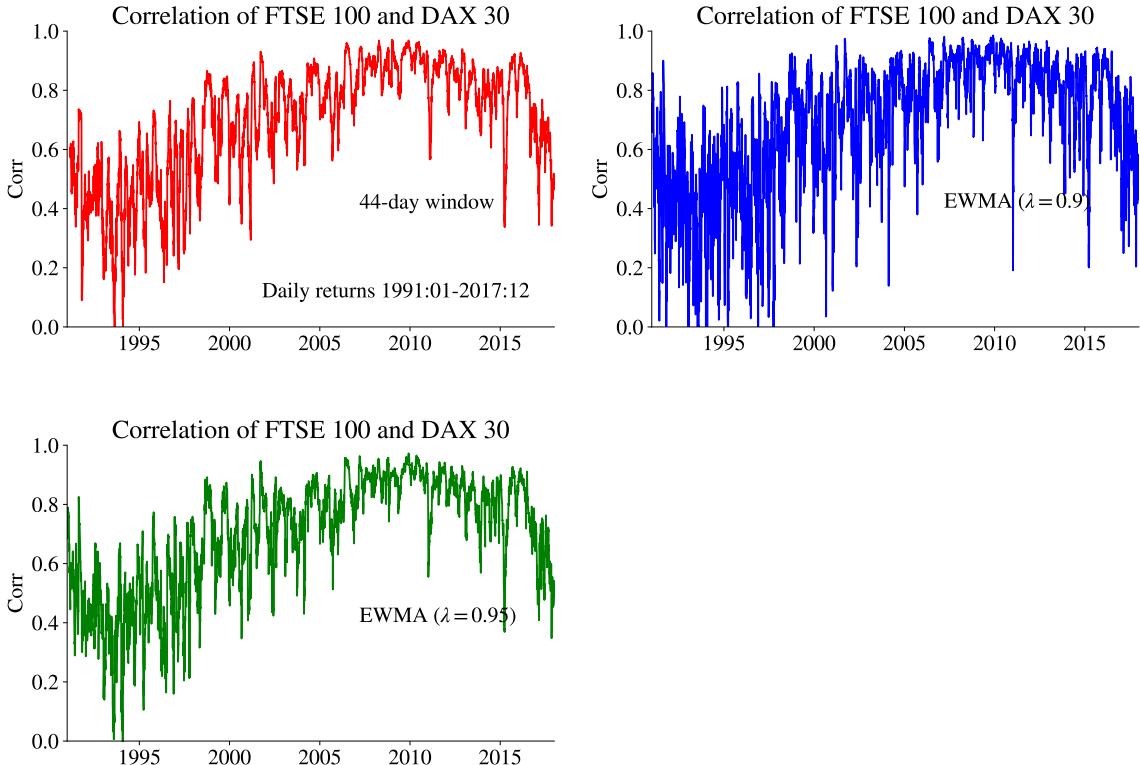


Figure 11.10: Time-varying correlations (different EWMA estimates)

Although we only need to care about the unique elements of Σ_t , that is, $\text{vech}(\Sigma_t)$, this still gives very many parameters

$$\text{vech}(\Sigma_t) = C + A \text{vech}(u_{t-1} u'_{t-1}) + B \text{vech}(\Sigma_{t-1}). \quad (11.20)$$

For instance, with $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = C + A \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1} u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + B \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \quad (11.21)$$

where C is 3×1 , A is 3×3 , and B is 3×3 . This gives 21 parameters, which is already hard to manage. We have to limit the number of parameters. We also have to find a way to impose restrictions so Σ_t is positive definite. Indeed a variance-covariance matrix must be positive definite in order to guarantee that every possible linear combination of the variables has a positive variance (compare the restrictions of positive coefficients in (11.11)). If not, something is wrong.

The Diagonal Model

The *diagonal model* assumes that A and B are diagonal. This means that every element of Σ_t follows a univariate process. With $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1}u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \quad (11.22)$$

which gives $3 + 3 + 3 = 9$ parameters. To make sure that Σ_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

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \quad (11.23)$$

and each of σ_{11t} and σ_{22t} follows a GARCH process. Assuming a GARCH(1,1) as in (11.11) gives 7 parameters (2×3 GARCH parameters and one correlation), which is convenient. The price is, of course, the assumption of no movements in the correlations. To get a positive definite Σ_t , each individual GARCH model must generate a positive variance (same restrictions as before), and that all the estimated (constant) correlations are between -1 and 1 .

Remark 11.9 (*Estimating the constant correlation model*) A quick (and dirty) method is to first estimate the individual 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.

See Figure 11.11 for an illustration and Figure 11.10 for a comparison with the EWMA approach.

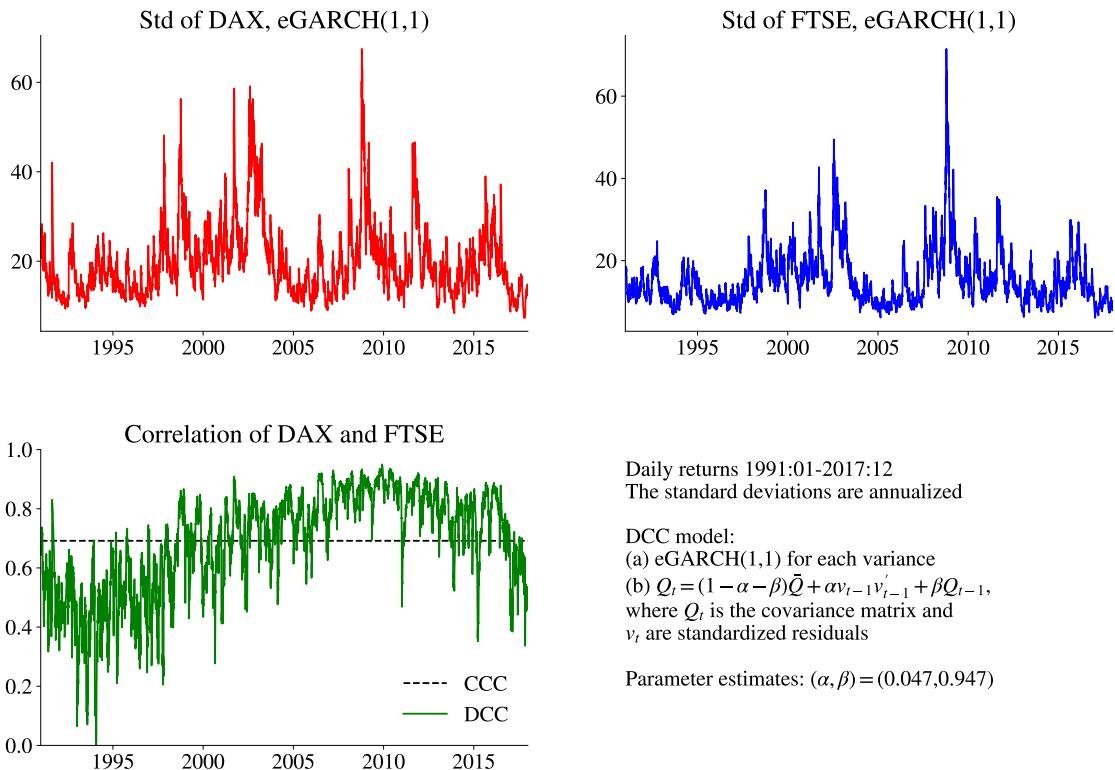


Figure 11.11: Results for multivariate eGARCH models

11.6.2 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 u_t is iid $N(0, \Sigma_t)$, then the log likelihood function is

$$\ln \mathcal{L} = -\frac{Tn}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^T \ln |\Sigma_t| - \frac{1}{2} \sum_{t=1}^T u_t' \Sigma_t^{-1} u_t. \quad (11.24)$$

In practice, the optimization problem can be difficult since there are typically many parameters. At least, good starting values are required.

Chapter 12

Risk Measures

Reference: Hull (2006) 18; McDonald (2006) 25; Fabozzi, Focardi, and Kolm (2006) 4–5; McNeil, Frey, and Embrechts (2005); Alexander (2008b)

12.1 Value at Risk

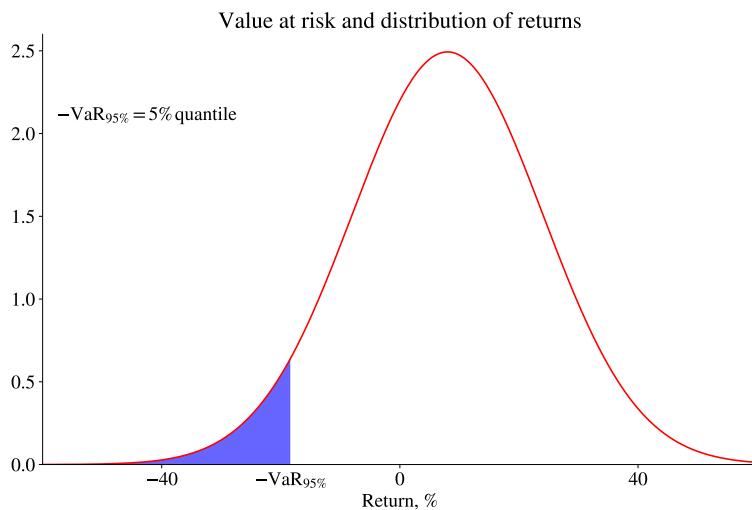


Figure 12.1: Value at risk

The mean-variance framework is often criticized for failing to distinguish between downside of the return distribution (considered to be risk) and upside (considered to be potential). The Value at Risk is one way of focusing on the downside.

Remark 12.1 (*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 ($-R$) is larger than $\text{VaR}_{95\%}$

$$\Pr(-R \geq \text{VaR}_{95\%}) = 5\%. \quad (12.1)$$

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 value at risk into value terms (CHF, say), just multiply the VaR for returns with the value of the investment (portfolio).

Clearly, $-R \geq \text{VaR}_{95\%}$ is true when (and only when) $R \leq -\text{VaR}_{95\%}$, so (12.1) can also be expressed as

$$\Pr(R \leq -\text{VaR}_{95\%}) = 5\%. \quad (12.2)$$

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 (12.2) allows us to work directly with the return distribution (not the loss distribution), which is often convenient. See Figure 12.1 for an illustration. If the return is normally distributed, $R \sim N(\mu, \sigma^2)$ then

$$\text{VaR}_{95\%} = -(\mu - 1.64\sigma). \quad (12.3)$$

Example 12.2 (*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, $\text{VaR}_{95\%} = 0.18$.

More generally, we can consider the confidence level α instead of just 0.95, so

$$\Pr(R \leq -\text{VaR}_\alpha) = 1 - \alpha, \text{ so} \quad (12.4)$$

$$\text{VaR}_\alpha = -(1 - \alpha)^{th} \text{ quantile of } R. \quad (12.5)$$

If the return is normally distributed, $R \sim N(\mu, \sigma^2)$, then

$$\text{VaR}_\alpha = -(\mu + c\sigma), \quad (12.6)$$

where c is the $(1 - \alpha)^{th}$ for a $N(0, 1)$ distribution, for instance, -1.64 for $1 - \alpha = 0.05$.

This is illustrated in Figure 12.2.

Remark 12.3 (*Critical values of $N(\mu, \sigma^2)$*) If $R \sim N(\mu, \sigma^2)$, then there is a 5% probability that $R \leq \mu - 1.64\sigma$, a 2.5% probability that $R \leq \mu - 1.96\sigma$, and a 1% probability that $R \leq \mu - 2.33\sigma$.

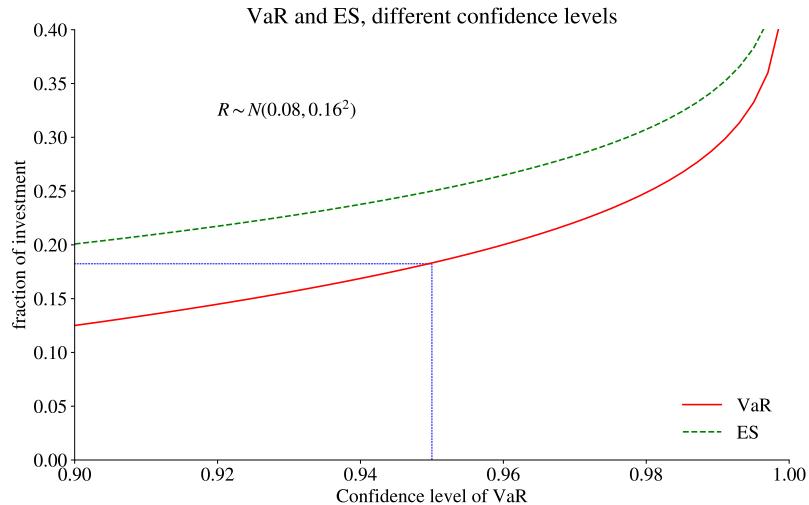


Figure 12.2: Value at risk, different probability levels

Example 12.4 (*VaR with $R \sim N(\mu, \sigma^2)$*) If $R \sim N(\mu, \sigma^2)$ with $\mu = 8\%$ and $\sigma = 16\%$, then $VaR_{97.5\%} = -(0.08 - 1.96 \times 0.16) \approx 0.24$.

Figure 12.3 shows the distribution and VaRs (for different probability levels) for the daily S&P 500 returns. Two different types of VaRs are shown: (i) based on a normal distribution and (ii) as the empirical VaR (from the empirical quantiles of the distribution).

Example 12.5 (*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 stated investment period (for instance, one day). Multi-period VaRs are calculated by either explicitly constructing the distribution of multi-period returns, or by making simplifying assumptions about the relation between returns in different periods (for instance, that they are iid).

Remark 12.6 (*Multi-period VaR*) If the returns are iid, 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.

Figure 12.4 shows VaR calculated from $N(\mu_t, \sigma_t^2)$, but where the μ_t, σ_t values are allowed to change over time.

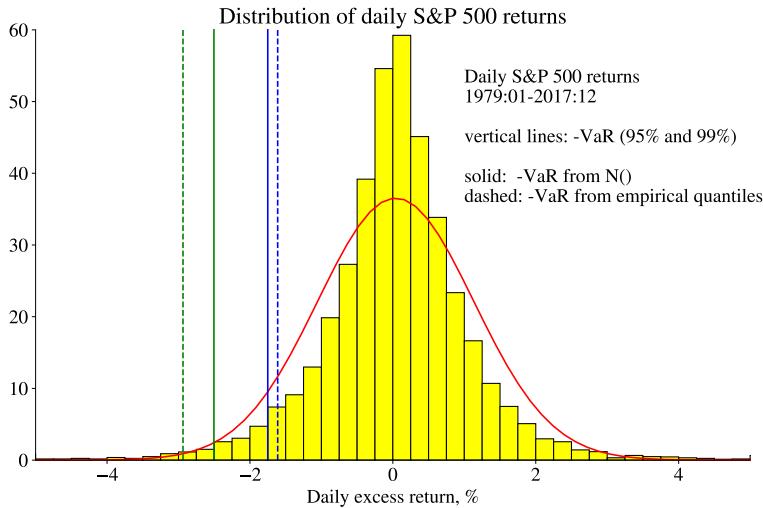


Figure 12.3: Return distribution and VaR for S&P 500

12.2 Backtesting a VaR Model

Backtesting a VaR model amounts to checking if (historical) data fits with the VaR numbers. For instance, we first find the $\text{VaR}_{95\%}$ and then calculate what fraction of returns that is actually below (the negative of) 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 (the negative of) this number. Figures 12.5–12.6 show results from backtesting a VaR model where the volatility follows a GARCH process (to capture the time varying volatility). The evidence suggests that this model, combined with the assumption that the return is normally distributed (but with time-varying parameters), works relatively well except for very high confidence levels.

It is also important to see if there are medium- to long-run deviations from the VaR confidence level. Figure 12.7 illustrates the importance of using a dynamic VaR to capture the swings in uncertainty.

Remark 12.7 (*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 $\text{Var}(y) = np(1-p)$.

To perform a statistical back test of a VaR model, define a variable that is one if the

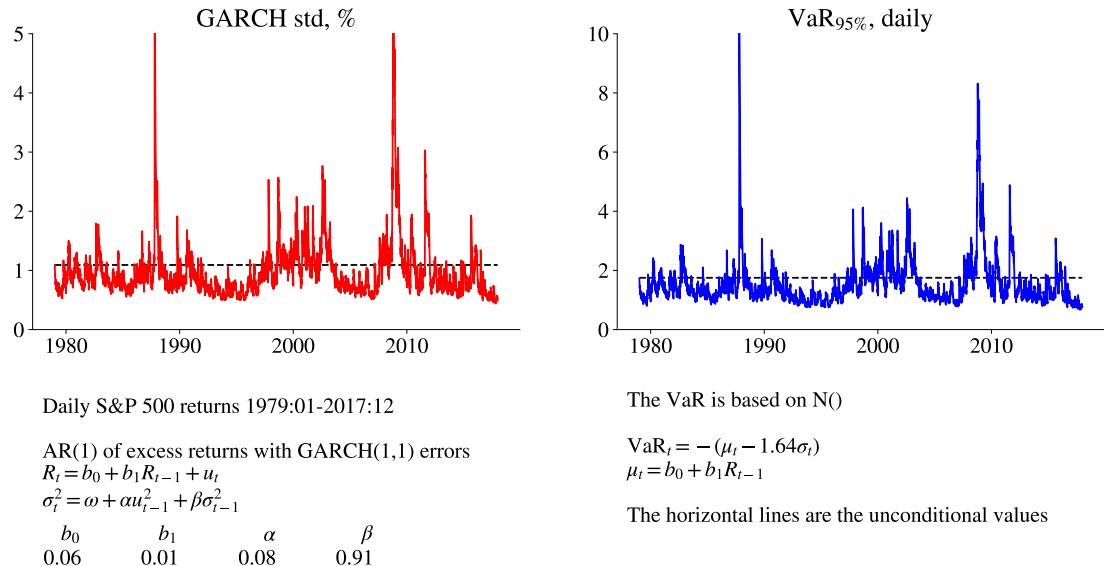


Figure 12.4: Conditional volatility and VaR

loss is greater than the VaR

$$d_t = \begin{cases} 1 & \text{if } R_t < -\text{VaR}_\alpha \\ 0 & \text{otherwise} \end{cases} \quad (12.7)$$

By using the properties of a binomial distribution, we can then form a test statistic for the null hypothesis that we have the number $(1 - \alpha)T$ of dates when the return is lower than $-\text{VaR}_\alpha$ (that is, the loss is larger than VaR_α)

$$\frac{\sum_{t=1}^T d_t - (1 - \alpha)T}{\sqrt{T\alpha(1 - \alpha)}} \xrightarrow{d} N(0, 1). \quad (12.8)$$

As indicated, the test statistic can be compared with a standard normal, provided our sample is large.

12.3 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 notion of diversification benefits. (To get this unfortunate property, the return distributions must be heavily skewed.) The expected shortfall has better

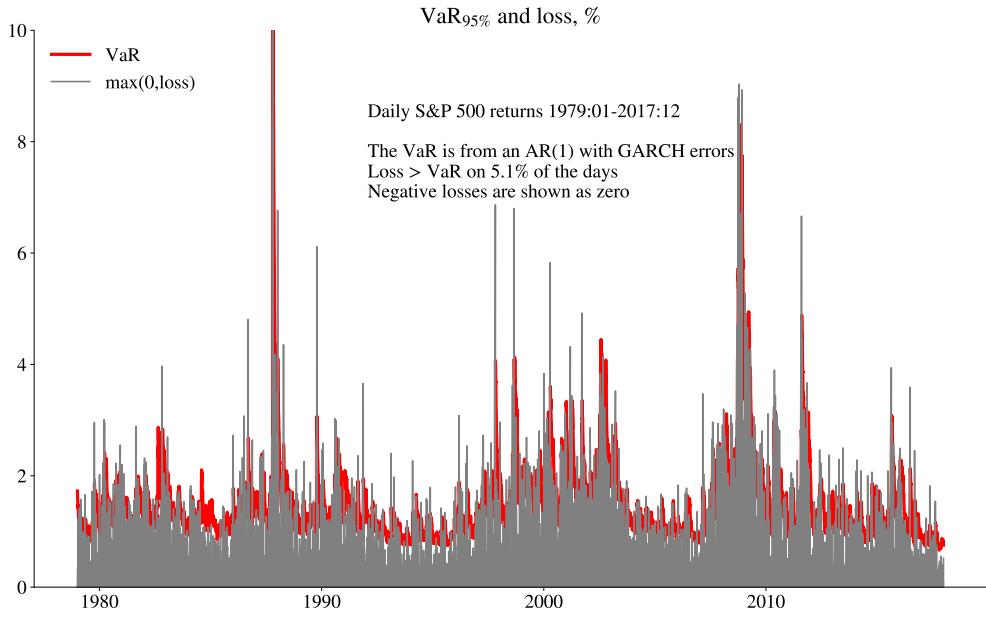


Figure 12.5: Backtesting VaR from a GARCH model, assuming normally distributed shocks

aggregation properties.

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,

$$\text{ES}_\alpha = -\mathbb{E}(R|R \leq -\text{VaR}_\alpha). \quad (12.9)$$

This might be more informative than the VaR_α , which is the *minimum loss* that will happen with a $1 - \alpha$ probability. See Figure 12.8 for an illustration.

For a normally distributed return $R \sim N(\mu, \sigma^2)$ we have

$$\text{ES}_\alpha = -\mu + \sigma \frac{\phi(c_{1-\alpha})}{1 - \alpha}, \quad (12.10)$$

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 $1 - \alpha = 0.05$).

Proof. (of (12.10)) If $x \sim N(\mu, \sigma^2)$, then $\mathbb{E}(x|x \leq b) = \mu - \sigma\phi(b_0)/\Phi(b_0)$ 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 . ■

Example 12.8 (ES) If $\mu = 8\%$ and $\sigma = 16\%$, the 95% expected shortfall is $\text{ES}_{95\%} =$

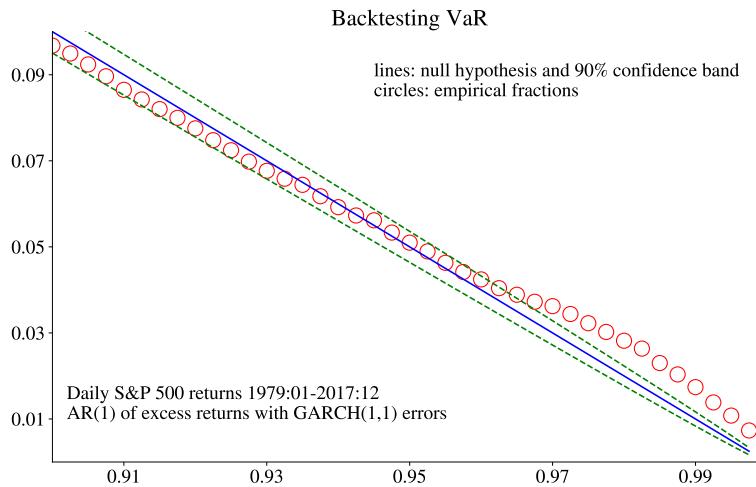


Figure 12.6: Backtesting VaR from a GARCH model, assuming normally distributed shocks

$$-0.08 + 0.16\phi(-1.64)/0.05 \approx 0.25 \text{ and the 97.5\% expected shortfall is } ES_{97.5\%} = \\ -0.08 + 0.16\phi(-1.96)/0.025 \approx 0.29.$$

Notice that the expected shortfall for a normally distributed return (12.10) is a strictly increasing function of the standard deviation (and the variance). Minimizing the expected shortfall at a given mean return therefore gives the same solution (portfolio weights) as minimizing the variance at the same given mean return. In other cases when returns are not normally distributed, the portfolio choice will be different (and perhaps complicated to perform).

See Table 12.1 for an empirical comparison of the VaR, ES and some more downside risk measures (discussed below).

	Small growth	Large value
Std	7.8	5.3
VaR (95%)	11.7	8.2
ES (95%)	16.8	11.5
SemiStd	5.4	3.5
Drawdown	79.4	59.3

Table 12.1: Risk measures of monthly returns of two stock indices (%), US data 1957:01-2017:12.

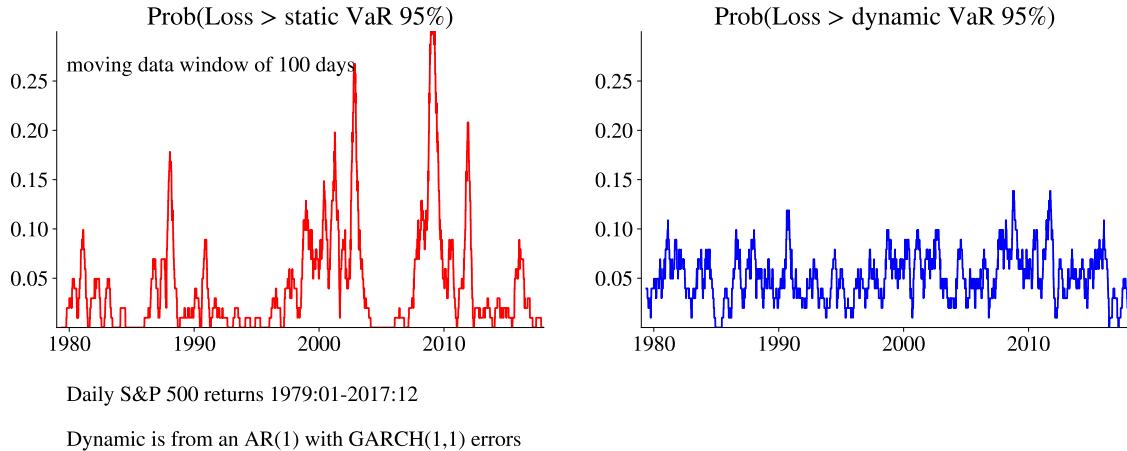


Figure 12.7: Backtesting VaR from a GARCH model on a moving data window, assuming normally distributed shocks

12.4 Semivariance and Max Drawdown

Reference: Bawa and Lindenberg (1977) and Nantell and Price (1979)

A semivariance is defined as

$$\lambda^2 = E[\min(R - E R, 0)^2]. \quad (12.11)$$

The square root of λ^2 is called the semi-standard deviation. In comparison with a variance only negative deviations from the mean are given any weight.

To estimate the target semivariance from data use

$$\lambda^2 = \frac{1}{T} \sum_{t=1}^T \delta(R_t \leq h)(R_t - \bar{R})^2, \quad (12.12)$$

where $\delta(q) = 1$ if q is true and zero otherwise.

An alternative measure is the (percentage) *maximum drawdown* over a given horizon, for instance, 5 years, say. This is the largest loss from peak to bottom within the given horizon—see Figure 12.10. This is a useful measure when the investor do not know exactly when he/she has to exit the investment—since it indicates the worst (peak to bottom) outcome over the sample.

See Figures 12.11–12.12 for an illustration of max drawdown.

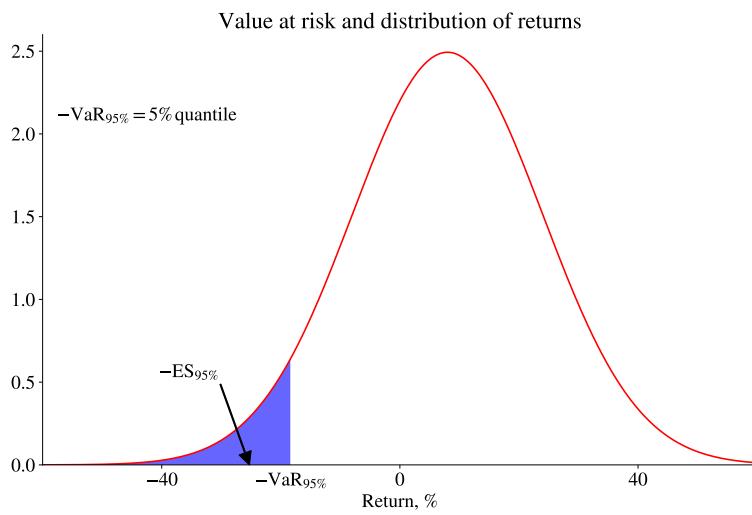


Figure 12.8: Value at risk and expected shortfall

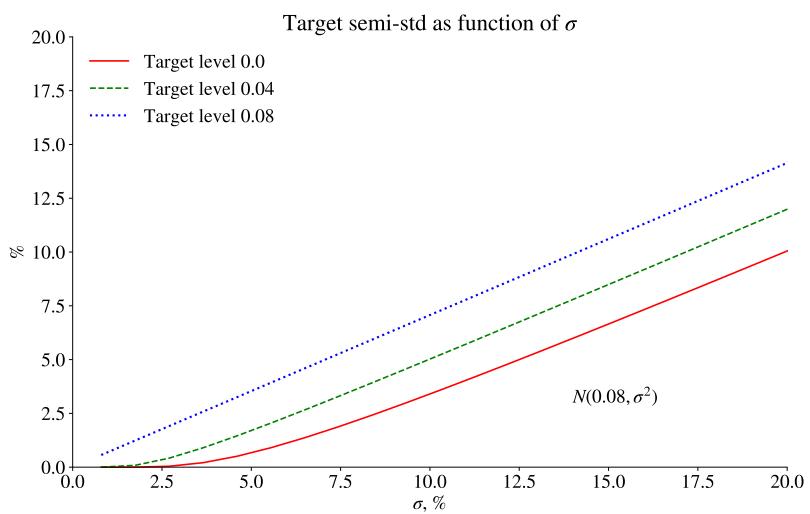


Figure 12.9: Target semivariance for a $N(\mu, \sigma^2)$ variable

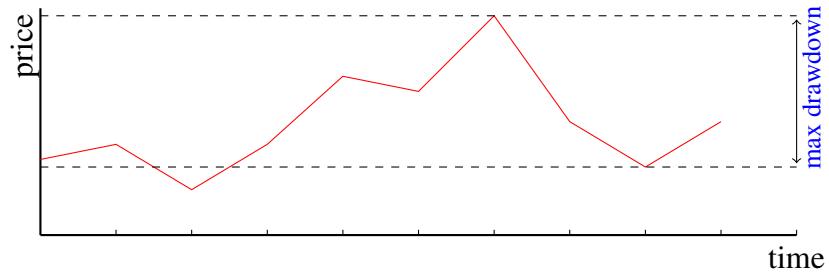


Figure 12.10: Max drawdown

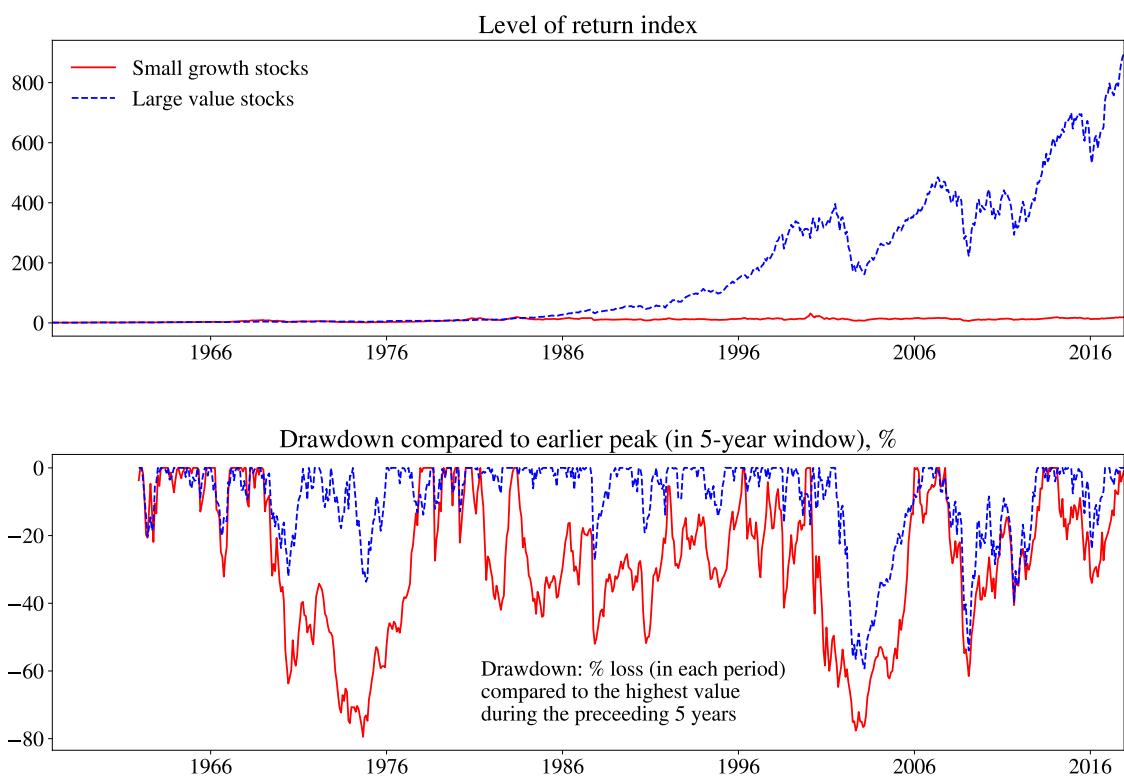


Figure 12.11: Drawdown

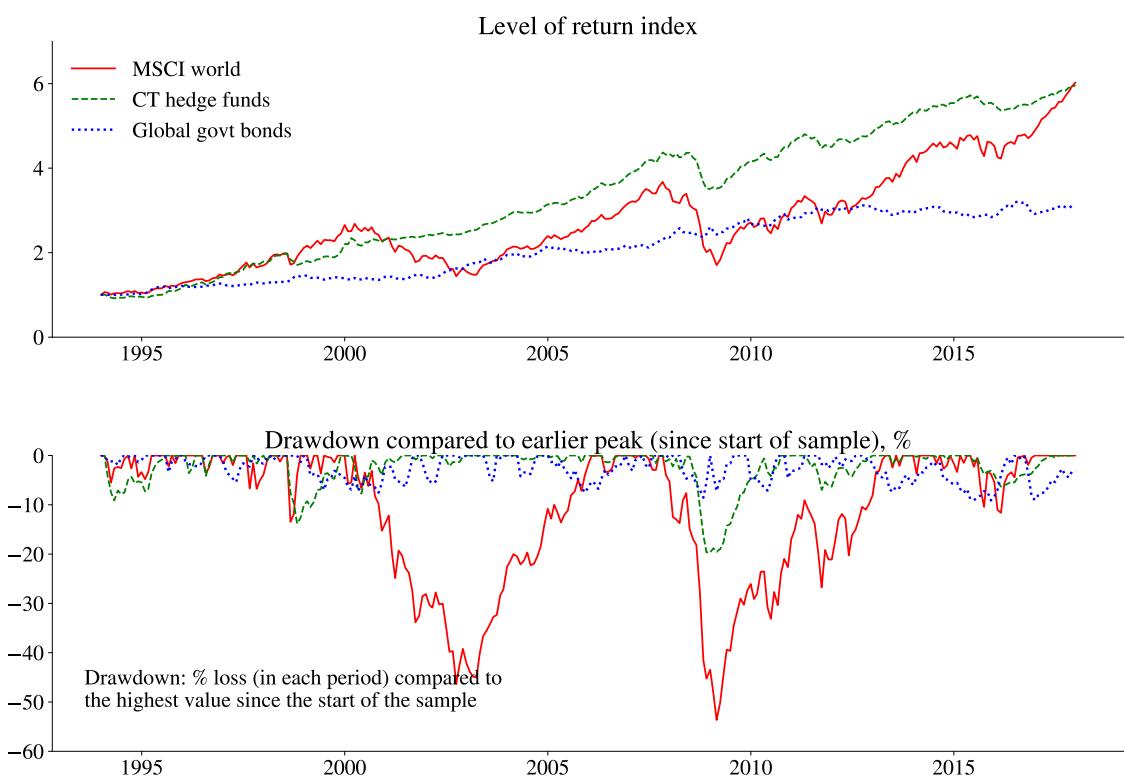


Figure 12.12: Drawdown

Chapter 13

Return Distributions (Univariate)

Sections denoted by a star (*) is not required reading.

13.1 Estimating and Testing Distributions

Reference: Harvey (1989) 260, Davidson and MacKinnon (1993) 267, Silverman (1986); Mittelhammer (1996), DeGroot (1986)

13.1.1 A Quick Recap of a Univariate Distribution

The cdf (cumulative distribution function) measures the probability that the random variable X_i is below or at some numerical value x_i ,

$$F_i(x_i) = \Pr(X_i \leq x_i). \quad (13.1)$$

For instance, with an $N(0, 1)$ distribution, $F(-1.64) = 0.05$. Clearly, the cdf values are between (and including) 0 and 1. The distribution of X_i is often called the *marginal distribution* of X_i —to distinguish it from the joint distribution of X_i and X_j . (See below for more information on joint distributions.)

The pdf (probability density function) $f_i(x_i)$ is the “height” of the distribution in the sense that the cdf $F(x_i)$ is the integral of the pdf from minus infinity to x_i

$$F_i(x_i) = \int_{s=-\infty}^{x_i} f_i(s)ds. \quad (13.2)$$

(This clearly means that the pdf is the derivative of the cdf, $f_i(x_i) = \partial F_i(x_i)/\partial x_i$.) A pdf must be non-negative, $f_i(s) \geq 0$. The Gaussian pdf (the normal distribution) is bell shaped.

Remark 13.1 (*Quantile of a distribution*) The α quantile of a distribution (ξ_α) is the value of x such that there is a probability of α of a lower value. We can solve for the quantile by inverting the cdf, $\alpha = F(\xi_\alpha)$ as $\xi_\alpha = F^{-1}(\alpha)$. For instance, the 5% quantile of a $N(0, 1)$ distribution is $-1.64 = \Phi^{-1}(0.05)$, where $\Phi^{-1}()$ denotes the inverse of an $N(0, 1)$ cdf. See Figure 13.1 for an illustration.

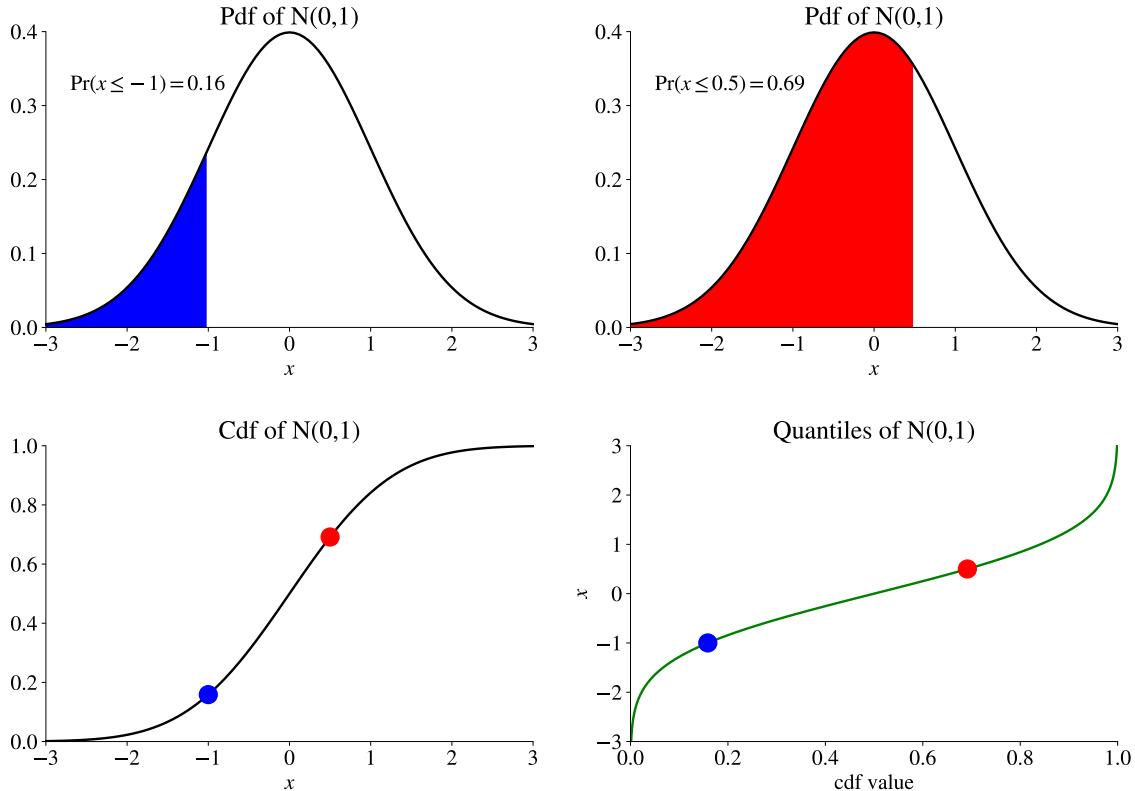


Figure 13.1: Finding quantiles of a $N(\mu, \sigma^2)$ distribution

13.1.2 QQ Plots

Are returns normally distributed? Mostly not, but it depends on the asset type and on the return horizon. Options returns typically have very non-normal distributions (in particular, since the return is -100% on many expiration days). Stock returns are typically distinctly non-linear at short horizons, but can look somewhat normal at longer horizons.

To assess the normality of returns, the usual econometric techniques (Bera–Jarque and Kolmogorov-Smirnov tests) are useful, but a visual inspection of the histogram and a QQ-plot also give useful clues. See Figures 13.2–13.5 for illustrations.

Remark 13.2 (*Reading a QQ plot*) 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. The *QQ plot* is a scatter plot where each point shows a specific percentile (quantile) according to the empirical as well as according to the theoretical distribution. For instance, if the 2th percentile (0.02 percentile) 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.

There is one important caveat to this way of studying data: it only provides evidence on the unconditional distribution. Suppose instead that we have estimated a model for time-variation in the mean and variance (denoted μ_t and σ_t^2 , respectively), then it makes more sense to study the distribution (QQ plot) of the standardised return

$$\tilde{R}_t = \frac{R_t - \mu_t}{\sigma_t}. \quad (13.3)$$

As a simple example, the mean could be estimated by an AR(1) model (so we would have $\mu_t = a + \rho R_{t-1}$) and the variance by a GARCH model (so we would have $\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2$ where u_{t-1} is the surprise to the return in $t-1$).

13.1.3 Parametric Tests of Normal Distribution

The skewness, kurtosis and Bera-Jarque test for normality are useful diagnostic tools. They are

	Test statistic	Distribution	
skewness	$= \frac{1}{T} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)^3$	$N(0, 6/T)$	(13.4)
kurtosis	$= \frac{1}{T} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)^4$	$N(3, 24/T)$	
Bera-Jarque	$= \frac{T}{6} \text{skewness}^2 + \frac{T}{24} (\text{kurtosis} - 3)^2$	χ_2^2	

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (13.4) 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 intuition for the χ_2^2 distribution of the Bera-Jarque 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 Bera-Jarque test statistic is therefore a sum of the square of two uncorrelated $N(0, 1)$ variables, which has a χ_2^2 distribution.

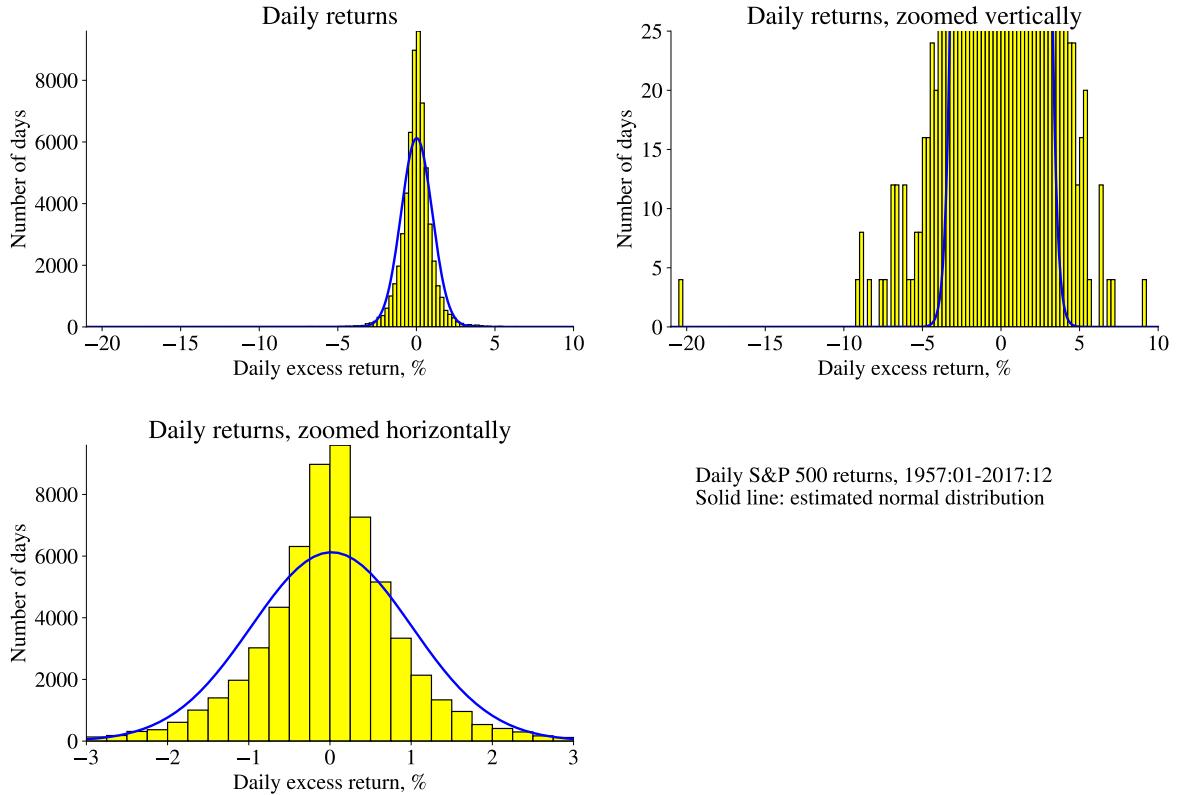


Figure 13.2: Distribution of daily S&P returns

13.1.4 Nonparametric Tests of General Distributions

The *Kolmogorov-Smirnov* test is designed to test if an empirical distribution function, $\text{EDF}(x)$, conforms with a theoretical cdf, $F(x)$. The empirical distribution function is defined as the fraction of observations which are less or equal to x , that is,

$$\text{EDF}(x) = \frac{1}{T} \sum_{t=1}^T \delta(x_t \leq x), \text{ where} \quad (13.5)$$

$$\delta(q) = \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases}$$

The $\text{EDF}(x_t)$ and $F(x_t)$ are often plotted against the sorted (in ascending order) sample $\{x_t\}_{t=1}^T$. See Figure 13.6 for an illustration.

Example 13.3 (EDF) Suppose we have a sample with three data points: $[x_1, x_2, x_3] = [5, 3.5, 4]$. The empirical distribution function is then as in Figure 13.6.

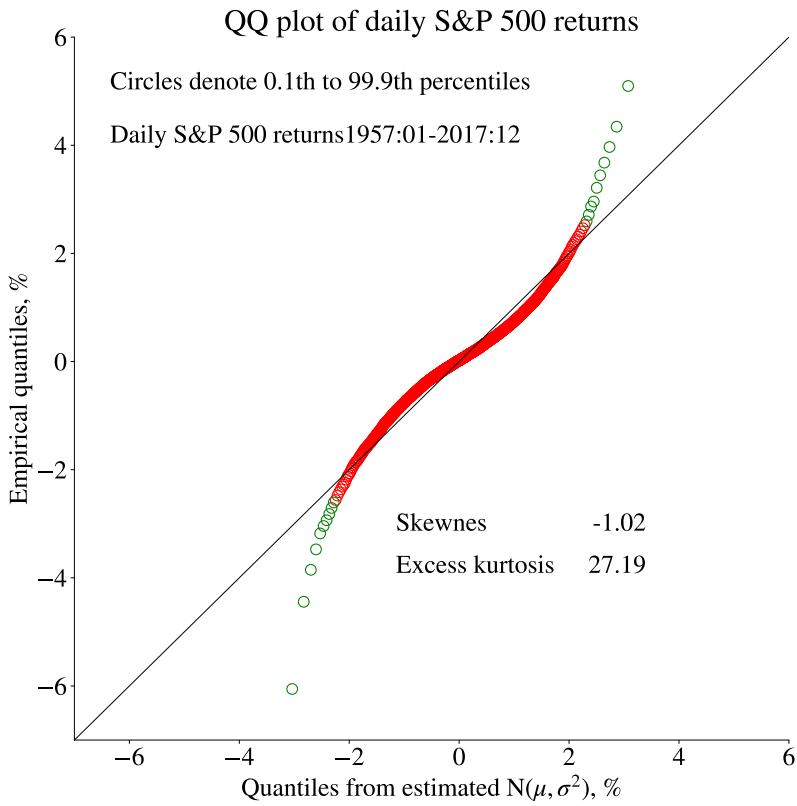


Figure 13.3: Quantiles of daily S&P returns

Define the absolute value of the maximum distance

$$D_T = \max_{x_t} |\text{EDF}(x_t) - F(x_t)|. \quad (13.6)$$

We reject the null hypothesis that $\text{EDF}(x) = F(x)$ if $\sqrt{T} D_t > c$, where c is a critical value ((1.36, 1.48, 1.63) are the critical values on the (5%, 2.5%, 1%) significance levels). There is a corresponding test for comparing two empirical cdfs. See Figure 13.7 for an illustration.

Example 13.4 (Kolmogorov-Smirnov test statistic) Figure 13.6 also shows the cumulative distribution function (cdf) of a normally distributed variable. The test statistic (13.6) is then the largest difference (in absolute terms) of the EDF and the cdf—among the observed values of x_t . See Figure 13.7 for an illustration.

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

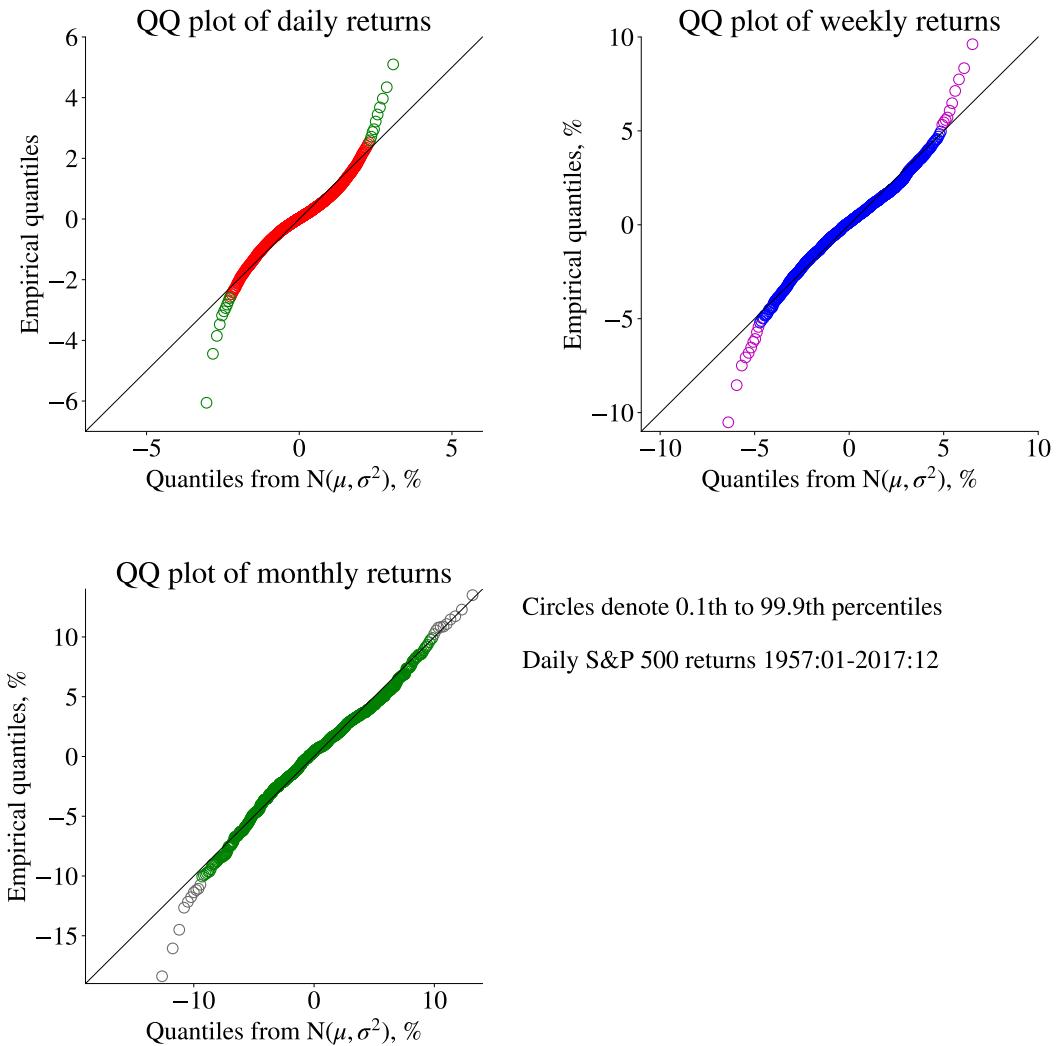


Figure 13.4: Distribution of S&P returns (different horizons)

predicts that the fraction p_i should be in category i , with $\sum_{i=1}^K p_i = 1$. Then

$$\sum_{i=1}^K \frac{(N_i - Tp_i)^2}{Tp_i} \sim \chi_{K-1}^2. \quad (13.7)$$

There is a corresponding test for comparing two empirical distributions.

13.1.5 Kernel Density Estimation

Reference: Silverman (1986)

A histogram is just a count of the relative number of observations that fall in (pre-

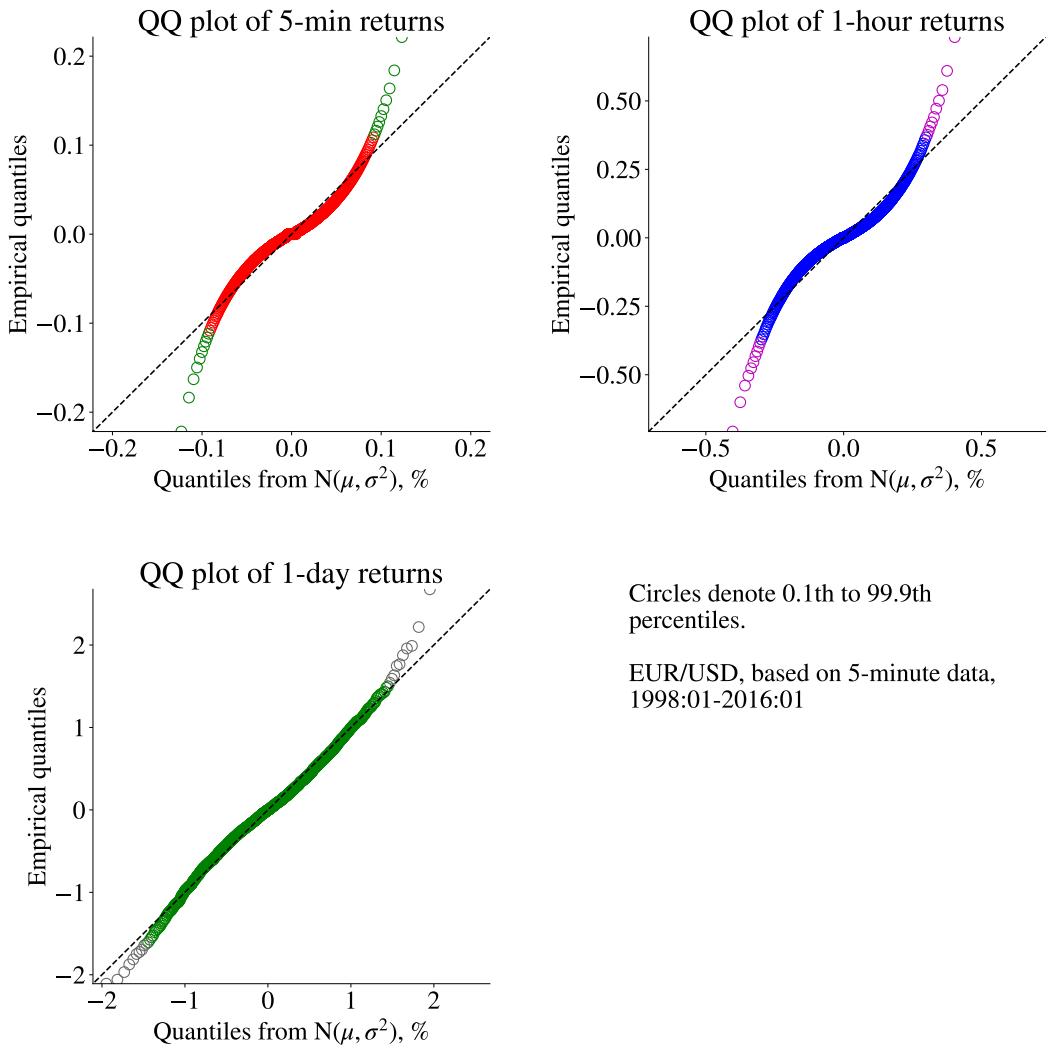


Figure 13.5: Distribution of exchange rate returns (different horizons)

specified) non-overlapping intervals. If we also divide by the width of the interval, then the area under the histogram is unity, so this scaled histogram can be interpreted as a density function. The scaled histogram at the point x (say, $x = 2.3$) can actually be defined as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{h} \delta\left(\left|\frac{x_t - x}{h}\right| \leq 1/2\right), \text{ where} \quad (13.8)$$

$$\delta(q) = \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases}$$

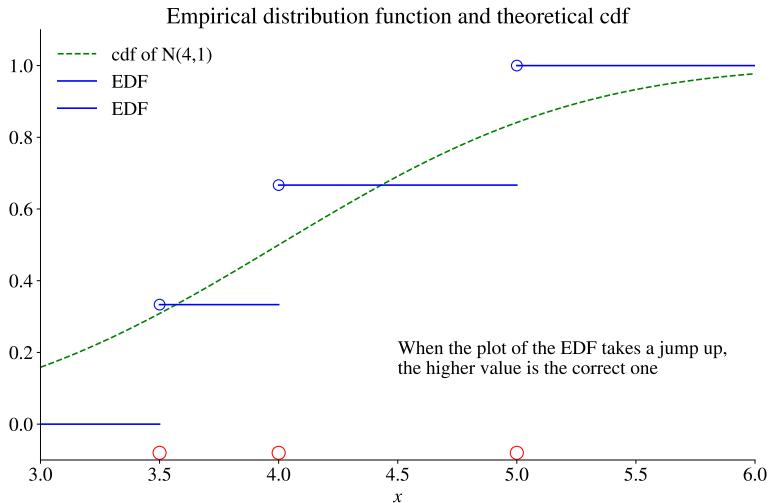


Figure 13.6: Example of empirical distribution function

In this case, the intervals (“bins”) are h wide around a point x : $x - h/2$ to $x + h/2$.

In fact, that $\frac{1}{h}\delta(|x_t - x| \leq h/2)$ 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 (here: the histogram) can be thought of as a average of hypothetical pdf values of the data in the neighbourhood of x .

However, we can gain efficiency and get a smoother (across x values) 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 (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

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - x}{h}\right)^2\right]. \quad (13.9)$$

Notice that the function in the summation is the density function of a $N(x, h^2)$ distribution. See Figure 17.1 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. See Figure 13.9 for an example

It can be shown that (with iid data and a Gaussian kernel) the asymptotic distribution

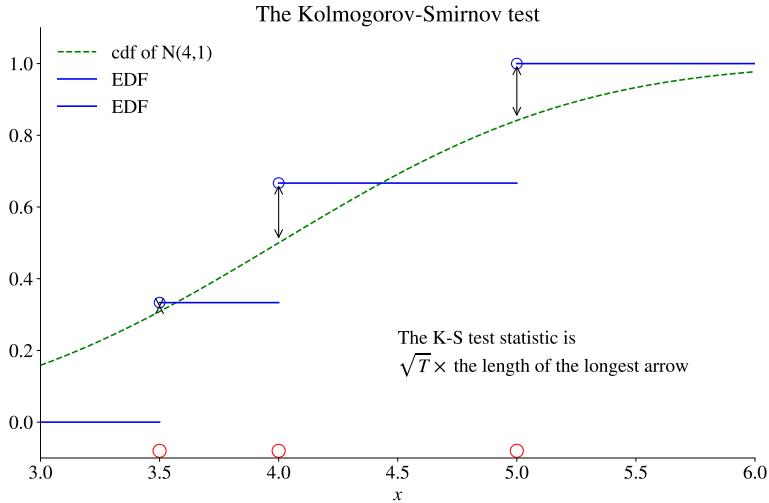


Figure 13.7: K-S test

is

$$\sqrt{Th}[\hat{f}(x) - f(x)] \xrightarrow{d} N\left[0, \frac{1}{2\sqrt{\pi}}f(x)\right]. \quad (13.10)$$

We can also estimate multivariate pdfs. Let x_t be a $d \times 1$ matrix and $\hat{\Omega}$ be the estimated covariance matrix of x_t . We can then estimate the pdf at a point x by using a multivariate Gaussian kernel as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{(2\pi)^{d/2} |H^2 \hat{\Omega}|^{1/2}} \exp\left[-\frac{1}{2}(x_t - x)'(H^2 \hat{\Omega})^{-1}(x_t - x)\right]. \quad (13.11)$$

Notice that the function in the summation is the (multivariate) density function of a $N(x, H^2 \hat{\Omega})$ distribution. The value $H = 1.06T^{-1/(d+4)}$ is sometimes recommended.

Remark 13.5 ((13.11) with $d = 1$) With just one variable, (13.11) becomes

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{H \text{Std}(x_t) \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - x}{H \text{Std}(x_t)}\right)^2\right],$$

which is the same as (13.9) if $h = H \text{Std}(x_t)$.

13.1.6 “Foundations of Technical Analysis...” by Lo, Mamaysky and Wang (2000)

Reference: Lo, Mamaysky, and Wang (2000)

Topic: is the distribution of the return different after a “signal?” This paper uses kernel regressions to identify and implement some technical trading rules, and then tests if the

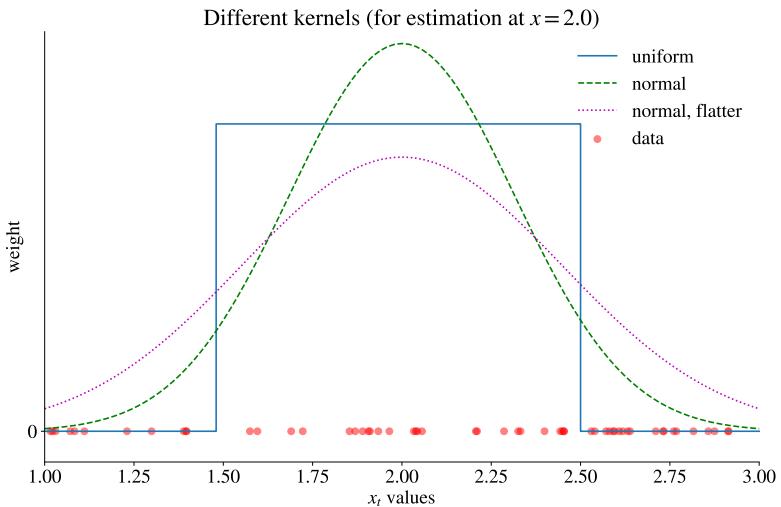


Figure 13.8: Different weighting functions for kernel density estimate

distribution (of the return) after a signal is the same as the unconditional distribution (using Pearson's χ^2 test and the Kolmogorov-Smirnov test). They reject that hypothesis in many cases, using daily data (1962–1996) for around 50 (randomly selected) stocks.

See Figures 13.11–13.12 for an illustration.

13.2 Tail Distribution*

Reference: McNeil, Frey, and Embrechts (2005) 7, Alexander (2008b) 3

In risk control, the focus is on the distribution of losses beyond some threshold level. This has three direct implications. First, the object under study is the loss

$$X = -R, \quad (13.12)$$

that is, the negative of the return. Second, the attention is on how the distribution looks like beyond a threshold and also on the probability of exceeding this threshold. In contrast, the exact shape of the distribution below that point is typically disregarded. Third, modelling the tail of the distribution is best done by using a distribution that allows for a much heavier tail than suggested by a normal distribution. The generalized Pareto (GP) distribution is often used.

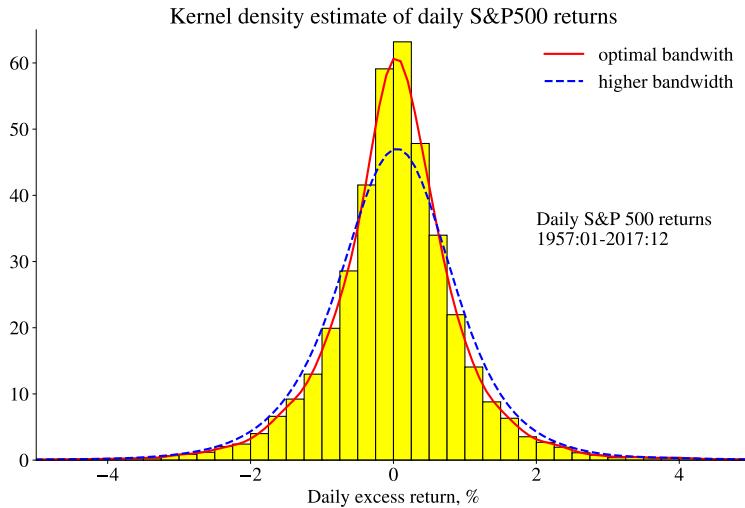


Figure 13.9: Stock returns

13.2.1 Loss Distribution and the Generalized Pareto Distribution

The generalized Pareto (GP) distribution is often used to model the tail of the loss distribution. See Figure 14.2 for an illustration.

Remark 13.6 (*Cdf and pdf of the generalized Pareto distribution*) The generalized Pareto distribution is described by a scale parameter ($\beta > 0$) and a shape parameter (ξ). The cdf ($\Pr(Z \leq z)$, where Z is the random variable and z is a value) is

$$G(z) = \begin{cases} 1 - (1 + \xi z / \beta)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-z / \beta) & \xi = 0, \end{cases}$$

for $0 \leq z$ if $\xi \geq 0$ and $0 \leq z \leq -\beta/\xi$ in case $\xi < 0$. The pdf is

$$g(z) = \begin{cases} \frac{1}{\beta} (1 + \xi z / \beta)^{-1/\xi - 1} & \text{if } \xi \neq 0 \\ \frac{1}{\beta} \exp(-z / \beta) & \xi = 0. \end{cases}$$

The mean is defined (finite) if $\xi < 1$ and is then $E(Z) = \beta / (1 - \xi)$. Similarly, the variance is finite if $\xi < 1/2$ and is then $\text{Var}(Z) = \beta^2 / [(1 - \xi)^2 (1 - 2\xi)]$. See Figure 14.3 for an illustration.

Remark 13.7 (*The location of the generalized Pareto distribution) In the application below we will use $z = x - u$ where x is a random variable and u a threshold level. In some texts this u is part of the definition of the distribution, and is often called the

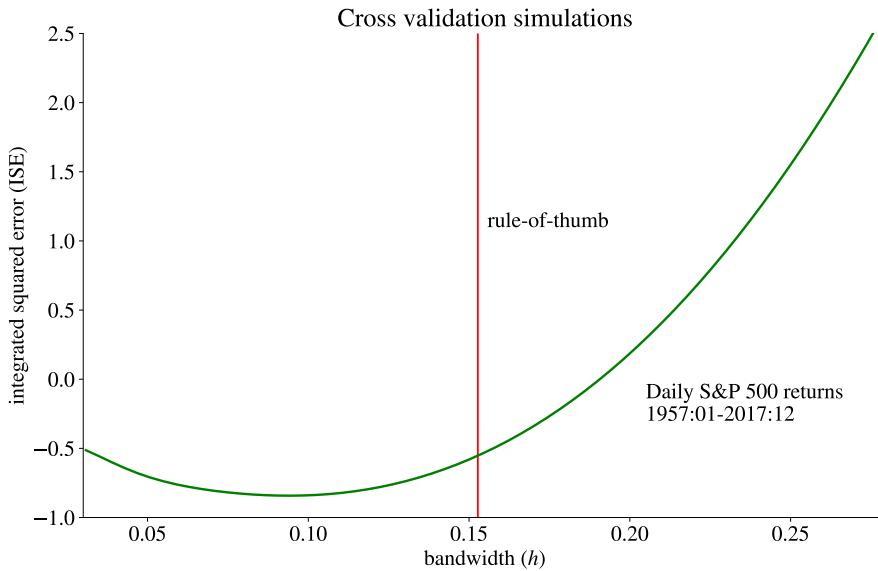


Figure 13.10: Stock returns

“location” parameter. This is not done in the text below, and it would not change any of the results.

Consider the loss X (the negative of the return) and let u be a threshold. Assume that the threshold exceedance ($X - u$) has a generalized Pareto distribution. Let P_u be probability of the loss being smaller than the threshold, that is, $P_u = \Pr(X \leq u)$. Then, the cdf of the loss for values greater than the threshold ($\Pr(X \leq x)$ for $x > u$) can be written

$$\Pr(X \leq x) = F(x) = P_u + G(x - u)(1 - P_u), \text{ for } x > u, \quad (13.13)$$

where $G(z)$ is the cdf of the generalized Pareto distribution. Noticed that, the cdf value is P_u at at $x = u$ (or just slightly above u), and that it becomes one as x goes to infinity.

Clearly, the pdf is

$$f(x) = g(x - u)(1 - P_u), \text{ for } x > u, \quad (13.14)$$

where $g(z)$ is the pdf of the generalized Pareto distribution. Notice that integrating the pdf from $x = u$ to infinity shows that the probability mass of X above u is $1 - P_u$. Since the probability mass below u is P_u , it adds up to unity (as it should). See Figures 14.2 and 14.5 for illustrations.

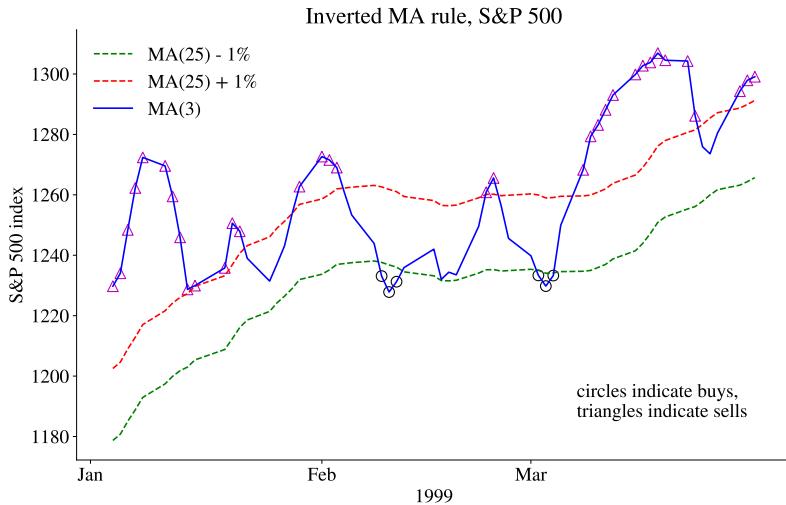


Figure 13.11: Examples of trading rules

It is often useful to calculate the *tail probability* $\Pr(X > x)$, which in the case of the cdf in (14.4) is

$$\Pr(X > x) = 1 - F(x) = (1 - P_u)[1 - G(x - u)], \quad (13.15)$$

where $G(z)$ is the cdf of the generalized Pareto distribution.

13.2.2 VaR and Expected Shortfall of a GP Distribution

The *value at risk*, VaR_α (say, $\alpha = 95\%$), is the α -th quantile of the loss distribution

$$\text{VaR}_\alpha = \text{cdf}_X^{-1}(\alpha), \quad (13.16)$$

where $\text{cdf}_X^{-1}()$ is the inverse cumulative distribution function of the losses. That is, VaR_α is the α quantile of the loss distribution. For instance, $\text{VaR}_{95\%}$ is the 0.95 quantile of the loss distribution. This clearly means that the probability of the loss to be less than VaR_α equals α

$$\Pr(X \leq \text{VaR}_\alpha) = \alpha. \quad (13.17)$$

(Equivalently, the $\Pr(X > \text{VaR}_\alpha) = 1 - \alpha$.)

Assuming $\text{VaR}_\alpha \geq u$ (that is, $\alpha \geq P_u$), the cdf (14.4) together with the form of the

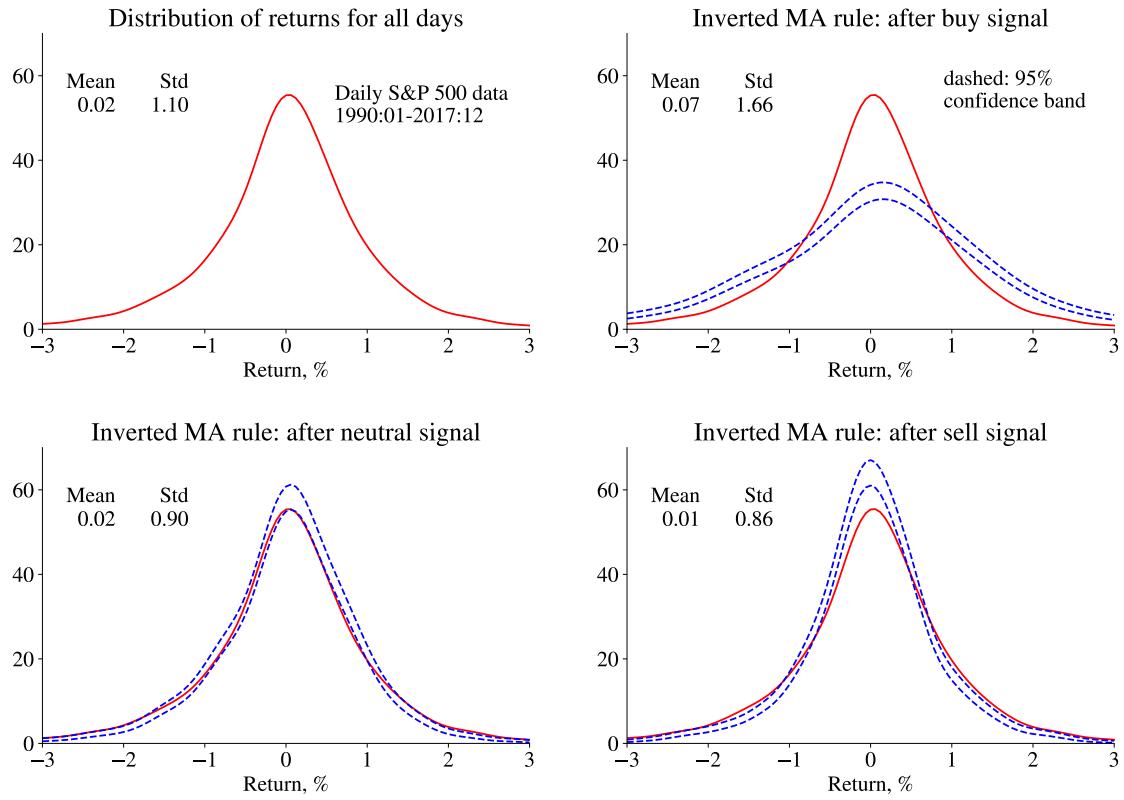


Figure 13.12: Examples of trading rules

generalized Pareto distribution give the VaR

$$\text{VaR}_\alpha = \begin{cases} u + \frac{\beta}{\xi} \left[\left(\frac{1-\alpha}{1-P_u} \right)^{-\xi} - 1 \right] & \text{if } \xi \neq 0 \\ u - \beta \ln \left(\frac{1-\alpha}{1-P_u} \right) & \xi = 0 \end{cases}, \text{ for } \alpha \geq P_u. \quad (13.18)$$

Proof. (of (14.7)) Set $F(x) = \alpha$ in (14.4) and use $z = x - u$ in the cdf from Remark 14.2 and solve for x . ■

If we assume $\xi < 1$ (to make sure that the mean is finite), then straightforward integration using (14.5) shows that the *expected shortfall* is

$$\begin{aligned} \text{ES}_\alpha &= \mathbb{E}(X | X \geq \text{VaR}_\alpha) \\ &= \frac{\text{VaR}_\alpha}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \text{ for } \alpha > P_u \text{ and } \xi < 1. \end{aligned} \quad (13.19)$$

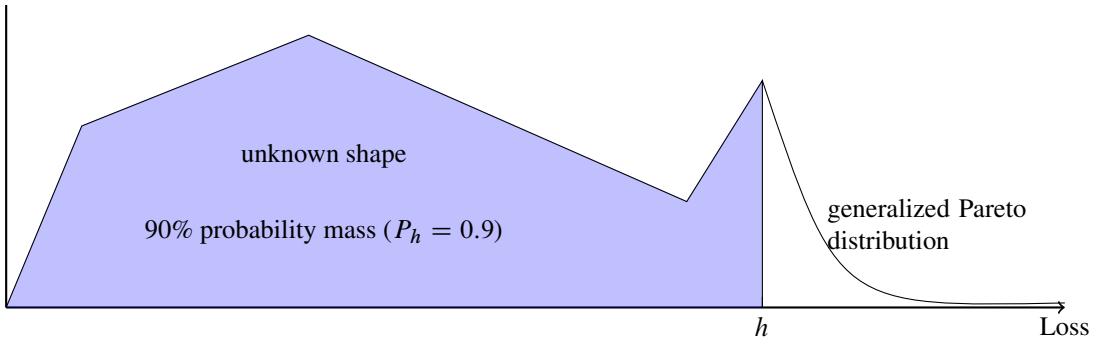


Figure 13.13: Loss distribution

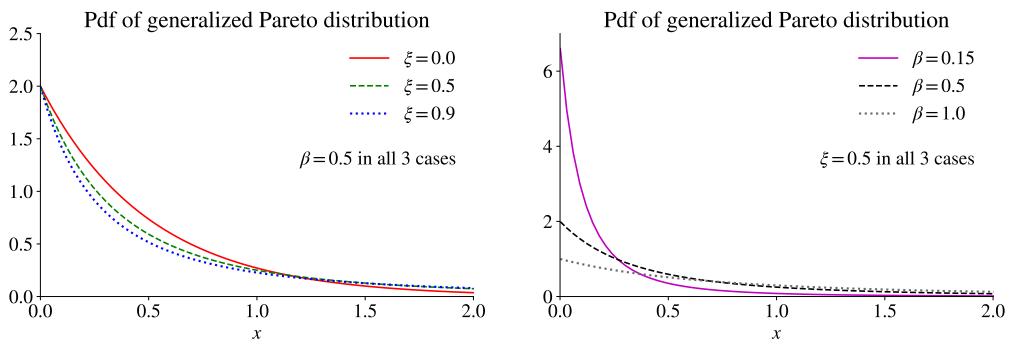


Figure 13.14: Generalized Pareto distributions

13.2.3 Expected Exceedance of a GP Distribution

The *expected exceedance* can help us to specify the cut-off level where the tail “starts,” that is, to choose the value of u . The average exceedance (in data) over some threshold level v is the mean of $X_t - v$ for those observations where $X_t > v$

$$\hat{e}(v) = \frac{\sum_{t=1}^T (X_t - v)\delta(X_t > v)}{\sum_{t=1}^T \delta(X_t > v)}, \text{ where } \quad (13.20)$$

$$\delta(q) = \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases}$$

(Notice that v denotes some general threshold level, while u denotes the threshold we eventually choose for the starting point of the GP distribution.)

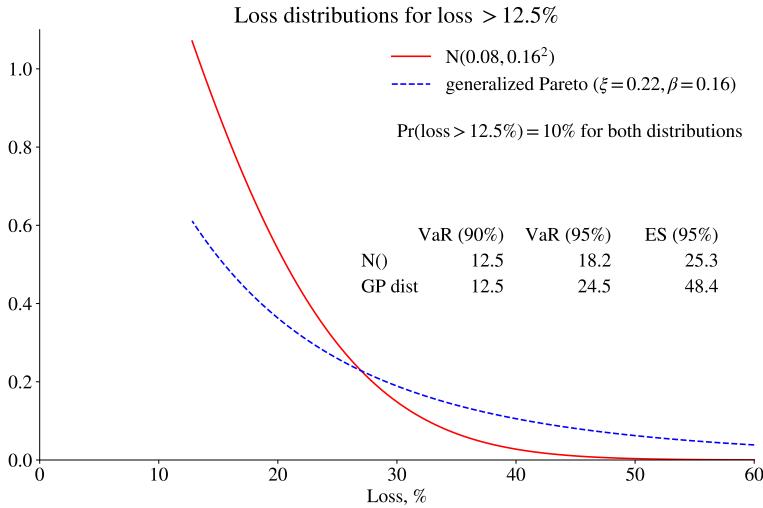


Figure 13.15: Comparison of a normal and a generalized Pareto distribution for the tail of losses

The expected exceedance of a GP distribution (with $\xi < 1$) for any threshold $v > u$ is

$$\begin{aligned} e(v) &= \mathbb{E}(X - v | X > v) \\ &= \frac{\xi v}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}, \text{ for } v > u \text{ and } \xi < 1. \end{aligned} \quad (13.21)$$

Proof. (of (14.9)) Substitute v for VaR_α in the expected shortfall (14.8)

$$\mathbb{E}(X | X \geq v) = \frac{v}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}$$

and subtract v from both sides to get (14.9). ■

The expected exceedance of a generalized Pareto distribution (with $0 < \xi < 1$) is increasing with the threshold level v , which indicates that the tail of the distribution is very long. To see that, notice that a normal distribution would typically show a negative relation: see Figure 14.5 for an illustration. This provides a way of assessing which distribution best fits the tail of the historical histogram. In addition, if we have decided to use the GP distribution for the tail, but does not know where the tail starts (the value of u), then it can be chosen as the lowest value (of v) after which the average exceedance in data (14.10) appears to be a linear function of the threshold. See Figure 14.6.

Remark 13.8 (*Expected exceedance from a normal distribution*) If $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} E(X - v | X > v) &= \mu + \sigma \frac{\phi(v_0)}{1 - \Phi(v_0)} - v, \\ \text{with } v_0 &= (v - \mu)/\sigma \end{aligned}$$

where $\phi()$ and Φ are the pdf and cdf of a $N(0, 1)$ variable respectively.

13.2.4 Estimating a GP Distribution

The estimation of the parameters of the distribution (ξ and β) is typically done by maximum likelihood. Alternatively, a comparison of the empirical exceedance (14.10) with the theoretical (14.9) can help. Suppose we calculate the empirical exceedance for different values of the threshold level (denoted v_i —all large enough so the relation looks linear), then we can estimate (by OLS)

$$\hat{e}(v_i) = a + b v_i + \varepsilon_i. \quad (13.22)$$

Then, the theoretical exceedance (14.9) for a given starting point of the GP distribution (u) is related to this regression according to (14.9) which implies

$$\begin{aligned} a &= \frac{\beta - \xi u}{1 - \xi} \text{ and } b = \frac{\xi}{1 - \xi}, \text{ or} \\ \xi &= \frac{b}{1 + b} \text{ and } \beta = a(1 - \xi) + \xi u. \end{aligned} \quad (13.23)$$

See Figure 14.6 for an illustration.

Remark 13.9 (*Log likelihood function of the loss distribution*) Since we have assumed that the threshold exceedance ($X - u$) has a generalized Pareto distribution, Remark 14.2 shows that the log likelihood for the observation of the loss above the threshold ($X_t > u$) is

$$\begin{aligned} L &= \sum_{t \text{ st. } X_t > u} L_t \\ \ln L_t &= \begin{cases} -\ln \beta - (1/\xi + 1) \ln [1 + \xi (X_t - u) / \beta] & \text{if } \xi \neq 0 \\ -\ln \beta - (X_t - u) / \beta & \xi = 0. \end{cases} \end{aligned}$$

This allows us to estimate ξ and β by maximum likelihood. Typically, u is not estimated, but imposed a priori (based on the expected exceedance).

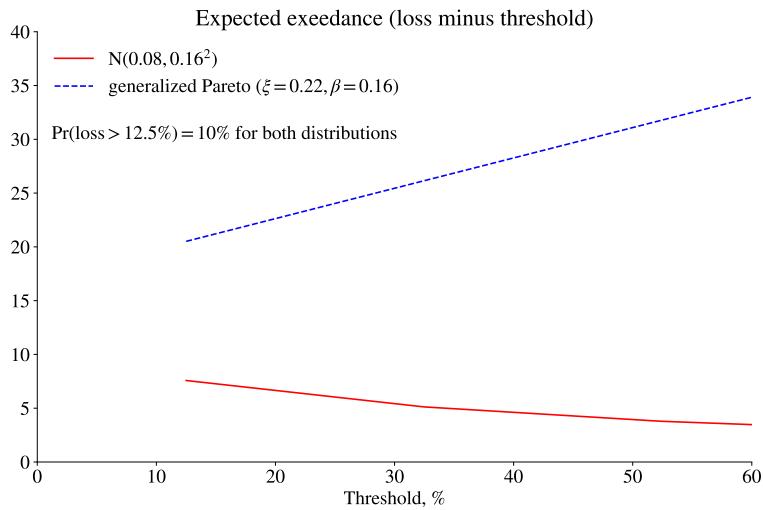


Figure 13.16: Expected exceedance, normal and generalized Pareto distribution

Example 13.10 (*Estimation of the generalized Pareto distribution on S&P daily returns*). *Figure 14.6* (upper left panel) shows that it may be reasonable to fit a GP distribution with a threshold $u = 1.3$ and that OLS estimation of (13.22) can provide useful parameter estimates. The upper right panel illustrates the estimated distribution, while the lower left panel shows that the highest quantiles are well captured by estimated distribution.

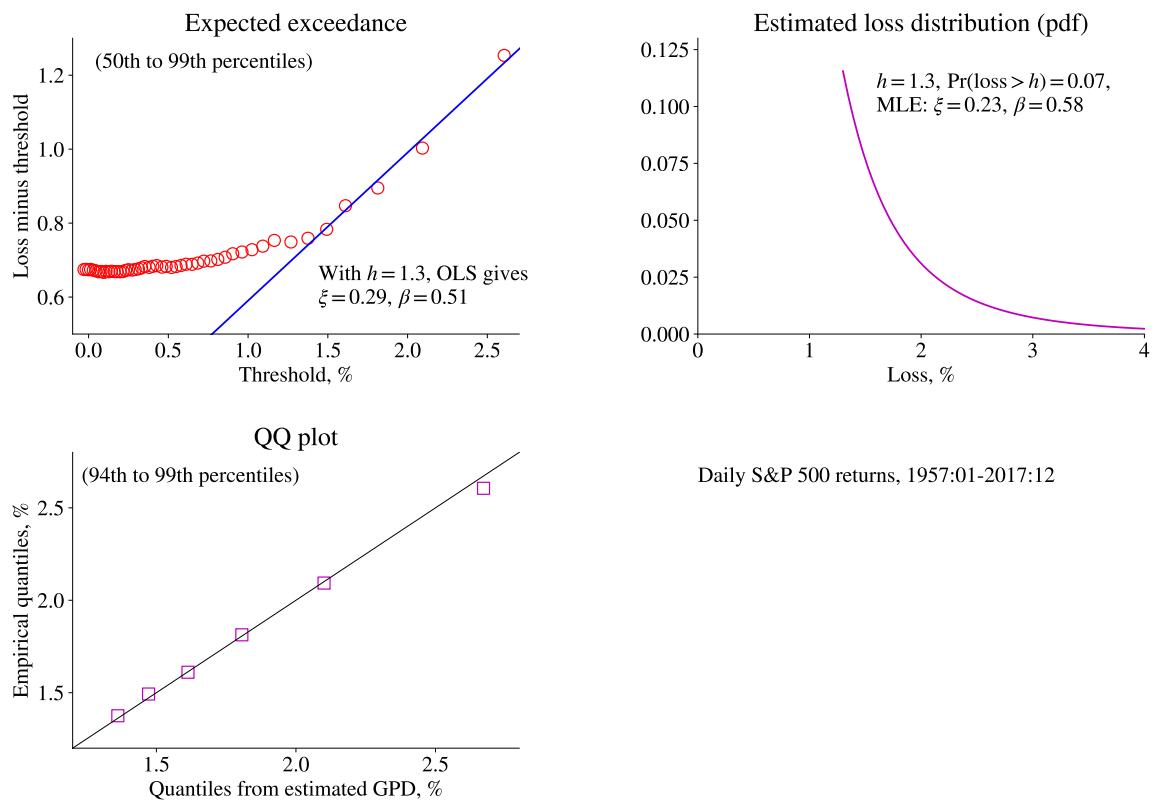


Figure 13.17: Results from S&P 500 data

Chapter 14

Return Distributions (Multivariate)^{*}

More advanced material is denoted by a star (*). It is not required reading.

14.1 Recap of Univariate Distributions

The cdf (cumulative distribution function) measures the probability that the random variable X_i is below or at some numerical value x_i ,

$$F_i(x_i) = \Pr(X_i \leq x_i). \quad (14.1)$$

For instance, with an $N(0, 1)$ distribution, $F(-1.64) = 0.05$. Clearly, the cdf values are between (and including) 0 and 1. The distribution of X_i is often called the *marginal distribution* of X_i —to distinguish it from the joint distribution of X_i and X_j . (See below for more information on joint distributions.)

The pdf (probability density function) $f_i(x_i)$ is the “height” of the distribution in the sense that the cdf $F(x_i)$ is the integral of the pdf from minus infinity to x_i

$$F_i(x_i) = \int_{s=-\infty}^{x_i} f_i(s)ds. \quad (14.2)$$

(Conversely, the pdf is the derivative of the cdf, $f_i(x_i) = \partial F_i(x_i)/\partial x_i$.) The Gaussian pdf (the normal distribution) is bell shaped.

Remark 14.1 (*Quantile of a distribution*) *The α quantile of a distribution (ξ_α) is the value of x such that there is a probability of α of a lower value. We can solve for the quantile by inverting the cdf, $\alpha = F(\xi_\alpha)$ as $\xi_\alpha = F^{-1}(\alpha)$. For instance, the 5% quantile of a $N(0, 1)$ distribution is $-1.64 = \Phi^{-1}(0.05)$, where $\Phi^{-1}()$ denotes the inverse of an $N(0, 1)$ cdf. See Figure 14.1 for an illustration.*

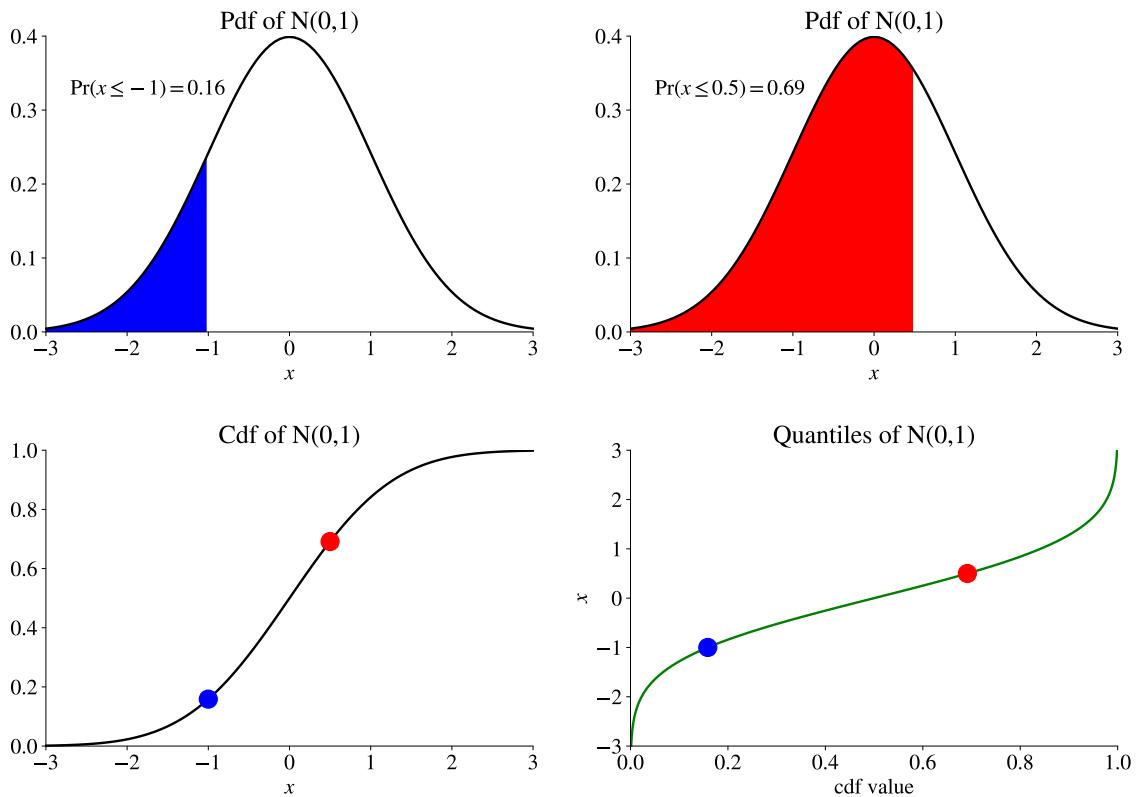


Figure 14.1: Finding quantiles of a $N(\mu, \sigma^2)$ distribution

14.2 Threshold Exceedance and Tail Distribution*

Reference: McNeil, Frey, and Embrechts (2005) 7

In risk control, the focus is on the distribution of losses beyond some threshold level. This has three direct implications. First, the object under study is the loss

$$X = -R, \quad (14.3)$$

that is, the negative of the return. Second, the attention is on how the distribution looks like beyond a threshold and also on the probability of exceeding this threshold. In contrast, the exact shape of the distribution below that point is typically disregarded. Third, modelling the tail of the distribution is best done by using a distribution that allows for a much heavier tail than suggested by a normal distribution. The generalized Pareto (GP) distribution is often used.

Remark 14.2 (*Cdf and pdf of the generalized Pareto distribution*) The generalized Pareto

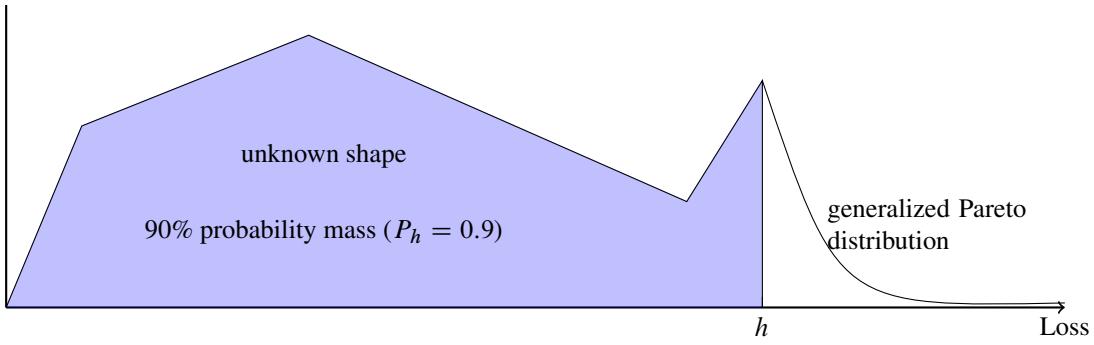


Figure 14.2: Loss distribution

distribution is described by a scale parameter ($\beta > 0$) and a shape parameter (ξ). The cdf ($\Pr(Z \leq z)$, where Z is the random variable and z is a value) is

$$G(z) = \begin{cases} 1 - (1 + \xi z / \beta)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-z / \beta) & \xi = 0, \end{cases}$$

for $0 \leq z$ if $\xi \geq 0$ and $z \leq -\beta/\xi$ in case $\xi < 0$. The pdf is therefore

$$g(z) = \begin{cases} \frac{1}{\beta} (1 + \xi z / \beta)^{-1/\xi - 1} & \text{if } \xi \neq 0 \\ \frac{1}{\beta} \exp(-z / \beta) & \xi = 0. \end{cases}$$

The mean is defined (finite) if $\xi < 1$ and is then $E(Z) = \beta / (1 - \xi)$. Similarly, the variance is finite if $\xi < 1/2$ and is then $\text{Var}(Z) = \beta^2 / [(1 - \xi)^2 (1 - 2\xi)]$. See Figure 14.3 for an illustration. In the application below we will use $z = x - h$ where x is a random variable and h a threshold level. In some texts this h is part of the definition of the distribution, and is often called the “location” parameter. This is not done in the text below, and it would not change any of the results.

Consider the loss X (the negative of the return) and let h be a threshold. Assume that the threshold exceedance ($X - h$) has a generalized Pareto distribution. Let P_h be probability of $X \leq h$, that is, $P_h = \Pr(X \leq h)$. Then, the cdf of the loss for values greater than the threshold ($\Pr(X \leq x)$ for $x > h$) can be written

$$F(x) = P_h + G(x - h)(1 - P_h), \text{ for } x > h, \quad (14.4)$$

where $G(z)$ is the cdf of the generalized Pareto distribution. Noticed that, the cdf value is P_h at $x = h$, and that it becomes one as x goes to infinity. See Figure 14.2 for an

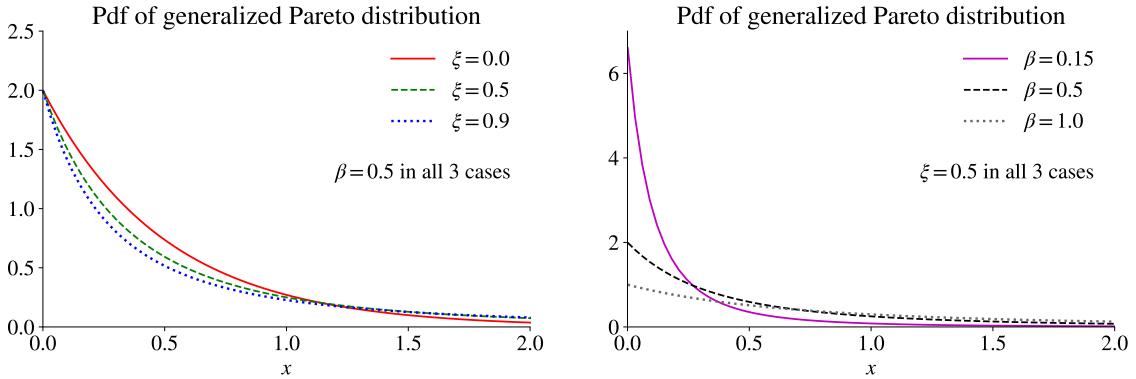


Figure 14.3: Generalized Pareto distributions

illustration.

Clearly, the pdf is

$$f(x) = g(x - h)(1 - P_h), \text{ for } x > h, \quad (14.5)$$

where $g(z)$ is the pdf of the generalized Pareto distribution. Notice that integrating the pdf from $x = h$ to infinity shows that the probability mass of X above h is $1 - P_h$. Since the probability mass below h is P_h , it adds up to unity (as it should). See Figure 14.5 for an illustration.

Remark 14.3 (*Tail probability) *The tail probability $\Pr(X > x)$ is simply the complement of the cdf, $1 - F(x)$.*

Remark 14.4 (*Distribution for all x values) *Suppose we want to specify the distribution for all x values, not just for $x > h$. For $x \leq h$ we pick a pdf of the type $j(x)$, for instance, from the pdf of an $N(\mu, \sigma^2)$, but we have to scale it so the total probability integrates to one. To do that, let $J(h)$ be the cdf value at $x = h$. Then, we set*

$$f(x) = j(x)P_h/J(h) \text{ for } x \leq h,$$

and according to (14.5) for $x > h$. The corresponding cdf is

$$F(x) = J(x)P_h/J(h) \text{ for } x \leq h,$$

and according to (14.4) for $x > h$. This last equation shows that $F(x) = P_h$ as required.

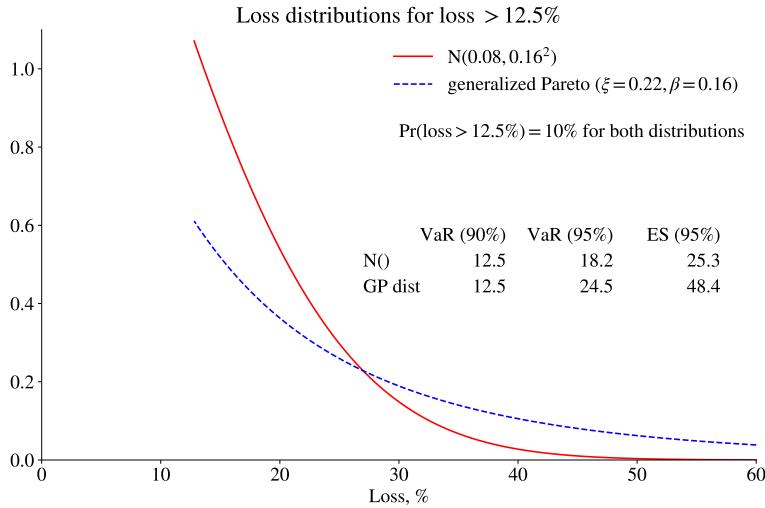


Figure 14.4: Comparison of a normal and a generalized Pareto distribution for the tail of losses

The VaR_α (say, $\alpha = 95\%$) is the α -th quantile of the loss distribution

$$\text{VaR}_\alpha = \text{cdf}_X^{-1}(\alpha), \quad (14.6)$$

where $\text{cdf}_X^{-1}()$ is the inverse cumulative distribution function of the losses. For instance, $\text{VaR}_{95\%}$ is the 0.95 quantile of the loss distribution. This clearly means that the probability of the loss to be less than VaR_α equals α .

Assuming $\text{VaR}_\alpha \geq h$ (that is, $\alpha \geq P_h$), the cdf (14.4) together with the form of the generalized Pareto distribution in give Remark 14.2

$$\text{VaR}_\alpha = \begin{cases} h + \frac{\beta}{\xi} \left[\left(\frac{1-\alpha}{1-P_h} \right)^{-\frac{1}{\xi}} - 1 \right] & \text{if } \xi \neq 0 \\ h - \beta \ln \left(\frac{1-\alpha}{1-P_h} \right) & \xi = 0 \end{cases}, \text{ for } \alpha \geq P_h. \quad (14.7)$$

Proof. (of (14.7)) Set $F(x) = \alpha$ in (14.4) and use $z = x - h$ in the cdf from Remark 14.2 and solve for x . ■

If we assume that the GDP parameter $\xi < 1$ (to make sure that the mean is finite), then straightforward integration using (14.5) shows that the expected shortfall is

$$\begin{aligned} \text{ES}_\alpha &= \text{E}(X | X \geq \text{VaR}_\alpha) \\ &= \frac{\text{VaR}_\alpha}{1-\xi} + \frac{\beta - \xi h}{1-\xi}, \text{ for } \alpha > P_h \text{ and } \xi < 1. \end{aligned} \quad (14.8)$$

It follows that the expected exceedance of a GP distribution (with $\xi < 1$) for any threshold $v > h$ (recall, h is the threshold where the GDP “starts”) is

$$\begin{aligned} e(v) &= \mathbb{E}(X - v | X > v) \\ &= \frac{\xi v}{1 - \xi} + \frac{\beta - \xi h}{1 - \xi}, \text{ for } v > h \text{ and } \xi < 1. \end{aligned} \quad (14.9)$$

Proof. (of (14.9)) Substitute v for VaR_α in the expected shortfall (14.8)

$$\mathbb{E}(X | X \geq v) = \frac{v}{1 - \xi} + \frac{\beta - \xi h}{1 - \xi}$$

and subtract v from both sides to get (14.9). ■

The expected exceedance of a generalized Pareto distribution (with $0 < \xi < 1$) is increasing with the threshold level v . This indicates that the tail of the distribution is very long. In contrast, a normal distribution would typically show a negative relation (see Figure 14.5 for an illustration). This provides a way of assessing which distribution that best fits the tail of the data.

Remark 14.5 (*Expected exceedance from a normal distribution) If $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} \mathbb{E}(X - v | X > v) &= \mu + \sigma \frac{\phi(v_0)}{1 - \Phi(v_0)} - v, \\ \text{with } v_0 &= (v - \mu)/\sigma \end{aligned}$$

where $\phi()$ and Φ are the pdf and cdf of a $N(0, 1)$ variable respectively.

The expected exceedance over v is often compared with an empirical estimate of the same thing: the mean of $X_t - v$ for those observations where $X_t > v$

$$\begin{aligned} \hat{e}(v) &= \frac{\sum_{t=1}^T (X_t - v) \delta(X_t > v)}{\sum_{t=1}^T \delta(X_t > v)}, \text{ where} \\ \delta(q) &= \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases} \end{aligned} \quad (14.10)$$

If it is found that $\hat{e}(v)$ is increasing (more or less) linearly with the threshold level (v) as in (14.9), then it is reasonable to model the tail of the distribution from that point as a generalized Pareto distribution.

The estimation of the parameters of the GP distribution (ξ and β) is typically done by maximum likelihood. Alternatively, a comparison of the empirical exceedance (14.10)

with the theoretical (14.9) can help (details in the following remark).

Remark 14.6 (*Estimating the GP parameters by OLS*) Suppose we calculate the empirical exceedance for different values of the threshold level (denoted v_i —all large enough so the relation looks linear), then we can estimate (by OLS)

$$\hat{e}(v_i) = a + bv_i + \varepsilon_i.$$

Then, the theoretical exceedance (14.9) for a given starting point of the GPD h is related to this regression according to (14.9) which implies

$$a = \frac{\beta - \xi h}{1 - \xi} \text{ and } b = \frac{\xi}{1 - \xi}, \text{ or}$$

$$\xi = \frac{b}{1 + b} \text{ and } \beta = a(1 - \xi) + \xi h.$$

See Figure 14.6 for an illustration.

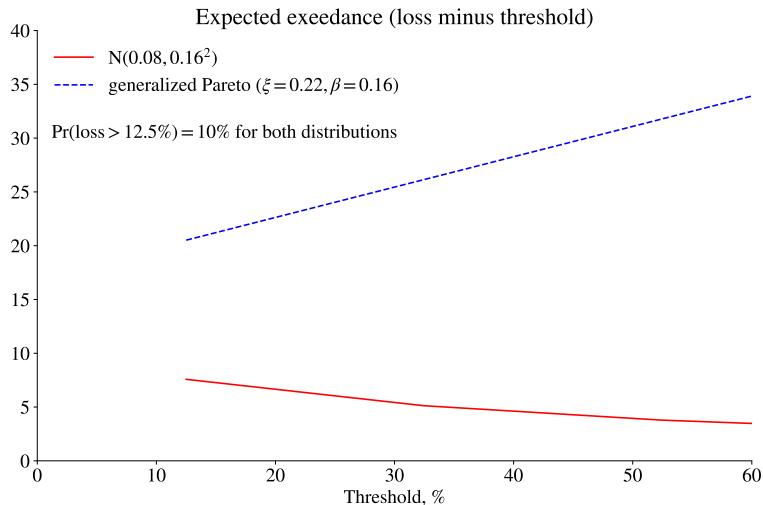


Figure 14.5: Expected exceedance, normal and generalized Pareto distribution

Remark 14.7 (*Log likelihood function of the loss distribution*) Since we have assumed that the threshold exceedance $(X - h)$ has a generalized Pareto distribution, Remark 14.2 shows that the log likelihood for the observation of the loss above the threshold $(X_t > h)$

is

$$L = \sum_{t \text{ st. } X_t > h} L_t$$

$$\ln L_t = \begin{cases} -\ln \beta - (1/\xi + 1) \ln [1 + \xi (X_t - h) / \beta] & \text{if } \xi \neq 0 \\ -\ln \beta - (X_t - h) / \beta & \xi = 0. \end{cases}$$

This allows us to estimate ξ and β by maximum likelihood. Typically, h is not estimated, but imposed *a priori* (based on the expected exceedance).

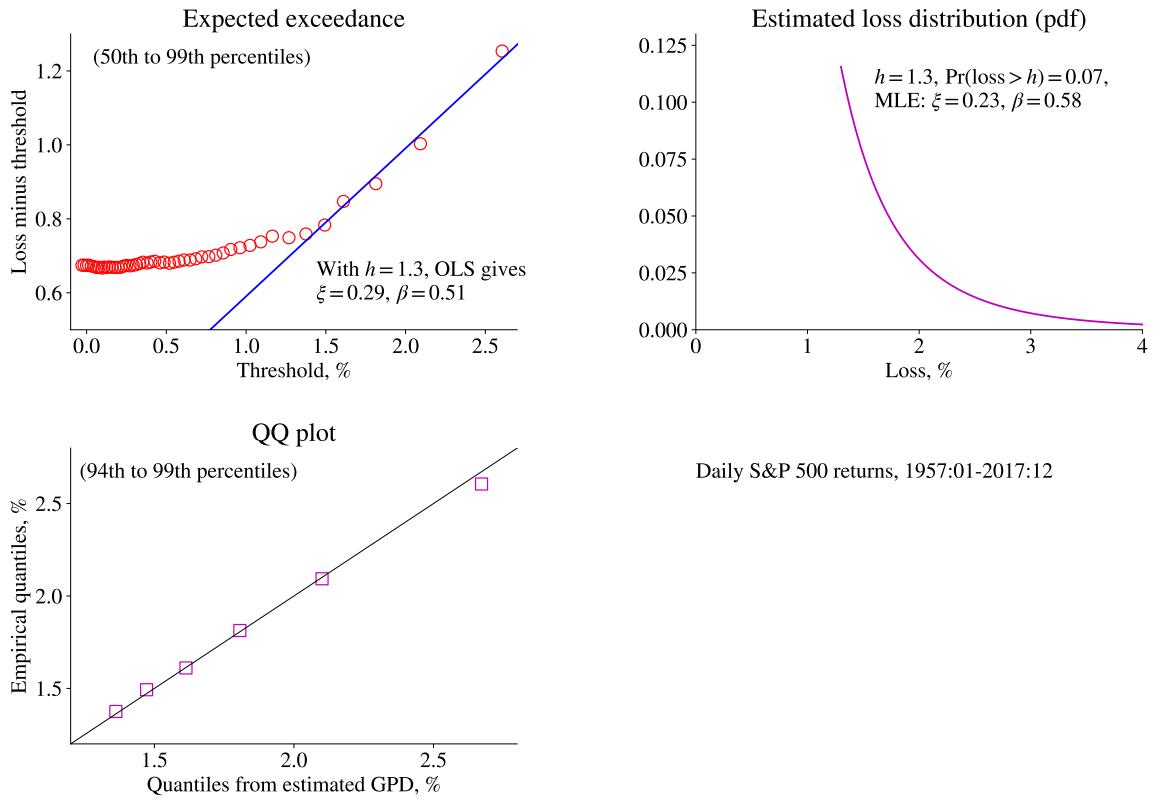


Figure 14.6: Results from S&P 500 data

Example 14.8 (*Estimation of the generalized Pareto distribution on S&P daily returns*).

Figure 14.6 (upper left panel) shows that it may be reasonable to fit a GP distribution with a threshold $h = 1.3$ and that OLS estimation as in Remark 14.6 can provide useful parameter estimates.. The upper right panel illustrates the estimated distribution, while the lower left panel shows that the highest quantiles are well captured by estimated distribution.

14.3 Exceedance Correlations*

Reference: Ang and Chen (2002)

It is often argued that most assets are more strongly correlated in down markets than in up markets. If so, diversification may not be such a powerful tool as what we would otherwise believe.

A straightforward way of examining this is to calculate the correlation of two returns (x and y , say) for specific intervals. For instance, we could specify that x_t should be between h_1 and h_2 and y_t between k_1 and k_2

$$\text{Corr}(x_t, y_t | h_1 < x_t \leq h_2, k_1 < y_t \leq k_2). \quad (14.11)$$

For instance, by setting the lower boundaries (h_1 and k_1) to $-\infty$ and the upper boundaries (h_2 and k_2) to 0, we get the correlation in down markets.

A (bivariate) normal distribution has little probability mass at very low returns, which leads to the correlation being squeezed towards zero as we only consider data far out in the tail. This means that the tail correlation of a normal distribution is always closer to zero than the correlation for all data points. This is illustrated in Figure 14.7.

In contrast, Figures 14.8–14.9 suggest (for two US portfolios) that the correlation in the lower tail is high. This suggests that the relation between the two returns in the tails is not well described by a normal distribution. In particular, we need to use a distribution that allows for much stronger dependence in the lower tail. Otherwise, the diversification benefits (in down markets) are likely to be exaggerated.

14.4 Beyond (Linear) Correlations*

Reference: Alexander (2008a) 6, McNeil, Frey, and Embrechts (2005)

The standard correlation (also called Pearson's correlation) measures the linear relation between two variables, that is, to what extent one variable can be explained by a linear function of the other variable (and a constant). That is adequate for most issues in finance, but we sometimes need to capture non-linear relations. It also turns out to be easier to calibrate/estimate copulas (see below) by using other measures of dependency.

Spearman's rank correlation (called Spearman's rho and often denoted ρ_S) measures to what degree two variables have a monotonic relation: it is the correlation of their respective ranks. It measures if one variable tends to be high when the other also is—without imposing the restriction that this relation must be linear. See Figure 14.10 for an

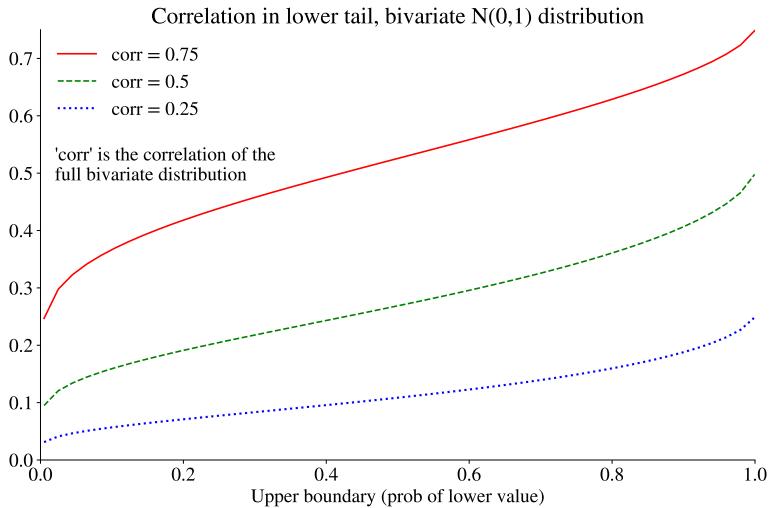


Figure 14.7: Correlation in lower tail when data is drawn from a normal distribution with correlation ρ

illustration.

It is computed in two steps. First, the data is *ranked* from the smallest (rank 1) to the largest (ranked T , where T is the sample size). Ties (when two or more observations have the same values) are handled by averaging the ranks. The following illustrates this for two variables

x_t	rank(x_t)	y_t	rank(y_t)
2	2.5	7	2
10	4	8	3
-3	1	2	1
2	2.5	10	4

(14.12)

In the second step, simply estimate the correlation of the ranks of two variables

$$\text{Spearman's } \rho = \text{Corr}[\text{rank}(x_t), \text{rank}(y_t)]. \quad (14.13)$$

Clearly, this correlation is between -1 and 1 .

Remark 14.9 (*Alternative way of calculating the rank correlation) There is an alternative way of calculating the rank correlation based on the difference of the ranks, $d_t = \text{rank}(x_t) - \text{rank}(y_t)$, $\rho = 1 - 6\sum_{t=1}^T d_t^2 / (T^3 - T)$. It gives the same result if there are no tied ranks.

Remark 14.10 (*Formal definition of Spearman's ρ) If we have specified the joint distribution of the random variables X and Y , then we can also calculate the implied Spear-

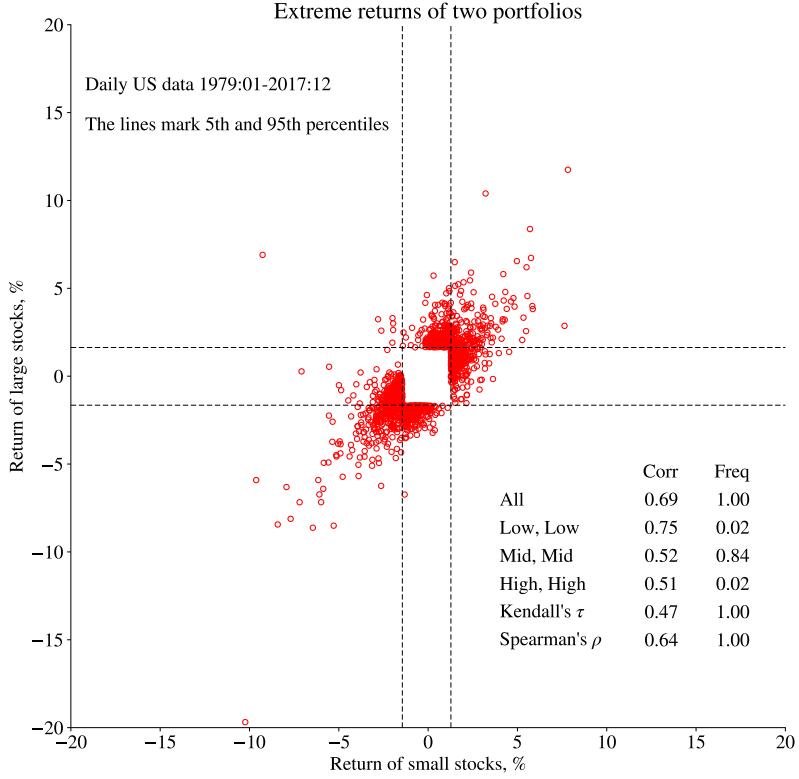


Figure 14.8: Correlation of two portfolios

man's ρ (sometimes only numerically) as $\text{Corr}[F_X(X), F_Y(Y)]$ where $F_X(X)$ is the cdf of X and $F_Y(Y)$ of Y .

The rank correlation can be tested by using the fact that (under the null hypothesis)

$$\sqrt{T-1} \times \text{Spearman's } \rho \xrightarrow{d} N(0, 1). \quad (14.14)$$

(For samples of 20 to 40 observations, it is often recommended to use $\sqrt{(T-2)/(1-\hat{\rho}_S^2)}\hat{\rho}_S$ where $\hat{\rho}_S$ denotes Spearman's ρ . This has a t_{T-2} distribution.)

Kendall's rank correlation (called Kendall's τ) is similar, but is based on comparing changes of x_t (compared to each of x_1, \dots, x_{t-1}) with the corresponding changes of y_t . For instance, with three data points $((x_1, y_1), (x_2, y_2), (x_3, y_3))$ we first calculate

Changes of x	Changes of y	(14.15)
$x_2 - x_1$	$y_2 - y_1$	
$x_3 - x_1$	$y_3 - y_1$	
$x_3 - x_2$	$y_3 - y_2$	

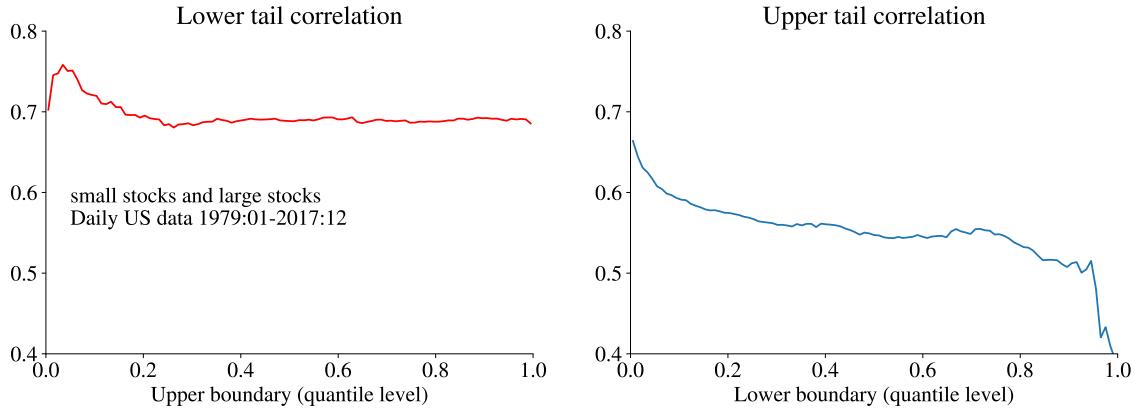


Figure 14.9: Correlation in the tails for two portfolios

which gives $T(T - 1)/2$ (here 3) pairs. Then, we investigate if the pairs are concordant (same sign of the change of x and y) or discordant (different signs) pairs

$$ij \text{ is concordant if } (x_j - x_i)(y_j - y_i) > 0 \quad (14.16)$$

$$ij \text{ is discordant if } (x_j - x_i)(y_j - y_i) < 0.$$

Finally, we count the number of concordant (T_c) and discordant (T_d) pairs and calculate Kendall's tau as

$$\text{Kendall's } \tau = \frac{T_c - T_d}{T(T - 1)/2}. \quad (14.17)$$

It can be shown that

$$\text{Kendall's } \tau \xrightarrow{d} N\left(0, \frac{4T + 10}{9T(T - 1)}\right), \quad (14.18)$$

so it is straightforward to test τ by a t-test.

Remark 14.11 (*Formal definition of Kendall's τ) Let x and \tilde{x} be independent draws (“independent copies”) from the same distribution, and similarly for y and \tilde{y} . Then, $\tau = E \text{sign}[(x - \tilde{x})(y - \tilde{y})]$, where $\text{sign}() = 1$ if the argument is positive, 0 if it is zero and -1 if it is negative.

Example 14.12 (Kendall's tau) Suppose the data is

$$\begin{array}{cc} \underline{x} & \underline{y} \\ 2 & 7 \\ 10 & 9 \\ -3 & 10. \end{array}$$

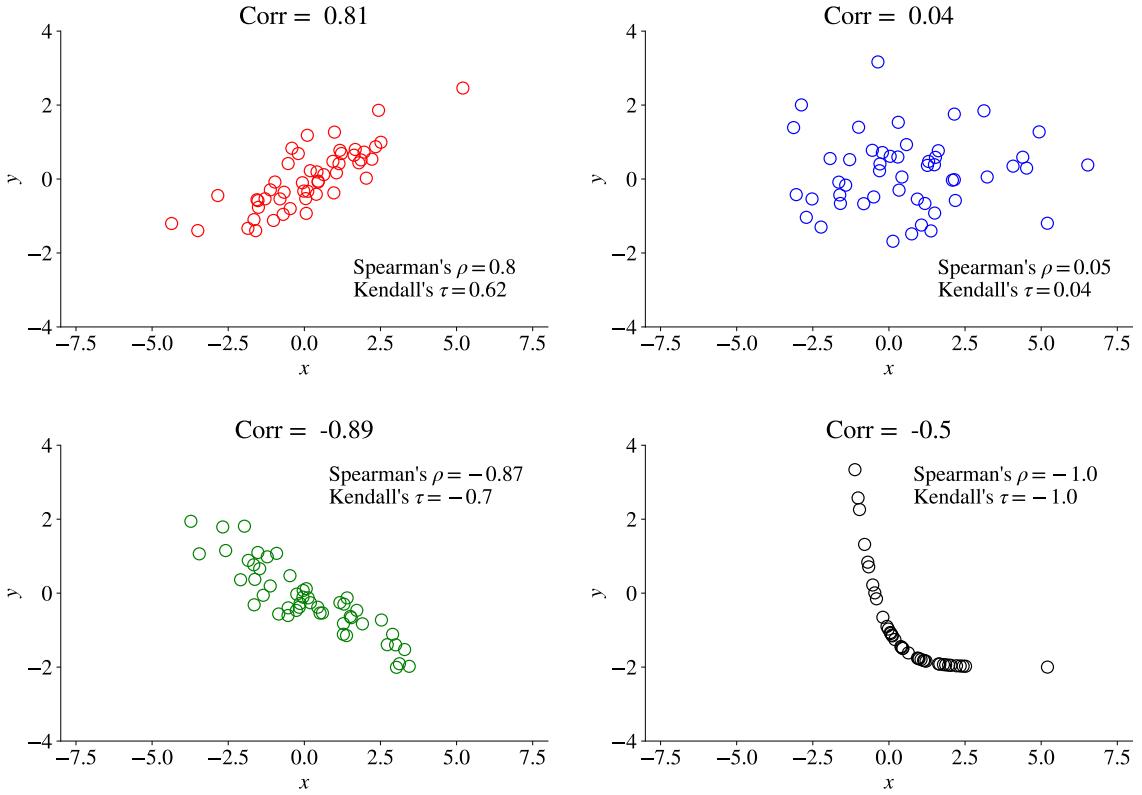


Figure 14.10: Illustration of correlation and rank correlation

We then get the following changes

<u>Changes of x</u>	<u>Changes of y</u>	
$x_2 - x_1 = 10 - 2 = 8$	$y_2 - y_1 = 9 - 7 = 2$	concordant
$x_3 - x_1 = -3 - 2 = -5$	$y_3 - y_1 = 10 - 7 = 3$	discordant
$x_3 - x_2 = -3 - 10 = -13$	$y_3 - y_2 = 10 - 9 = 1,$	discordant.

Kendall's tau is therefore

$$\tau = \frac{1 - 2}{3(3 - 1)/2} = -\frac{1}{3}.$$

If x and y actually has bivariate normal distribution with correlation ρ , then it can be shown that on average we have

$$\text{Spearman's rho} = \frac{6}{\pi} \arcsin(\rho/2) \approx \rho \quad (14.19)$$

$$\text{Kendall's tau} = \frac{2}{\pi} \arcsin(\rho). \quad (14.20)$$

In this case, all three measures give similar messages (although the Kendall's tau tends to be lower than the linear correlation and Spearman's rho). This is illustrated in Figure 14.11. Clearly, when data is not normally distributed, then these measures can give distinctly different answers.

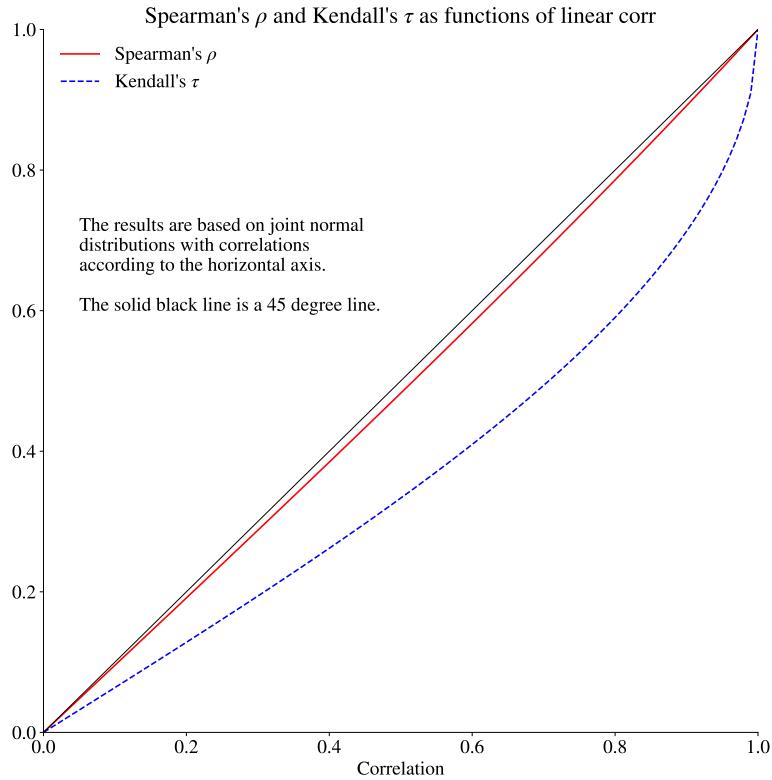


Figure 14.11: Spearman's rho and Kendall's tau if data has a bivariate normal distribution

A *joint α -quantile exceedance probability* measures how often two random variables (x and y , say) are both above their respective α quantiles. Similarly, we can also define the probability that they are *both* below their respective α quantiles

$$G_\alpha = \Pr(x \leq \xi_{x,\alpha}, y \leq \xi_{y,\alpha}), \quad (14.21)$$

$\xi_{x,\alpha}$ and $\xi_{y,\alpha}$ are α -quantile of the x - and y -distribution respectively.

In practice, this can be estimated from data by first finding the empirical α -quantiles ($\hat{\xi}_{x,\alpha}$ and $\hat{\xi}_{y,\alpha}$) by simply sorting the data and then picking out the value of observation

αT of this sorted list (do this individually for x and y). Then, calculate the estimate

$$\hat{G}_\alpha = \frac{1}{T} \sum_{t=1}^T \delta_t, \text{ where} \quad (14.22)$$

$$\delta_t = \begin{cases} 1 & \text{if } x_t \leq \hat{\xi}_{x,\alpha} \text{ and } y_t \leq \hat{\xi}_{y,\alpha} \\ 0 & \text{otherwise.} \end{cases}$$

See Figure 14.15 for an illustration based on a joint normal distribution and Figure 14.12 for an empirical example.

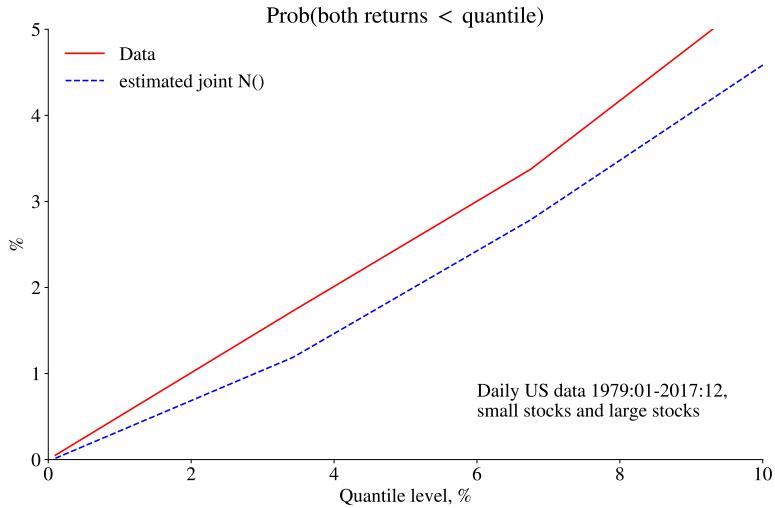


Figure 14.12: Probability of joint low returns

14.5 Copulas*

Reference: McNeil, Frey, and Embrechts (2005), Alexander (2008a) 6, Jondeau, Poon, and Rockinger (2007) 6

Portfolio choice and risk analysis depend crucially on the joint distribution of asset returns. Empirical evidence suggest that many returns have non-normal distribution, especially when we focus on the tails. There are several ways of estimating complicated joint (non-normal) distributions: using copulas is one. This approach has the advantage that it proceeds in two steps: first we estimate the marginal distribution of each return separately, then we model the comovements by a copula.

14.5.1 Multivariate Distributions and Copulas

Any joint (bivariate) pdf can be written as

$$f_{1,2}(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2), \text{ with} \quad (14.23)$$

$$u_i = F_i(x_i),$$

where $c()$ is a *copula density* function, $u_i = F_i(x_i)$ is short-hand notation for the cdf value as in (14.1) and $f_i(x_i)$ for the pdf values as in (14.2). The extension to three or more random variables is straightforward. Notice that $u_i = F_i(x_i)$ and $f_i(x_i)$ both depend refer to the same x_i value.

Equation (14.23) means that if we know the joint pdf $f_{1,2}(x_1, x_2)$ —and thus also the cdfs $F_1(x_1)$ and $F_2(x_2)$ —then we can figure out what the copula density function is. Alternatively, if we know the marginal (univariate) distributions ($f_i(x_i)$) and thus $F_i(x_i)$ —and the copula function, then we can construct the joint distribution. (This is called Sklar's theorem.) This latter approach will turn out to be useful as an alternative way of modelling the joint distribution.

The correlation of x_1 and x_2 depends on both the copula and the marginal distributions. In contrast, both Spearman's rho and Kendall's tau are determined by the copula only. They therefore provide a way of calibrating/estimating the copula without having to involve the marginal distributions directly.

Example 14.13 (Independent X and Y) If X and Y are independent, then we know that $f_{1,2}(x_1, x_2) = f_1(x_1) f_2(x_2)$, so the copula density function is just a constant equal to one.

Remark 14.14 (Joint cdf) A joint cdf of two random variables (X_1 and X_2) is defined as

$$F_{1,2}(x_1, x_2) = \Pr(X_1 \leq x_1 \text{ and } X_2 \leq x_2).$$

This cdf is obtained by integrating the joint pdf $f_{1,2}(x_1, x_2)$ over both variables

$$F_{1,2}(x_1, x_2) = \int_{s=-\infty}^{x_1} \int_{t=-\infty}^{x_2} f_{1,2}(s, t) ds dt.$$

(Conversely, the pdf is the mixed derivative of the cdf, $f_{1,2}(x_1, x_2) = \partial^2 F_{1,2}(x_1, x_2) / \partial x_1 \partial x_2$.) See Figure 14.13 for an illustration.

Remark 14.15 (From joint to univariate pdf) The pdf of x_1 (also called the marginal pdf of x_1) can be calculate from the joint pdf as $f_1(x_1) = \int_{x_2=-\infty}^{\infty} f_{1,2}(x_1, x_2) dx_2$.

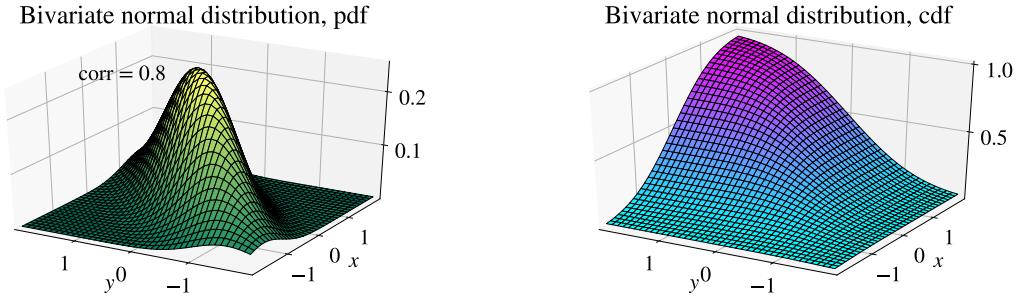


Figure 14.13: Bivariate normal distributions

Remark 14.16 (*Joint pdf and copula density, n variables*) For n variables (14.23) generalizes to

$$f_{1,2,\dots,n}(x_1, x_2, \dots, x_n) = c(u_1, u_2, \dots, u_n) f_1(x_1) f_2(x_2) \dots f_n(x_n), \text{ with} \\ u_i = F_i(x_i),$$

Remark 14.17 (*Cdfs and copulas**) The joint cdf can be written as

$$F_{1,2}(x_1, x_2) = C(u_1, u_2), \text{ with } u_i = F_i(x_i),$$

where $C()$ is the unique copula function. Taking derivatives gives (14.23) where

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}.$$

Notice the derivatives are with respect to $u_i = F_i(x_i)$, not x_i . Conversely, integrating the density over both u_1 and u_2 gives the copula function $C()$.

14.5.2 The Gaussian Copula

The bivariate Gaussian copula density function is

$$c(u) = \frac{1}{|R|^{1/2}} \exp\left[-\frac{1}{2} v' (R^{-1} - I_2) v\right], \text{ with } v = \begin{bmatrix} \Phi^{-1}(u_1) \\ \Phi^{-1}(u_2) \end{bmatrix}, \quad (14.24)$$

where $\Phi^{-1}(u_1)$ is the inverse of the $N(0, 1)$ cdf and R is the correlation matrix of u . With n variables, change I_2 to I_n . See Figure 14.14 for an illustration of how the copula density depends on the correlation.

Remark 14.18 (*The Gaussian copula function*) It can be shown that integrating (14.24) twice (over u_1 and u_2), gives the Gaussian copula function

$$C(u_1, u_2) = \Phi_R(v),$$

where Φ_R is the cdf of an joint $N(\mathbf{0}, R)$ distribution and v is defined in (14.24). (This result might look surprising given the form of the density in (14.24). The explanation is that we integrate with respect to u , not v .)

Notice that when using this function in (14.23) to construct the joint pdf, we take the following steps: (1) calculate the cdf values $u_i = F_i(x_i)$ from the univariate distribution of x_i (which may be non-normal); (2) calculate the quantiles of the same x_i values according to a standard normal distribution $\xi_i = \Phi^{-1}(u_i)$; (3) use in (14.24); (4) combine the results as in (14.23)). See Figures 14.16–14.17 for illustrations of the resulting joint distribution. These figures show results for two different choices of the marginal distributions—to highlight that the marginal distributions clearly matter for the joint pdf.

It can be shown that assuming that the marginal pdfs ($f_1(x_1)$ and $f_2(x_2)$) are $N(\mu_i, \sigma_i^2)$ and then combining with the Gaussian copula density (14.24) recovers a bivariate normal distribution. However, the way we typically use copulas is to assume (and estimate) some other type of univariate marginal distributions, for instance, with fat tails—and then combine with a (perhaps Gaussian) copula density to create the joint distribution.

A zero correlation ($R = I_2$) makes the copula density (14.24) equal to unity—so the joint density is just the product of the marginal densities. A positive correlation makes the copula density high when both x_1 and x_2 deviate from their means in the same direction. The easiest way to calibrate a Gaussian copula is therefore to set

$$\rho = \text{Spearman's rho}(x_1, x_2) \tag{14.25}$$

as suggested by (14.19). Notice that the rank correlation (Spearman's rho) is the same for (x_1, x_2) as for (u_1, u_2) .

Alternatively, the ρ parameter can be calibrated to fit the properties of the joint tail distribution of data. For instance, to match the joint α -quantile exceedance probability in data (14.22) and as implied by the copula.

To calculate joint α -quantile exceedance probability implied by a specific copula (and its parameters), we need to find the copula function (essentially the cdf) corresponding to a copula density. For some copulas (see below), there are analytical results, but in other cases we need to use numerical integration and/or simulations. Some results are given

in remarks below. See Figure 14.15: for the Gaussian copula, the figure shows that a higher correlation implies a larger probability that both variables are very low—but that the probabilities quickly become very small as we move towards lower quantiles (lower returns).

14.5.3 The Clayton Copula

The Gaussian copula is useful, but it has the drawback that it is symmetric—so the downside and the upside look the same. This is at odds with evidence from many financial markets that show higher correlations across assets in down markets. The *Clayton copula density* is therefore an interesting alternative

$$c(u_1, u_2) = (-1 + u_1^{-\alpha} + u_2^{-\alpha})^{-2-1/\alpha} (u_1 u_2)^{-\alpha-1} (1 + \alpha), \quad (14.26)$$

where $\alpha \neq 0$. When $\alpha > 0$, then correlation on the downside is much higher than on the upside (where it goes to zero as we move further out in the tail). See Figure 14.14 for an illustration of how the copula density differs from the Gaussian copula. Also, see Figures 14.16–14.17 for illustrations of the resulting joint distribution (contour plots).

Remark 14.19 (*Multivariate Clayton copula density**) *The Clayton copula density for n variables is*

$$c(u) = (1 - n + \sum_{i=1}^n u_i^{-\alpha})^{-n-1/\alpha} (\prod_{i=1}^n u_i)^{-\alpha-1} (\prod_{i=1}^n [1 + (i-1)\alpha]).$$

With Clayton copula, the α parameter is related to Kendall's τ (denoted τ) according to

$$\text{Kendall's } \tau = \frac{\alpha}{\alpha + 2}, \text{ so} \quad (14.27)$$

$$\alpha = \frac{2\tau}{1 - \tau}. \quad (14.28)$$

The easiest way to *calibrate a Clayton copula* is therefore to set the parameter α according to (14.28), where Kendall's τ is estimated from the “data” on (u_1, u_2) . The latter are obtained by (1) estimate the marginal univariate distributions of x_1 and x_2 ; and (2) calculate the $u_{i,t} = \hat{F}_i(x_{i,t})$ values (and then estimate τ of those).

Figure 14.15 (right subfigure) illustrates how the probability of both variables to be below their respective quantiles depend on the α parameter. These parameters generate correlations similar to those for the Gaussian copula (see (14.19)–(14.20)) in the left sub-figure. The subfigures are therefore comparable—and the main point is that Clayton's

copula gives probabilities of joint low values (both variables being low) that do not decay as quickly as according to the Gaussian copulas. Intuitively, this means that the Clayton copula exhibits much higher “correlations” in the lower tail than the Gaussian copula does—although they imply the same overall correlation. That is, according to the Clayton copula more of the overall correlation of data is driven by synchronized movements in the lower tail. This could be interpreted as if the correlation is higher in market crashes than during normal times. See Figures 14.16–14.17 for illustrations.

Remark 14.20 (*Clayton copula function**) *The copula function (the cdf) corresponding to (14.26) is*

$$C(u_1, u_2) = (-1 + u_1^{-\alpha} + u_2^{-\alpha})^{-1/\alpha}$$

and with n components

$$C(u) = (1 - n + \sum_{i=1}^n u_i^{-\alpha})^{-1/\alpha}.$$

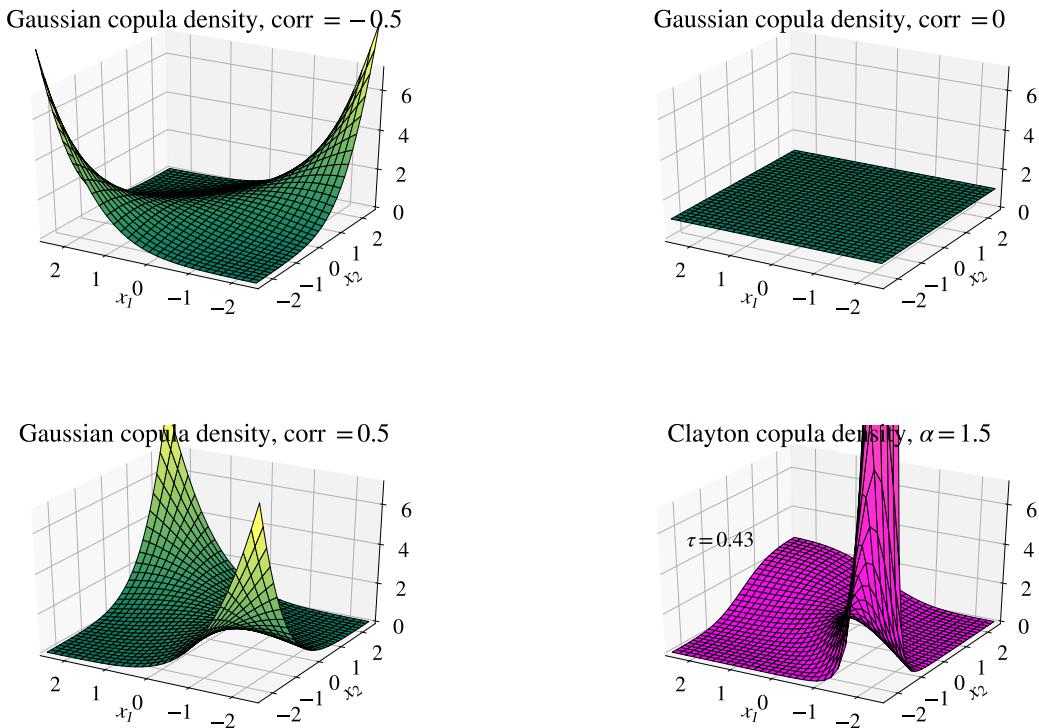


Figure 14.14: Copula densities (as functions of x_i)

Remark 14.21 (*Tail Dependence**) *The measure of lower tail dependence starts by finding the probability that X_1 is lower than its q th quantile ($X_1 \leq F_1^{-1}(q)$) given that X_2 is*

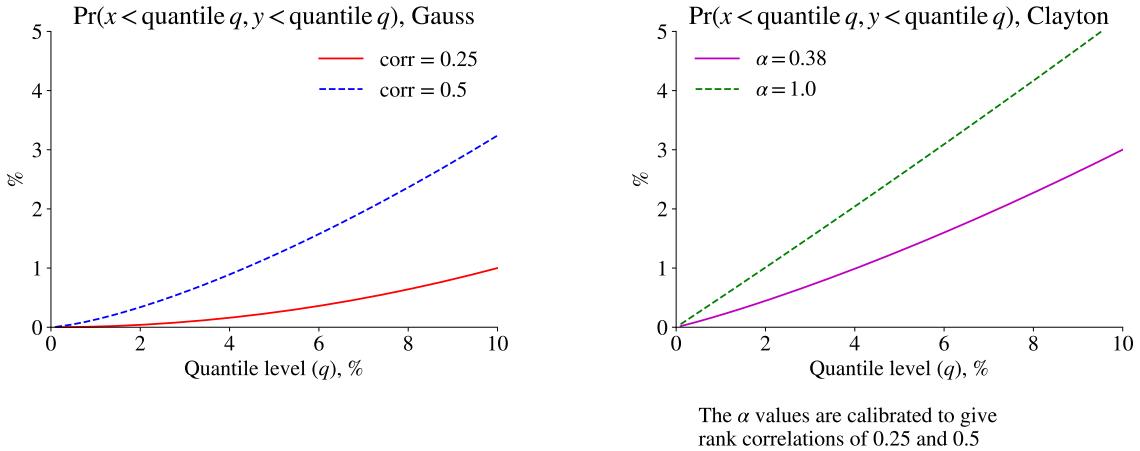


Figure 14.15: Probability of joint low returns, Gaussian and Clayton copulas

lower than its q th quantile ($X_2 \leq F_2^{-1}(q)$)

$$\Lambda_l = \Pr[X_1 \leq F_1^{-1}(q) | X_2 \leq F_2^{-1}(q)],$$

and then takes the limit as the quantile goes to zero

$$\lambda_l = \lim_{q \rightarrow 0} \Pr[X_1 \leq F_1^{-1}(q) | X_2 \leq F_2^{-1}(q)].$$

It can be shown that a Gaussian copula gives zero or very weak tail dependence, unless the correlation is 1. It can also be shown that the lower tail dependence of the Clayton copula is

$$\lambda_l = 2^{-1/\alpha} \text{ if } \alpha > 0$$

and zero otherwise.

14.6 Simulating Joint Distributions*

To find the implication for a portfolio of several assets (modelled using copula and perhaps tail distribution approaches), we often resort to simulations. That is, we draw random numbers (returns for each of the assets) from the joint tail distribution and then study the properties of the portfolio (with say, equal weights or whatever). The reason we simulate is that it is difficult to analytically calculate the distribution of the portfolio.

The approach proceeds in two steps. First, draw n values for the copula ($u_i, i = 1, \dots, n$). Second, calculate the random number (“return”) by inverting the marginal cdfs

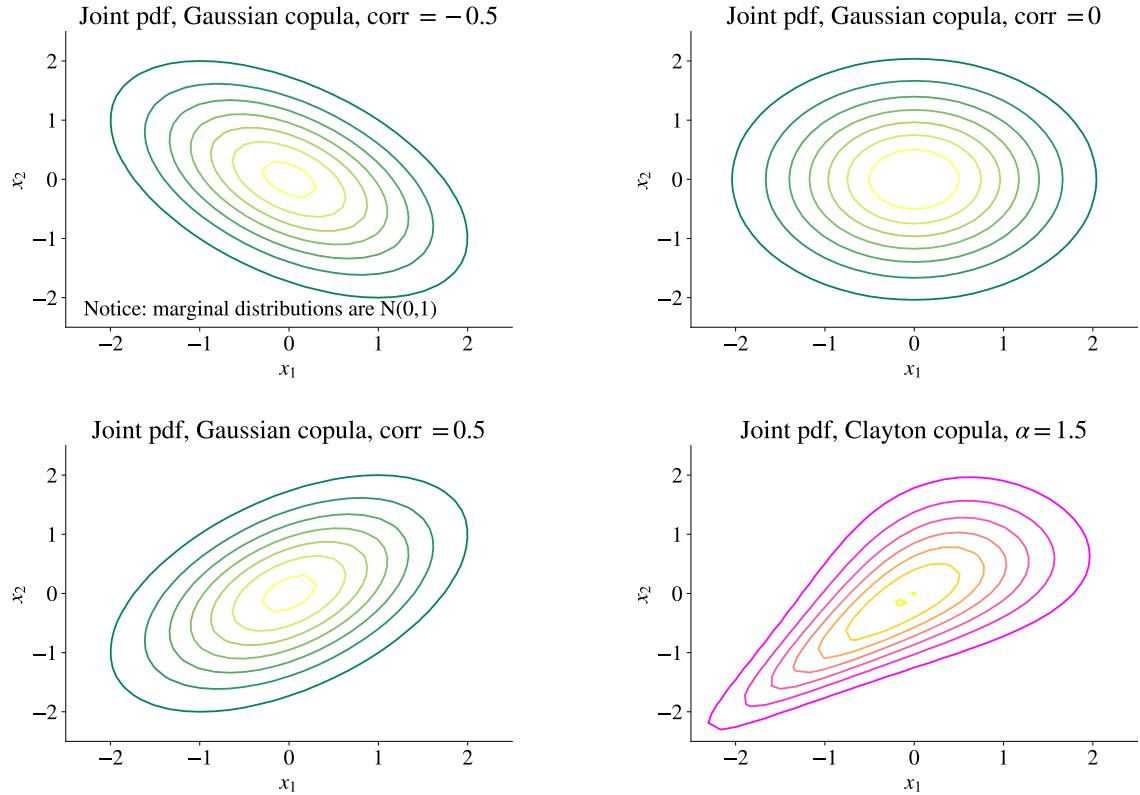


Figure 14.16: Contours of bivariate pdfs

$u_i = F_i(x_i)$ in (14.23) as

$$x_i = F_i^{-1}(u_i), \quad (14.29)$$

where $F_i^{-1}()$ is the inverse of the marginal cdf of x_i . The following remarks discuss the details.

Remark 14.22 (*To draw n random numbers from a Gaussian copula*) First, draw n numbers (v_i for $i = 1, \dots, n$) from an multivariate $N(\mathbf{0}, R)$ distribution, where R is the correlation matrix. Second, calculate $u_i = \Phi(v_i)$, where Φ is the cdf of a standard normal distribution. Third, calculate $x_i = F_i^{-1}(u_i)$ for each of the n variables. Notice that the parameters of the $F_i()$ distribution differ across i .

Remark 14.23 (*To draw n random numbers from a Clayton copula*) First, draw z_i for $i = 1, \dots, n$ from a uniform distribution (between 0 and 1). Second, draw X from a $\text{gamma}(1/\alpha, 1)$ distribution. Third, calculate $u_i = [1 - \ln(z_i)/X]^{-1/\alpha}$ for $i = 1, \dots, n$. Fourth, calculate $x_i = F_i^{-1}(u_i)$ for each of the n variables.

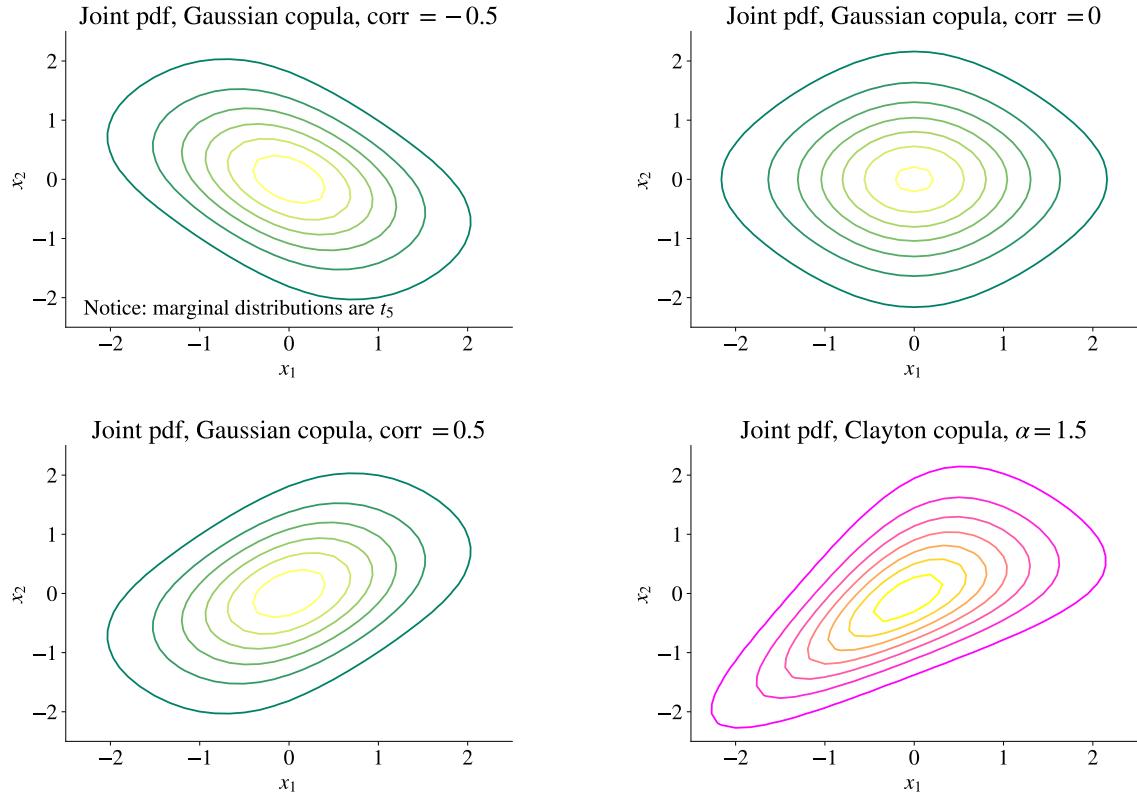


Figure 14.17: Contours of bivariate pdfs

Remark 14.24 (*Simulating returns from the GP distribution*) Consider the case of Remark 14.4 where the cdf for losses $x \leq h$ is $F(x) = J(x)P_h/J(h)$, but for $x > h$ it is $F(x) = P_h + G(x-h)(1-P_h)$. With u from the copula, let $\tilde{u} = 1-u$ (since u from the copula is for returns and we need to apply GP which is for losses). For $\tilde{u} \leq P_h$ use

$$\tilde{u} = J(x)P_h/J(h) \text{ and solve as } x = J^{-1}[\tilde{u}J(h)/P_h],$$

where $J^{-1}[]$ is the inverse of the $J()$ cdf. Instead, for $\tilde{u} > P_h$ set

$$\tilde{u} = P_h + G(x-h)(1-P_h) \text{ and solve as } x = h + G^{-1}[(\tilde{u}-P_h)/(1-P_h)],$$

where $G^{-1}[z]$ is the inverse of the GP cdf in Remark 14.2 (which is easily calculated). Once the loss X is simulated, the simulated return is just $R = -X$.

Such simulations can be used to quickly calculate the VaR and other risk measures for different portfolios. Figures 14.18–14.19 give an illustration of how the movements in the

lower tails get more synchronized as the α parameter in the Clayton copula increases.

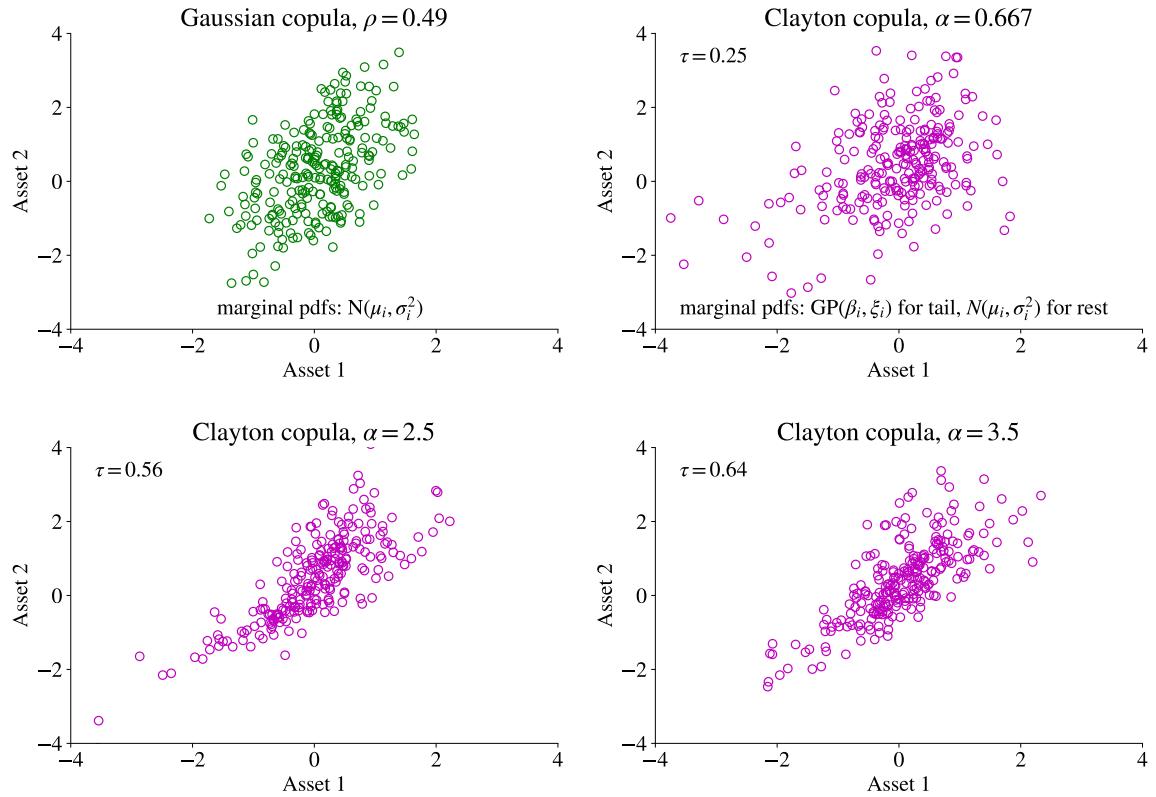


Figure 14.18: Example of scatter plots of two asset returns drawn from different copulas

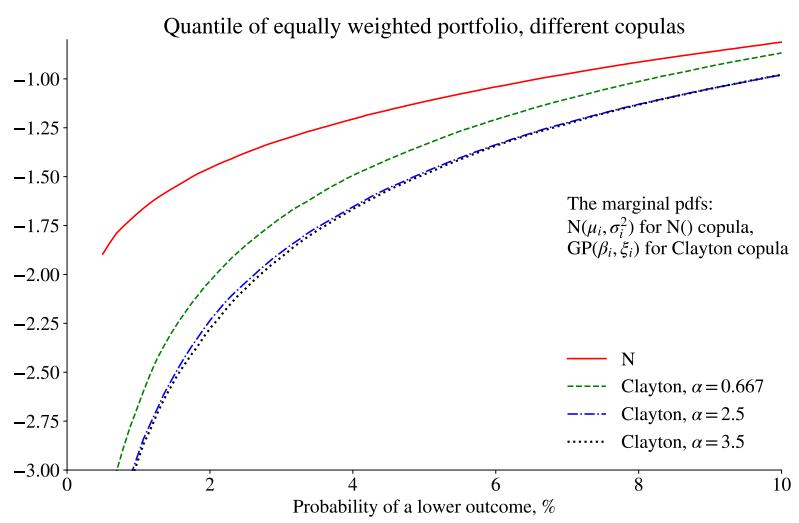


Figure 14.19: Quantiles of an equally weighted portfolio of two asset returns drawn from different copulas

Chapter 15

Option Pricing and Estimation of Continuous Time Processes

Reference: Hull (2006) 19, Elton, Gruber, Brown, and Goetzmann (2003) 22 or Bodie, Kane, and Marcus (2005) 21

Reference (advanced): Taylor (2005) 13–14; Campbell, Lo, and MacKinlay (1997) 9; Gourieroux and Jasiak (2001) 12–13

More advanced material is denoted by a star (*). It is not required reading.

15.1 The Black-Scholes Model

15.1.1 The Black-Scholes Option Price Model

A European call option contract traded (contracted and paid) in t may stipulate 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 $t + m$ at the strike price K . The option payoff (in $t + m$) is clearly $\max(0, S_{t+m} - K)$, where S_{t+m} is the asset price at expiration. See Figure 15.1 for the timing convention.

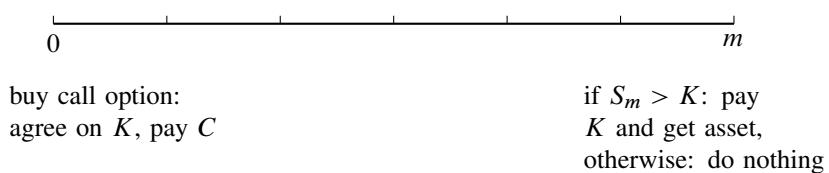


Figure 15.1: Timing convention of option contract

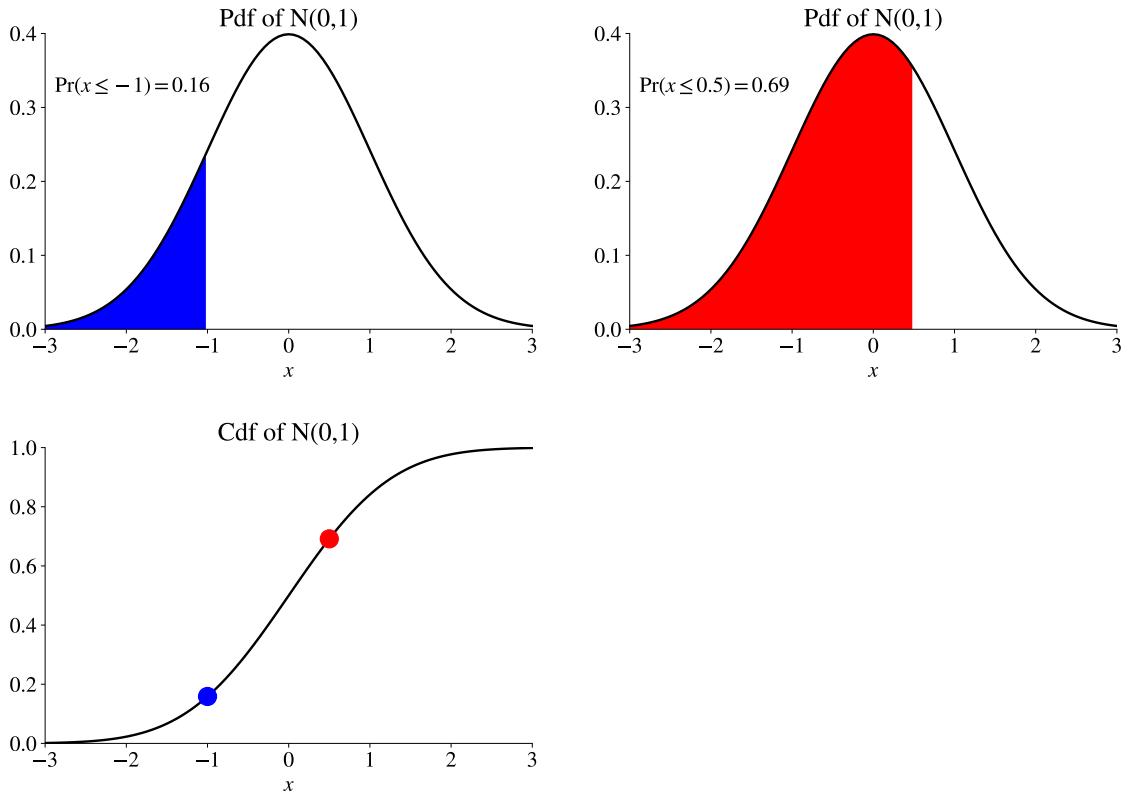


Figure 15.2: Pdf and cdf of $N(0,1)$

The Black-Scholes formula for a European call option price is

$$C_t = S_t \Phi(d_1) - K e^{-rm} \Phi(d_1 - \sigma \sqrt{m}), \text{ where} \quad (15.1)$$

$$d_1 = \frac{\ln(S_t/K) + (r + \sigma^2/2)m}{\sigma \sqrt{m}}.$$

where S_t is the price of the underlying asset in period t , and r is the continuously compounded interest rate. Also, $\Phi()$ is the cumulative distribution function of a standard normal, $N(0, 1)$, variable. For instance, $\Phi(2)$ is the probability that the variable is less or equal to two, see Figure 15.2.

Some basic properties of the model are illustrated in Figure 15.3. In particular, the call option price is increasing in the volatility and decreasing in the strike price.

The B-S formula can be derived from several stochastic processes of the underlying asset price (discussed below), but they all imply that the distribution of log asset price in $t + m$ (conditional on the information in t) is normal with some mean α (not important

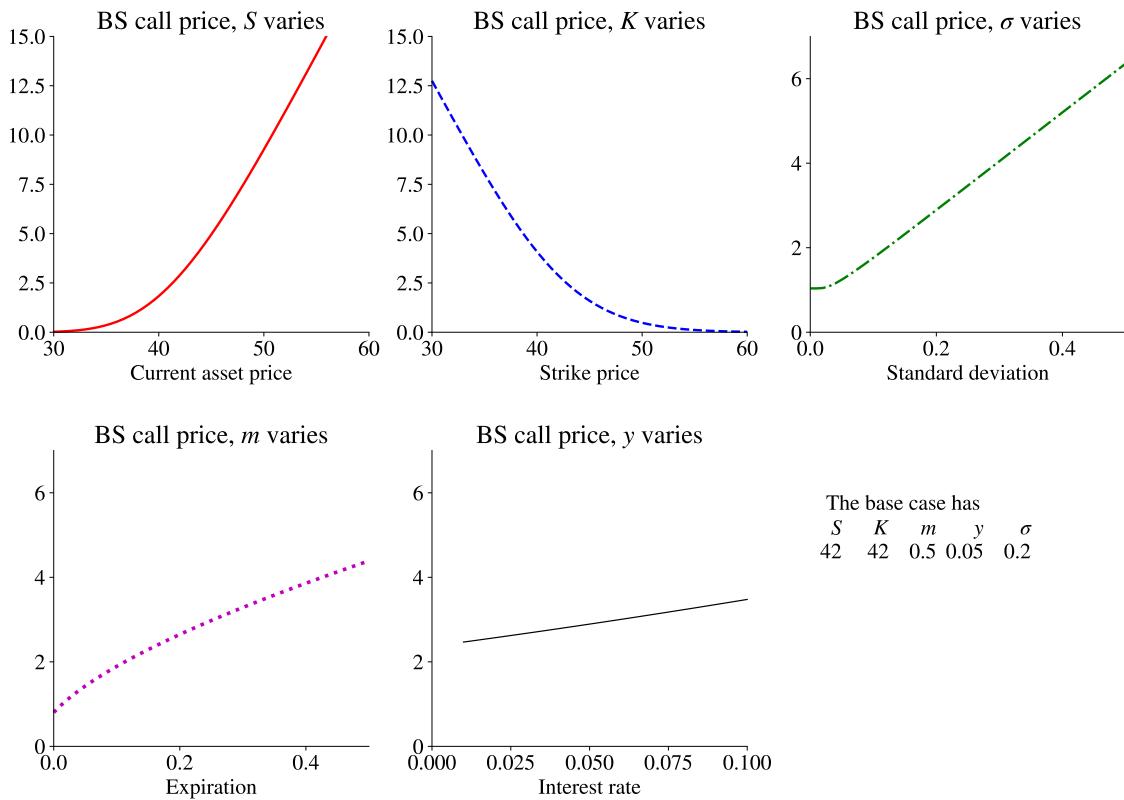


Figure 15.3: Call option price, Black-Scholes model

for the option price) and the variance $m\sigma^2$

$$\ln S_{t+m} \sim N(\alpha, m\sigma^2). \quad (15.2)$$

Option pricing is basically about forecasting the volatility (until expiration of the option) of the underlying asset. This is clear from the Black-Scholes model where the only unknown parameter is the volatility. It is also true more generally—which can be seen in at least two ways. First, a higher volatility is good for an owner of a call option since it increases the upside potential (higher probability of a really good outcome), at the same time as the down side is protected. Second, many option portfolios highlight how volatility matters for the potential profits. For instance, a straddle (a long position in both a call and a put at the same strike price) pays off if the price of the underlying asset moves a lot (in either direction) from the strike price, that is, when volatility is high. See Figures 15.4–15.5 for illustrations.

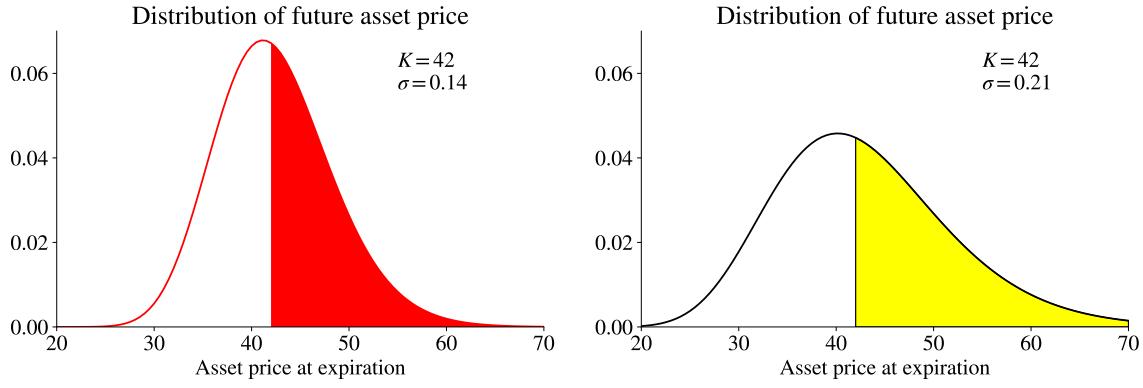


Figure 15.4: Distribution of future stock price

15.1.2 Implied Volatility

The pricing formula (15.1) contains only one unknown parameter: the standard deviation σ in the distribution of $\ln S_{t+m}$, see (15.2). 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 should not be thought of as an estimation of an unknown parameter—rather as just a transformation of the option price. Notice that we can solve (by trial-and-error or some numerical routine) for one implied volatility for each available strike price.

If the Black-Scholes formula is correct, that is, if the assumption in (15.2) is correct, then these volatilities should be the same across strike prices—and it should also be constant over time.

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. See Figures 15.6–15.7 for illustrations. 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 move considerably over time—see Figure 15.8 for a time series of implied volatilities

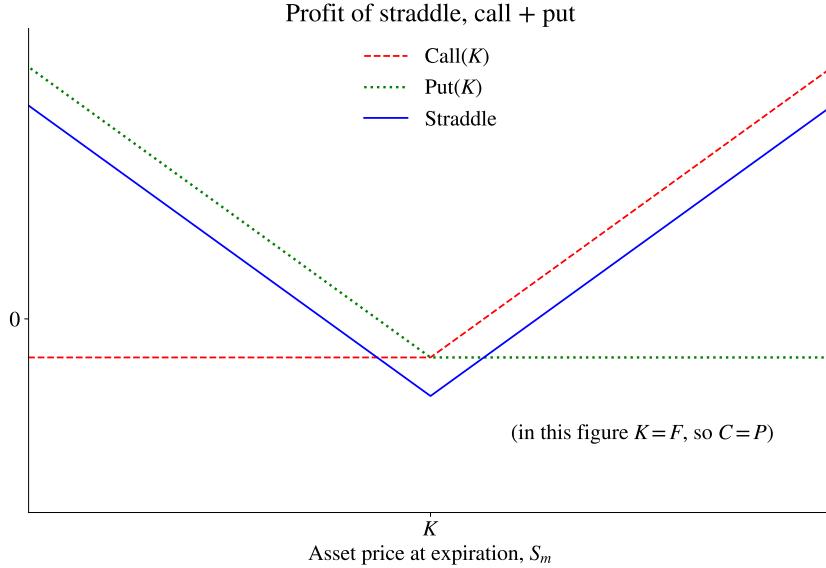


Figure 15.5: Profit of straddle portfolio

15.1.3 Brownian Motion without Mean Reversion: The Random Walk

The basic assumption behind the B-S formula (15.1) is that the log price of the underlying asset, $\ln S_t$, follows a geometric Brownian motion—with or without mean reversion.

This section discusses the standard geometric Brownian motion without mean reversion

$$d \ln S_t = \mu dt + \sigma dW_t, \quad (15.3)$$

where $d \ln S_t$ is the change in the log price (the return) over a very short time interval. On the right hand side, μ is the drift (typically expressed on annual basis), dt just indicates the change in time, σ is the standard deviation (per year), and dW_t is a random component (Wiener process) that has an $N(0, 1)$ distribution if we cumulate dW_t over a year ($\int_0^1 dW_t \sim N(0, 1)$). By comparing (15.1) and (15.3) we notice that only the volatility (σ), not the drift (μ), show up in the option pricing formula. In essence, the drift is already accounted for by the current spot price in the option pricing formula (as the spot price certainly depends on the expected drift of the asset price).

Remark 15.1 (Alternative stock price process*) *If we instead of (15.3) assume the process $dS_t = \tilde{\mu}S_t dt + \sigma S_t dW_t$, then we get the same option price. The reason is that Itô's lemma tells us that (15.3) implies this second process with $\tilde{\mu} = \mu + \sigma^2/2$. The difference is only in terms of the drift, which does not show up (directly, at least) in the B-S formula.*

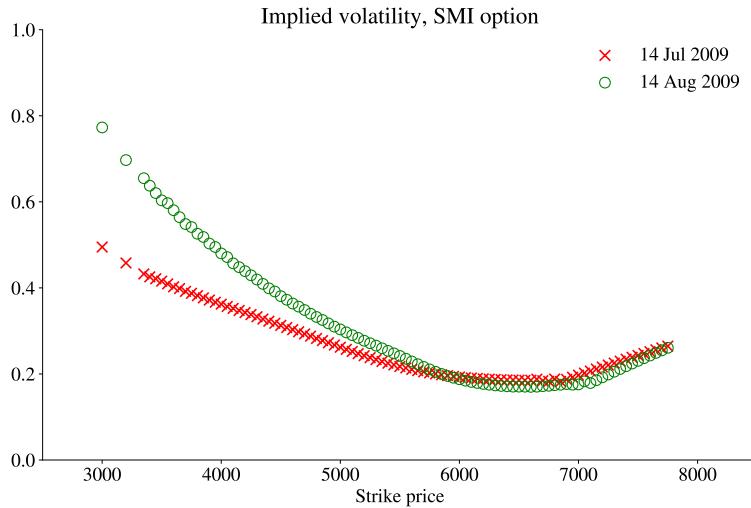


Figure 15.6: Implied volatilities of SMI options, selected dates

Remark 15.2 ((15.3) as a limit of a discrete time process*) (15.3) can be thought of as the limit of the discrete time process $\ln S_{t+h} - \ln S_t = \mu h + \sigma \sqrt{h} \varepsilon_{t+h}$ (where ε_t is iid $N(0, 1)$) as the time interval h becomes very small.

We can only observe the value of the asset price at a limited number of times, so we need to understand what (15.3) implies for discrete time intervals $(\tau - 1, \tau, \dots)$ which are h years apart (for instance, $h = 1/52$ with weekly data and $h = 252$ with daily data if we only count trading days). Let Δ denote the first difference, so $\Delta x_\tau = x_\tau - x_{\tau-1}$. Then (15.3) implies that

$$\Delta \ln S_\tau = \alpha + \varepsilon_\tau, \text{ where} \quad (15.4)$$

$$\varepsilon_\tau \text{ is iid } N(0, \sigma^2 h), \text{ and} \quad (15.5)$$

$$\alpha = \mu h. \quad (15.6)$$

That ε_τ is iid implies that $\text{Cov}(\varepsilon_\tau, \varepsilon_{\tau-1}) = 0$, which is the same as saying that $\text{Cov}(\Delta \ln S_\tau, \Delta \ln S_{\tau-1}) = 0$. Notice that both the drift and the variance scale linearly with the horizon h . The reason is that the growth rates are iid.

Example 15.3 Suppose μ and σ^2 are 0.1, and 0.25 respectively—and the periods are years. Equations (15.4)–(15.6) then gives the following for weekly ($h = 1/52$) data

$$\ln S_\tau - \ln S_{\tau-1} = 0.1/52 + \varepsilon_\tau \text{ with } \text{Var}(\varepsilon_\tau) = 0.25/52 \approx 0.0048.$$

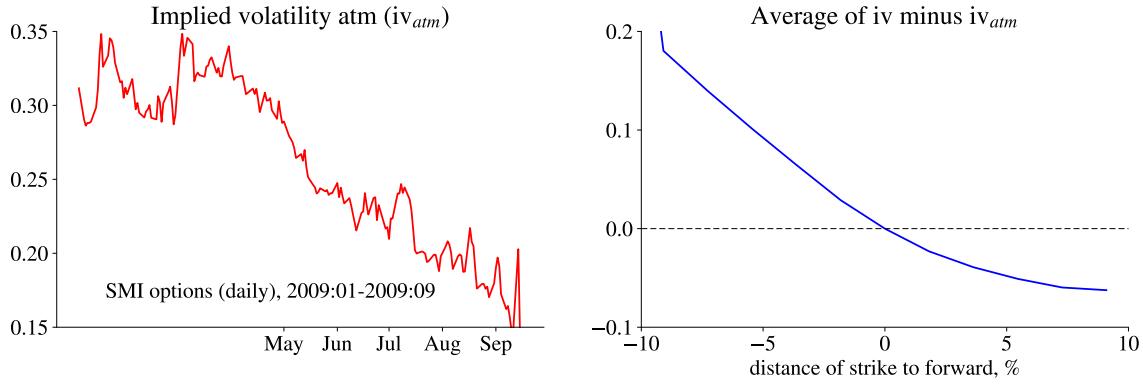


Figure 15.7: Implied volatilities

Remark 15.4 (*iid random variable in discrete time*) Suppose x_t has the constant mean μ and a variance σ^2 . Then $E(x_t + x_{t-1}) = 2\mu$ and $\text{Var}(x_t + x_{t-1}) = 2\sigma^2 + 2 \text{Cov}(x_t, x_{t-1})$. If x_t is iid, then the covariance is zero, so $\text{Var}(x_t + x_{t-1}) = 2\sigma^2$. In this case, both mean and variance scale linearly with the horizon.

15.1.4 Brownian Motion with Mean Reversion*

The mean reverting Ornstein-Uhlenbeck process is

$$d \ln S_t = \lambda(\mu - \ln S_t)dt + \sigma dW_t, \text{ with } \lambda > 0. \quad (15.7)$$

This process makes $\ln S_t$ revert back to the mean μ , and the mean reversion is faster if λ is large. It is used in, for instance, the Vasicek model of interest rates.

To estimate the parameters in (15.7) on real life data, we (once again) have to understand what the model implies for discretely sampled data (where $\tau - 1$ and τ are h years apart). It can be shown that it implies a discrete time AR(1)

$$\ln S_\tau = \alpha + \rho \ln S_{\tau-1} + \varepsilon_\tau, \text{ with} \quad (15.8)$$

$$\rho = e^{-\lambda h}, \alpha = \mu(1 - \rho), \text{ and } \varepsilon_t \sim N[0, \sigma^2(1 - \rho^2)/(2\lambda)]. \quad (15.9)$$

We know that the maximum likelihood estimator (MLE) of the discrete AR(1) is least squares combined with the traditional estimator of the residual variance. MLE has the further advantage of being invariant to parameter transformations, which here means that the MLE of λ , μ and σ^2 can be backed out from the LS estimates of ρ , α and $\text{Var}(\varepsilon_t)$ by using (15.9).

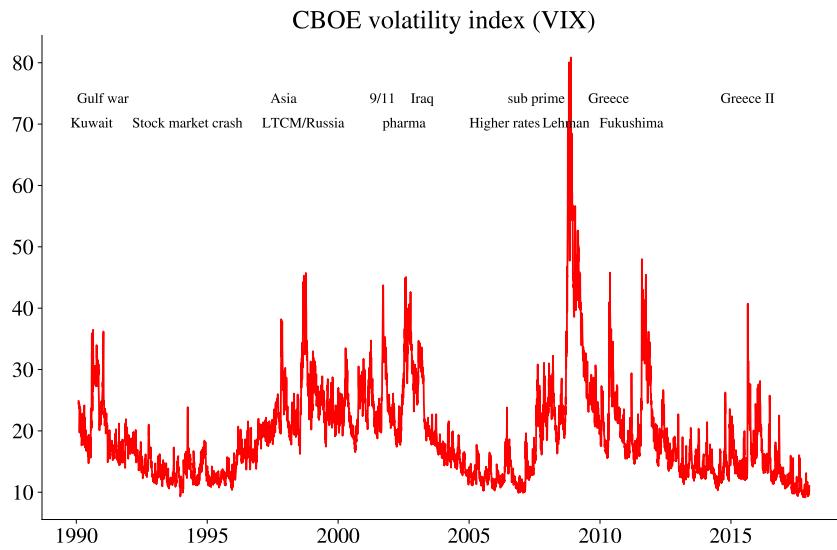


Figure 15.8: CBOE VIX, summary measure of implied volatilities (30 days) on US stock markets

Example 15.5 Suppose λ , μ and σ^2 are 2, 0, and 0.25 respectively—and the periods are years. Equations (15.8)–(15.9) then gives the following AR(1) for weekly ($h = 1/52$) data

$$\ln S_\tau = 0.96 \ln S_{\tau-1} + \varepsilon_\tau \text{ with } \text{Var}(\varepsilon_\tau) \approx 0.0046.$$

15.2 Estimation of the Volatility of a Random Walk Process

This section discusses different ways of estimating the volatility, since it is the key to option pricing. We will assume that we have data for observations in $\tau = 1, 2, \dots, n$. This could be 5-minute intervals, days, weeks or whatever. Let the time between τ and $\tau + 1$ be h (years). The sample therefore stretches over $T = nh$ periods (years). For instance, for daily data $h = 1/365$ (or possibly something like 1/252 if only trading days are counted). Instead, with weekly data $h = 1/52$. See Figure 15.9 for an illustration.

15.2.1 Standard Approach

We first estimate the variance for the sampling frequency we have, and then convert to the annual frequency.

According to (15.4) the growth rates, $\ln(S_t/S_{t-h})$, are iid over any sampling frequency. To simplify the notation, let $y_\tau = \ln(S_\tau/S_{\tau-1})$ be the observed growth rates.

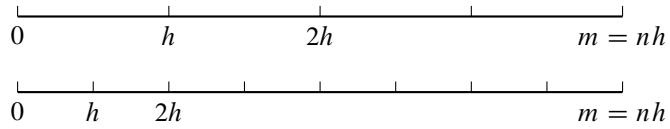


Figure 15.9: Two different samplings with same time span T

The classical estimator of the variance of an iid data series is

$$\hat{s}^2 = \sum_{\tau=1}^n (y_\tau - \bar{y})^2 / n, \text{ where} \quad (15.10)$$

$$\bar{y} = \sum_{\tau=1}^n y_\tau / n. \quad (15.11)$$

(This is also the maximum likelihood estimator.) To annualise these numbers, use

$$\hat{\sigma}^2 = \hat{s}^2 / h, \text{ and } \hat{\mu} = \bar{y} / h. \quad (15.12)$$

Example 15.6 If $(\bar{y}, \hat{s}^2) = (0.001, 0.03)$ on daily data, then the annualized values are $(\mu, \sigma^2) = (0.001 \times 250, 0.03 \times 250) = (0.25, 7.5)$ if we assume 250 trading days per year.

Notice that it can be quite important to subtract the mean drift, \bar{y} . Recall that for any random variable, we have

$$\sigma^2 = E(x^2) - \mu^2, \quad (15.13)$$

so a non-zero mean drives a wedge between the variance (which we want) and the second moment (which we estimate if we assume $\bar{y} = 0$).

Example 15.7 (US stock market volatility) For the US stock market index excess return since WWII we have approximately a variance of 0.16^2 and a mean of 0.08. In this case, (15.13) becomes

$$0.16^2 = E(x^2) - 0.08^2, \text{ so } E(x^2) \approx 0.18^2.$$

Assuming that the drift is zero gives an estimate of the variance equal to 0.18^2 which is 25% too high.

Remark 15.8 (*Variance vs second moment, the effect of the maturity) Suppose we are interested in the variance over an m -period horizon, for instance, because we want to

price an option that matures in $t + m$. How important is it then to use the variance ($m\sigma^2$) rather than the second moment? The relative error is

$$\frac{\text{Second moment - variance}}{\text{variance}} = \frac{m^2\mu^2}{m\sigma^2} = \frac{m\mu^2}{\sigma^2},$$

where we have used the fact that the second moment equals the variance plus the squared mean (cf (15.13)). Clearly, this relative exaggeration is zero if the mean is zero. The relative exaggeration is small if the maturity is small.

If we have high frequency data on the asset price or the return, then we can choose which sampling frequency to use in (15.10)–(15.11). Recall that a sample with n observations (where the length of time between the observations is h) covers $T = nh$ periods (years). It can be shown that the asymptotic variances (that is, the variances in a very large sample) of the estimators of μ and σ^2 in (15.10)–(15.12) are

$$\text{Var}(\hat{\mu}) = \sigma^2/T \text{ and } \text{Var}(\hat{\sigma}^2) = 2\sigma^4/n. \quad (15.14)$$

Therefore, to get a precise estimator of the mean drift, μ , we need a sample that stretches over a long period: it does not help to just sample more frequently. However, the sampling frequency is crucial for getting a precise estimator of σ^2 , while a sample that stretches over a long period is less important. For estimating the volatility (to use in the B-S model) we should therefore use high-frequency data.

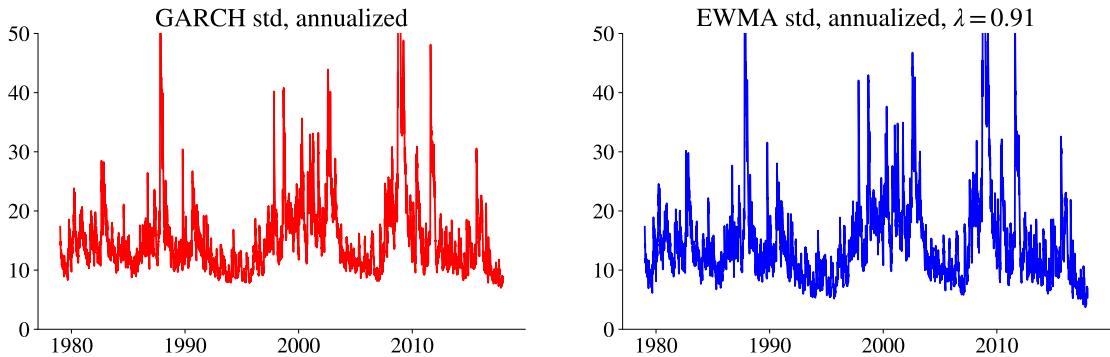
15.2.2 Exponentially Weighted Moving Average

The traditional estimator is based on the assumption that volatility is constant—which is consistent with the assumptions of the B-S model. In reality, volatility is time varying.

A practical ad hoc approach to estimate time varying volatility is to modify (15.10)–(15.11) so that recent observations carry larger weight. The exponentially weighted moving average (EWMA) model lets the weight for lag s be $(1 - \lambda)\lambda^s$ where $0 < \lambda < 1$. If we assume that \bar{y} is the same in all periods, then we have

$$\hat{s}_\tau^2 = \lambda \hat{s}_{\tau-1}^2 + (1 - \lambda)(y_{\tau-1} - \bar{y})^2, \quad (15.15)$$

where τ is the current period and $\tau - 1$ the previous period (say, today and yesterday). Clearly, a higher λ means that old data plays a larger role—and at the limit as λ goes towards one, we have the traditional estimator. See Figure 15.10 for a comparison using daily US equity returns. This method is commonly used by practitioners. For instance,



Daily S&P 500 returns 1979:01-2017:12
 AR(1) of excess returns with GARCH(1,1) errors
 $\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2$
 estimate
 α 0.08
 β 0.91

$$\sigma_t^2 = (1 - \lambda) u_{t-1}^2 + \lambda \sigma_{t-1}^2$$

Figure 15.10: Different estimates of US equity market volatility

the RISK Metrics is based on $\lambda = 0.94$ on daily data. Alternatively, λ can be chosen to minimize some criterion function.

Remark 15.9 (*EWMA with time-variation in the mean**) If we want also the mean to be time-varying, then we can use the estimator

$$\hat{s}_\tau^2 = (1 - \lambda) [(y_{\tau-1} - \bar{y}_\tau)^2 + \lambda (y_{\tau-2} - \bar{y}_\tau)^2 + \lambda^2 (y_{\tau-3} - \bar{y}_\tau)^2 + \dots]$$

$$\bar{y}_\tau = [y_{\tau-1} + y_{\tau-2} + y_{\tau-3} + \dots] / (\tau - 1).$$

Notice that the mean is estimated as a traditional sample mean, using observations 1 to $\tau - 1$. This guarantees that the variance will always be a non-negative number.

It should be noted, however, that the B-S formula is, strictly speaking, not compatible with time-varying volatility. Still, it can be used as an approximation.

15.2.3 Autoregressive Conditional Heteroskedasticity

The model with Autoregressive Conditional Heteroskedasticity (ARCH) is a useful tool for estimating the properties of volatility clustering. The first-order ARCH expresses volatility as a function of the latest squared shock

$$s_\tau^2 = \alpha_0 + \alpha_1 u_{\tau-1}^2, \quad (15.16)$$

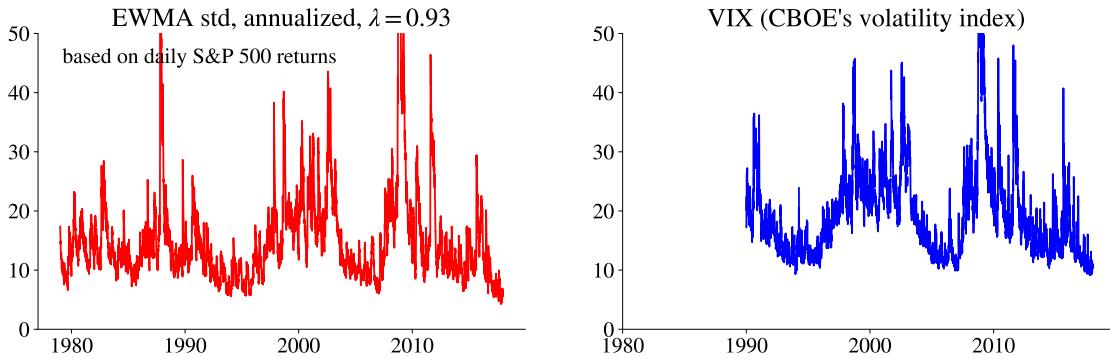


Figure 15.11: Different estimates of US equity market volatility

where u_τ is a zero-mean variable. The model requires $\alpha_0 > 0$ and $0 \leq \alpha_1 < 1$ to guarantee that the volatility stays positive and finite. The variance reverts back to an average variance ($\alpha_0/(1 - \alpha_1)$). The rate of mean reversion is α_1 , that is, the variance behaves much like an AR(1) model with an autocorrelation parameter of α_1 . The model parameters are typically estimated by maximum likelihood. Higher-order ARCH models include further lags of the squared shocks (for instance, $u_{\tau-2}^2$).

Instead of using a high-order ARCH model, it is often convenient to use a first-order generalized ARCH model, the GARCH(1,1) model. It adds a term that directly captures autoregression of the volatility

$$s_\tau^2 = \alpha_0 + \alpha_1 u_{\tau-1}^2 + \beta_1 s_{\tau-1}^2. \quad (15.17)$$

We require that $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, and $\alpha_1 + \beta_1 < 1$ to guarantee that the volatility stays positive and finite. This is very similar to the EWMA in (15.15), except that the variance reverts back to the mean ($\alpha_0/(1 - \alpha_1 - \beta_1)$). The rate of mean reversion is $\alpha_1 + \beta_1$, that is, the variance behaves much like an AR(1) model with an autocorrelation parameter of $\alpha_1 + \beta_1$.

15.2.4 Time-Variation in Volatility and the B-S Formula

The ARCH and GARCH models imply that volatility is random, so they are (strictly speaking) not consistent with the B-S model. However, they are often combined with the B-S model to provide an approximate option price. See Figure 15.12 for a comparison of the actual distribution of the log asset price (actually, cumulated returns, so assuming that the initial log asset price is zero) at different horizons (1 and 10 days) when the daily

returns are generated by a GARCH model—and a normal distribution with the same mean and variance. To be specific, the figure shows the distribution of the future log asset price calculated as

$$\ln S_{t+m} = \ln S_t + r_{t+h}, \text{ or} \quad (15.18)$$

$$= \ln S_t + \sum_{i=1}^{10} r_{t+ih}, \quad (15.19)$$

where each of the returns (r_{t+ih}) is drawn from an $N(0, s_{t+ih}^2)$ distribution where the variance follows the GARCH(1,1) process like in (15.17).

It is clear the normal distribution is a good approximation unless the ARCH component ($\alpha_1 \times$ lagged squared shock) dominates the GARCH component ($\beta_1 \times$ lagged variance).

Intuitively, we get (almost) a normal distribution when the random part of the volatility (the ARCH component) is relatively small compared to the non-random part (the GARCH component). For instance, if there is no random part at all, then we get exactly a normal distribution (the sum of independent normally distributed variables is normally distributed—if all the variances are deterministic).

However, to get an option price that is perfectly consistent with a GARCH process, we need to go beyond the B-S model (see, for instance, Heston and Nandi (2000)).

Remark 15.10 (*Time-varying, but deterministic volatility**) A time-varying, but non-random volatility could be consistent with (15.2): if $\ln S_{t+m}$ is the sum (integral) of normally distributed changes with known (but time-varying variances), then this sum has a normal distribution (recall: if the random variables x and y are normally distributed, so is $x+y$). A random variance does not fit this case, since a variable with a random variance is not normally distributed.

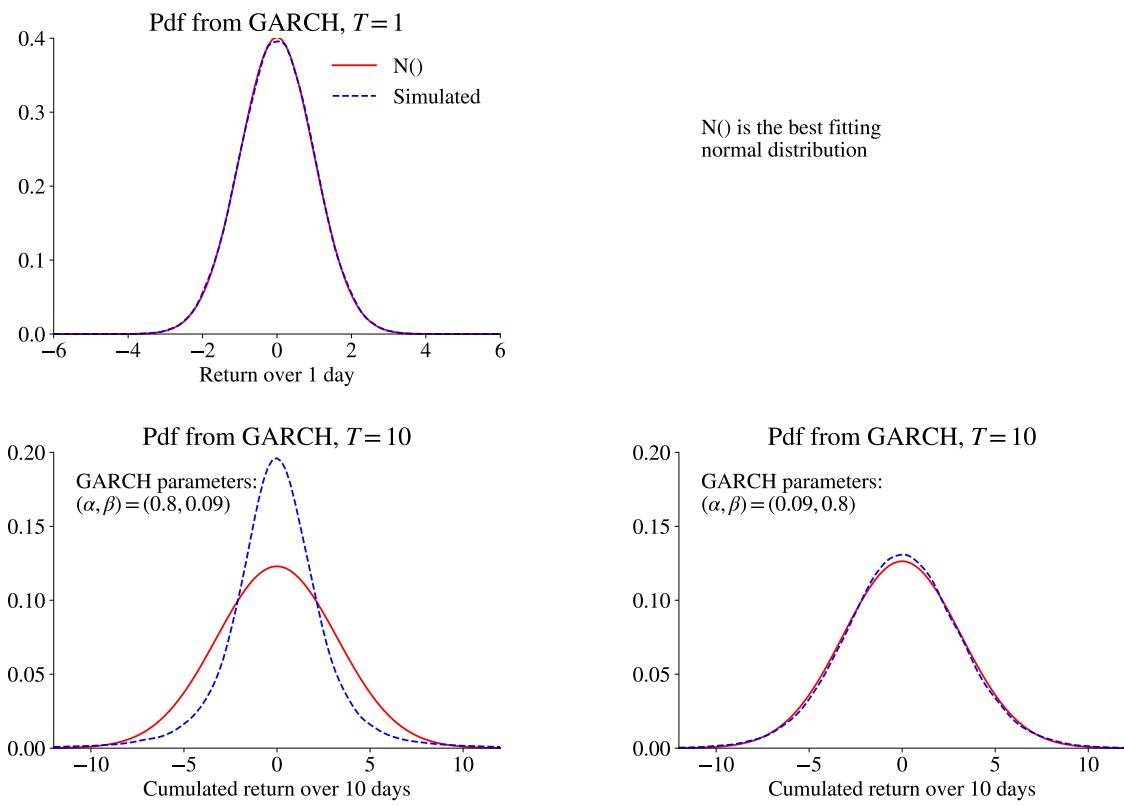


Figure 15.12: Comparison of normal and simulated distribution of m -period returns

Chapter 16

Event Studies

Reference: Bodie, Kane, and Marcus (2005) 12.3 or Copeland, Weston, and Shastri (2005) 11

Reference (advanced): Campbell, Lo, and MacKinlay (1997) 4

More advanced material is denoted by a star (*). It is not required reading.

16.1 Basic Structure of Event Studies

The idea of an event study is to study the effect (on stock prices or returns) of a special event by using a cross-section of such events. For instance, what is the average (across firms) 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 not or if the monetary policy surprise was positive (lower interest rates than expected) or not. 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 (some average of) analysts' forecast. Similarly, for monetary policy moves, the event could be specified as the interest rate decision minus previous forward rates (as a measure of previous expectations).

Similarly, we typically study the *abnormal return* around such events, defined as (for asset i in period t)

$$u_{it} = R_{it} - R_{it}^{normal}, \quad (16.1)$$

where R_{it} is the actual return and the last term is the normal return (which may differ

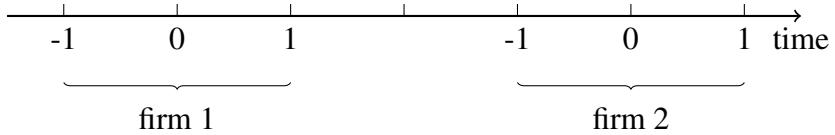


Figure 16.1: Event days and windows

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 (the normal return).

Suppose we have a sample of n such events (“assets”). To keep the notation (reasonably) simple, 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 16.1 for an illustration.

To control for 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

$$\bar{u}_s = \sum_{i=1}^n u_{is}/n. \quad (16.2)$$

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 (16.1) over some period around the event. It is often calculated from the beginning of the event window (or from day 0). For instance, if the event window starts at -20 , then the 3-period (day?) car for firm i is

$$\text{car}_{i3} = u_{i,-20} + u_{i,-19} + u_{i,-18}. \quad (16.3)$$

More generally, if the event window starts at w (say, -20), then the q -period car for firm i is

$$\text{car}_{iq} = \sum_{\tau=w}^{w+q-1} u_{i,\tau}. \quad (16.4)$$

The cross sectional average of the q -period car is

$$\overline{\text{car}}_q = \sum_{i=1}^n \text{car}_{iq} / n. \quad (16.5)$$

See Figure 16.2 for an empirical example.

Example 16.1 (*Abnormal returns for ± 1 day around 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

Time	Firm 1	Firm 2	Cross-sectional Average
-1	0.2	-0.1	0.05
0	1.0	2.0	1.5
1	0.1	0.3	0.2

We have the following cumulative returns

Time	Firm 1	Firm 2	Cross-sectional Average
-1	0.2	-0.1	0.05
0	1.2	1.9	1.55
1	1.3	2.2	1.75

16.2 Models of Normal Returns

This section summarizes the most common ways of calculating the normal return in (16.1). The parameters in these models are typically estimated on a recent sample, the “estimation window,” which ends before the event window. See Figure 16.3 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. This allows us to get a clean estimate of the normal return.

The *constant mean return model* assumes that the return of asset i fluctuates randomly

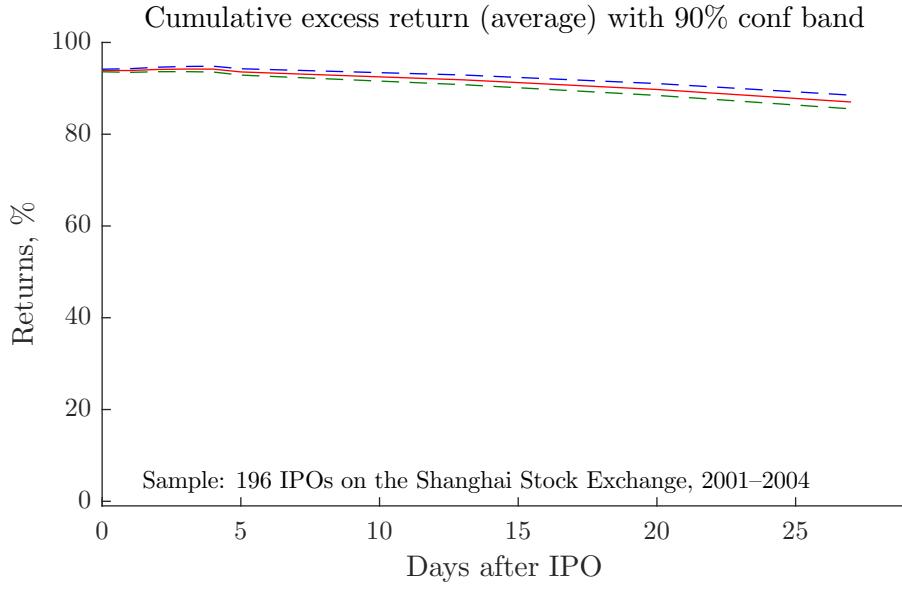


Figure 16.2: Event study of IPOs in Shanghai 2001–2004. (Data from Nou Lai.)

around some mean μ_i

$$R_{it} = \mu_i + \varepsilon_{it} \text{ with} \\ E(\varepsilon_{it}) = 0 \text{ and } \text{Cov}(\varepsilon_{it}, \varepsilon_{i,t-s}) = 0.$$
(16.6)

This mean is estimated by the sample average (during the estimation window). The normal return in (16.1) is then the estimated mean, $\hat{\mu}_i$. During the event window, we calculate the abnormal return as

$$u_{it} = R_{it} - \hat{\mu}_i. \quad (16.7)$$

The standard error of this 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

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it} \text{ with} \\ E(\varepsilon_{it}) = 0, \text{Cov}(\varepsilon_{it}, \varepsilon_{i,t-s}) = 0 \text{ and } \text{Cov}(\varepsilon_{it}, R_{mt}) = 0.$$
(16.8)

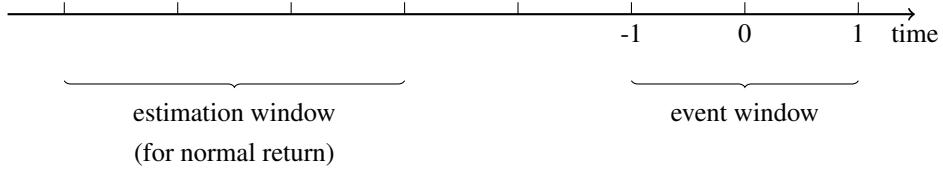


Figure 16.3: Event and estimation windows

Notice that we typically do not impose the CAPM restrictions on the intercept in (16.8). The normal return in (16.1) 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

$$u_{it} = R_{it} - \hat{\alpha}_i - \hat{\beta}_i R_{mt}. \quad (16.9)$$

The standard error of this is estimated by the standard error of $\hat{\varepsilon}_{it}$ (in the estimation window).

Recently, the market model has increasingly been replaced by a multi-factor model which uses several regressors instead of only the market return. For instance, Fama and French (1993) argue that (16.8) needs to be augmented by a portfolio that captures the different returns of small and large firms and also by a portfolio that captures the different returns of firms with high and low book-to-market ratios.

When we restrict $\alpha_i = 0$ and $\beta_i = 1$ in (16.8), 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 the abnormal return as

$$u_{it} = R_{it} - R_{mt} \quad (16.10)$$

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 a normal return as the actual return on assets which are very 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). In this case, the abnormal return becomes the difference between the actual return and the average return on the matching portfolio. This type of *matching portfolio* is becoming increasingly popular. For the event window we calculate the abnormal return

as

$$u_{it} = R_{it} - R_{pt}, \quad (16.11)$$

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 mode, 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.

To illustrate the importance of the model for normal returns, consider the market model (16.8). Under the null hypothesis that the event has no effect on the return, the abnormal return would be just the residual in the regression (16.8). It has the variance (assuming we know the model parameters)

$$\text{Var}(u_{it}) = \text{Var}(\varepsilon_{it}) = (1 - R^2) \text{Var}(R_{it}), \quad (16.12)$$

where R^2 is the coefficient of determination of the regression (16.8).

Proof. (of (16.12)) Recall that R^2 is defined as

$$R^2 = 1 - \frac{\text{Var}(\varepsilon_{it})}{\text{Var}(R_{it})}.$$

Rearrange to get (16.12). ■

This variance is crucial for testing the hypothesis of no abnormal returns: the smaller is the variance, the easier it is to reject a false null hypothesis (see Section 16.3). 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 $\sqrt{1 - 0.75} = 0.5$). More realistically, R^2 might be 0.43 (or less), so the market model gives a 25% decrease in the standard deviation. 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.

16.3 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 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* (across the n assets) *average*, \bar{u}_s in (16.2), is then

$$\text{Var}(\bar{u}_s) = (\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2) / n^2 = \sum_{i=1}^n \sigma_i^2 / n^2, \quad (16.13)$$

since all covariances are assumed to be zero. In a large sample (where the asymptotic normality of a sample average starts to kick in), we can therefore use a t -test

$$\bar{u}_s / \text{Std}(\bar{u}_s) \xrightarrow{d} N(0, 1). \quad (16.14)$$

In most applications the σ_i^2 values used in (16.13) are from the estimation window, as discussed in Section 16.2.

Remark 16.2 (*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 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 (16.13) and σ_{CR}^2 / n are conceptually the same. (The extension to cumulative returns over q period is straightforward: just multiply the one-period variance by q .) 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 speaks in favour of using (16.13) and a large estimation window for each of the assets.

The *cumulative abnormal return* over q period, $car_{i,q}$, can also be tested with a t -test. Since the returns are assumed to have no autocorrelation the variance of the $car_{i,q}$

$$\text{Var}(car_{i,q}) = q\sigma_i^2. \quad (16.15)$$

This variance is increasing in q since we are considering cumulative returns (not the time average of returns).

If the abnormal returns are uncorrelated across time, then the *cross-sectional average car_{i,q}* is

$$\text{Var}(\bar{car}_q) = q \text{Var}(\bar{u}_s), \quad (16.16)$$

where $\text{Var}(\bar{u}_s)$ is defined in (16.13).

Example 16.3 (Variances of abnormal returns) If the standard deviations of the daily abnormal returns of the two firms in Example 16.1 are $\sigma_1 = 0.1$ and $\sigma_2 = 0.2$, then we have the following variances for the abnormal returns at different days

Time	Firm 1	Firm 2	Cross-sectional Average
-1	0.1^2	0.2^2	$(0.1^2 + 0.2^2) / 4$
0	0.1^2	0.2^2	$(0.1^2 + 0.2^2) / 4$
1	0.1^2	0.2^2	$(0.1^2 + 0.2^2) / 4$

Similarly, the variances for the cumulative abnormal returns are

Time	Firm 1	Firm 2	Cross-sectional Average
-1	0.1^2	0.2^2	$(0.1^2 + 0.2^2) / 4$
0	2×0.1^2	2×0.2^2	$2 \times (0.1^2 + 0.2^2) / 4$
1	3×0.1^2	3×0.2^2	$3 \times (0.1^2 + 0.2^2) / 4$

Example 16.4 (Tests of abnormal returns) By dividing the numbers in Example 16.1 by the square root of the numbers in Example 16.3 (that is, the standard deviations) we get

the test statistics for the abnormal returns

<u>Time</u>	<u>Firm 1</u>	<u>Firm 2</u>	<u>Cross-sectional Average</u>
-1	2	-0.5	0.4
0	10	10	13.4
1	1	1.5	1.8

Similarly, the variances for the cumulative abnormal returns we have

<u>Time</u>	<u>Firm 1</u>	<u>Firm 2</u>	<u>Cross-sectional Average</u>
-1	2	-0.5	0.4
0	8.5	6.7	9.8
1	7.5	6.4	9.0

Remark 16.5 (*Normalized cumulative abnormal returns**). We now consider a normalization of the results whereby we report the cumulative abnormal return from the start of the event window until period s , denoted $\text{car}(-w, s)$, minus the cumulative abnormal return up to and including the event day, denoted $\text{car}(-w, 0)$,

$$\text{car}^n(-w, s) = \text{car}(-w, s) - \text{car}(-w, 0).$$

For instance, with $(w = -2)$ we clearly have

$$\begin{aligned}\text{car}^n(-2, -2) &= u_{-2} - (u_{-2} + u_{-1} + u_0) = -(u_{-1} + u_0) \\ \text{car}^n(-2, -1) &= u_{-2} + u_{-1} - (u_{-2} + u_{-1} + u_0) = -u_0 \\ \text{car}^n(-2, 0) &= 0 \\ \text{car}^n(-2, 1) &= u_{-2} + u_{-1} + u_0 + u_1 - (u_{-2} + u_{-1} + u_0) = u_1 \\ \text{car}^n(-2, 2) &= u_{-2} + u_{-1} + u_0 + u_1 + u_2 - (u_{-2} + u_{-1} + u_0) = u_1 + u_2.\end{aligned}$$

If the abnormal returns are uncorrelated across time, then it is clear that the variances of the $\text{car}^n(-w, s)$ is zero on the event day and scales with $|s|$ for the days (fanning out from $s = 0$)

$$\text{Var}[\text{car}^n(-w, s)] = |s|\sigma^2.$$

16.3.1 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

$$\bar{u}_s = (u_{1s} + u_{2s})/2, \quad (16.17)$$

where u_{is} is the abnormal return on asset i on day s . In general, the variance of this average is

$$\text{Var}(\bar{u}_s) = (\sigma_1^2 + \sigma_2^2 + 2\sigma_{12})/4, \quad (16.18)$$

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 refer 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

$$\text{Var}(\bar{u}_s) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}' \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} / 4, \quad (16.19)$$

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

$$\text{Var}(\bar{u}_s) = \mathbf{1}' \Sigma \mathbf{1} / n^2, \quad (16.20)$$

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 (16.13) and (16.16) with (16.20). The rest of the analysis is unchanged.

16.4 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)

$$\text{car}_{iq} = a + bx_i + \xi_i. \quad (16.21)$$

The slope coefficient is then a measure of how much the cumulative abnormal return reacts to a change of one unit of x_i .

Chapter 17

Kernel Density Estimation and Regression

17.1 Non-Parametric Regression

Reference: Campbell, Lo, and MacKinlay (1997) 12.3; Härdle (1990); Pagan and Ullah (1999); Mittelhammer, Judge, and Miller (2000) 21

17.1.1 Simple Kernel Regression

Non-parametric regressions are used when we are unwilling to impose a parametric form on the regression equation—and we have a lot of data.

Let the scalars y_t and x_t be related as

$$y_t = b(x_t) + \varepsilon_t, \quad (17.1)$$

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 will give quick estimates, but 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. The approach in this section is more “local” by down weighting information from data points where x_t is far from x .

Suppose the sample had 3 observations (say, $t = 3, 27$, and 99) with exactly the same value of x_t , say 1.9. A natural way of estimating $b(x)$ at $x = 1.9$ would then be to average over these 3 observations as we can expect average of the error terms to be close to zero (iid and zero mean).

Unfortunately, we seldom have repeated observations of this type. Instead, we may try to approximate the value of $b(x)$ (x is a single value, 1.9, say) by averaging over

observations where x_t is close to x . The general form of this type of estimator is

$$\hat{b}(x) = \frac{\sum_{t=1}^T w(x_t - x)y_t}{\sum_{t=1}^T w(x_t - x)}, \quad (17.2)$$

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 (17.2) is that the $b(x)$ function is smooth so local averaging (around x) makes sense.

As an example of a $w(\cdot)$ function, it could give equal weight to the k values of x_t which are closest to x and zero weight 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, $E[d^2\hat{b}(x)/dx^2]^2$. This defines a cubic spline (and is often used in macroeconomics, where $x_t = t$ and is then called the Hodrick-Prescott filter).

A *Kernel regression* uses a probability density function (pdf) as the weight function $w(\cdot)$. The perhaps simplest choice is a uniform density function over $x - h/2$ to $x + h/2$ (and zero outside this interval). In this case, the weighting function is

$$w(x_t - x) = \frac{1}{h} \delta \left(\left| \frac{x_t - x}{h} \right| \leq 1/2 \right), \text{ where } \delta(q) = \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases} \quad (17.3)$$

This weighting function puts the weight $1/h$ on all data points in the interval $x \pm h/2$ and zero on all other data points. See Figure 17.1 for an illustration of the weights.

However, we can gain efficiency and get a smoother (across x values) 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. The pdf of $N(x, h^2)$ is commonly used as a kernel, where the choice of h allows us to easily vary the relative weights of different observations. With this particular kernel, we get the following weights (for estimation of $b(x)$ at point x , for instance, $x = 1.9$)

$$w(x_t - x) = \frac{\exp \left[- \left(\frac{x_t - x}{h} \right)^2 / 2 \right]}{h \sqrt{2\pi}}. \quad (17.4)$$

This weighting function is positive so all observations get a positive weight, but the weights are highest for observations close to x and then tapers off in a bell-shaped way. See Figure 17.1 for an illustration.

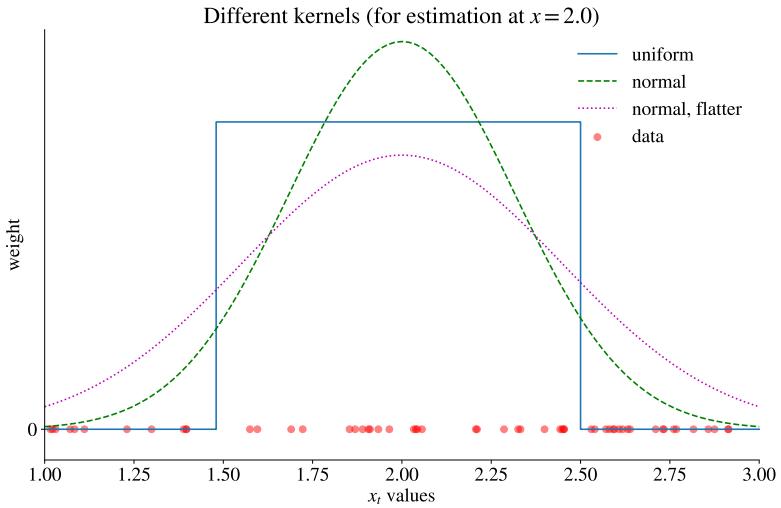


Figure 17.1: Different weighting functions for non-parametric regression

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 17.3 for an illustration and Figure 17.5 for an example.

Example 17.1 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 $x = 1.9$. With $h = 1$, the numerator in (17.4) is

$$\begin{aligned} \sum_{t=1}^T w(x_t - x)y_t &= \left(e^{-(1.5-1.9)^2/2} \times 5 + e^{-(2-1.9)^2/2} \times 4 + e^{-(2.5-1.9)^2/2} \times 3.5 \right) / \sqrt{2\pi} \\ &\approx (0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5) / \sqrt{2\pi} \\ &= 11.52 / \sqrt{2\pi}. \end{aligned}$$

The denominator is

$$\begin{aligned} \sum_{t=1}^T w(x_t - x) &= \left(e^{-(1.5-1.9)^2/2} + e^{-(2-1.9)^2/2} + e^{-(2.5-1.9)^2/2} \right) / \sqrt{2\pi} \\ &\approx 2.75 / \sqrt{2\pi}. \end{aligned}$$

The estimate at $x = 1.9$ is therefore

$$\hat{b}(1.9) \approx 11.52 / 2.75 \approx 4.19.$$

A low value of h means that the weights taper off fast—the weight function is then a

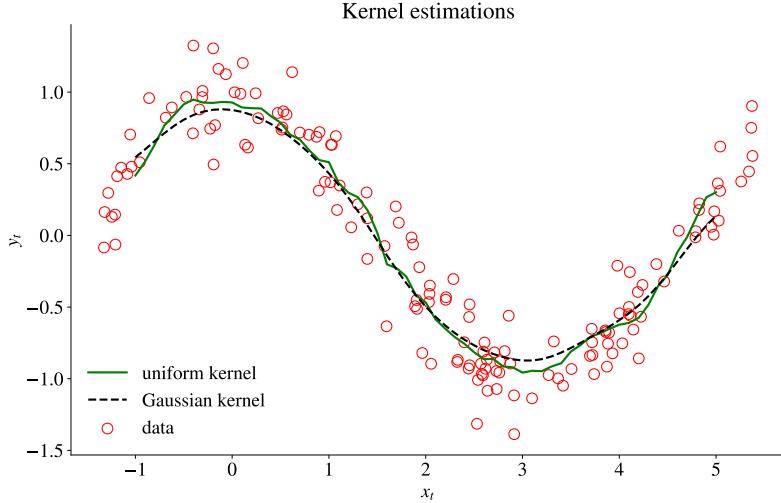


Figure 17.2: Example of kernel estimations

normal pdf with a low variance. When $h \rightarrow 0$, then $\hat{b}(x)$ evaluated at $x = x_t$ becomes just y_t , so no averaging is done. In contrast, as $h \rightarrow \infty$, $\hat{b}(x)$ becomes the sample average of y_t , so we have global averaging. Clearly, some value of h in between is needed. See Figure 17.4 for an illustration and Figure 17.5 for an empirical example.

A rule of thumb value of h is

$$h = T^{-1/5} |\gamma|^{-2/5} \sigma_\varepsilon^{2/5} (x_{\max} - x_{\min})^{1/5} \times 0.6, \quad (17.5)$$

where γ is a from the regression $y = \alpha + \beta x + \gamma x^2 + \varepsilon$ and σ_ε^2 is the variance of those fitted residuals. In practice, we replace $x_{\max} - x_{\min}$ by the difference between the 90th and 10th percentiles of x .

A good (but computationally intensive) approach to choose h is by the leave-one-out *cross-validation* technique. This approach would, for instance, choose h to minimize the expected (or average) prediction error

$$\text{EPE}(h) = \sum_{t=1}^T [y_t - \hat{b}_{-t}(x_t, h)]^2 / T, \quad (17.6)$$

where $\hat{b}_{-t}(x_t, h)$ is the fitted value at x_t when we use a regression function estimated on a sample that excludes observation t , and a bandwidth h . This means that each prediction is out-of-sample. To calculate (17.6) we clearly need to make T estimations (for each x_t)—and then repeat this for different values of h to find the minimum. See Figure 17.6 for an example.

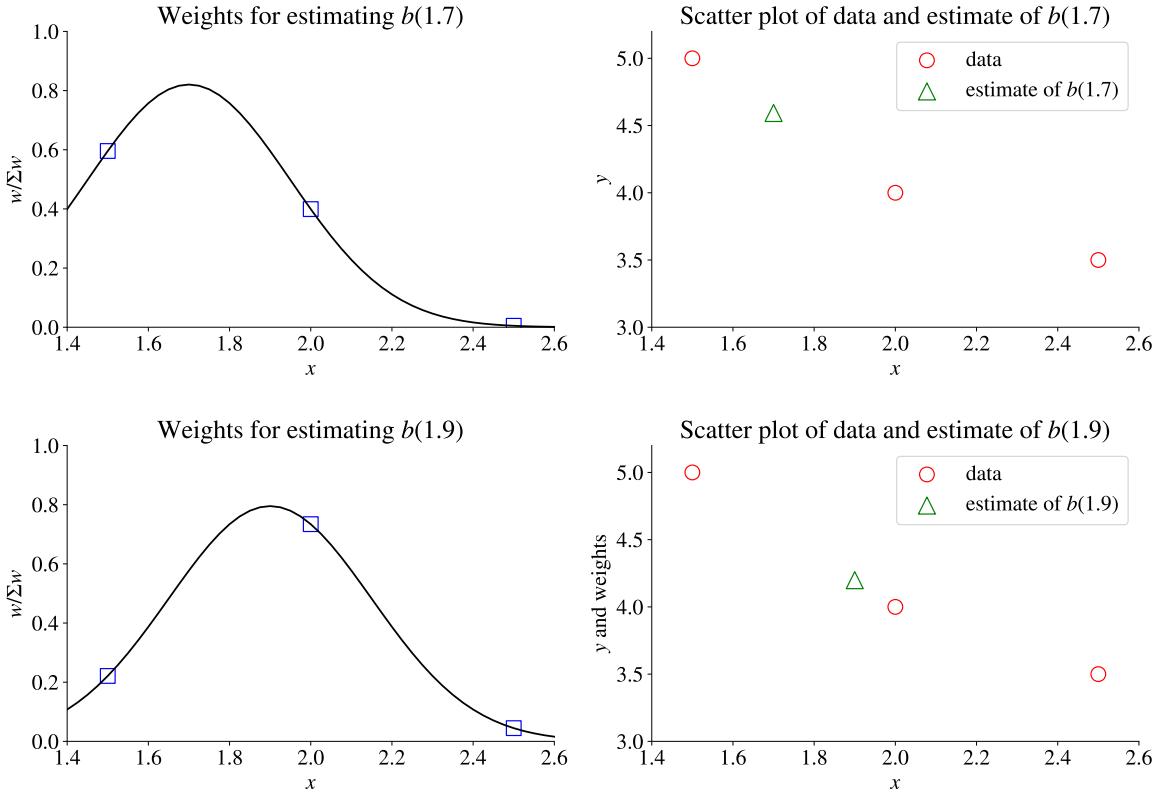


Figure 17.3: Example of kernel regression with three data points

Remark 17.2 (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 $t = 2$ and. calculate the error $y_2 - \hat{b}_{-2}(x_2)$. Repeat this for $t = 3, 4, \dots, T$. Calculate the EPE as in (17.6).

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)

If the observations are independent, then it can be shown (see Härdle (1990) 4.2 and Pagan and Ullah (1999) 3.3–6) that, with a Gaussian kernel, the estimator at point x is asymptotically normally distributed

$$\sqrt{Th} [\hat{b}(x) - b(x)] \xrightarrow{d} N \left[0, \frac{1}{2\sqrt{\pi}} \frac{\sigma^2(x)}{f(x)} \right], \quad (17.7)$$

where $\sigma^2(x)$ is the variance of the residuals in (17.1) and $f(x)$ the marginal density of

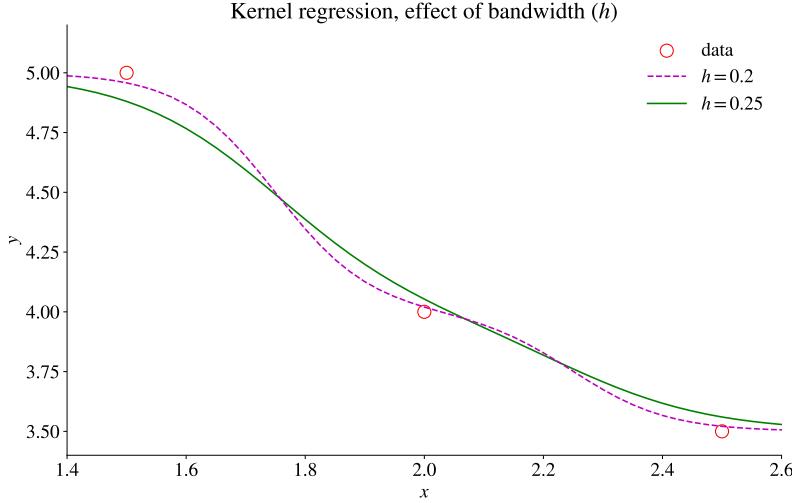


Figure 17.4: Example of kernel regression with three data points

x . For this to hold, h must be decreased (as T increases) slightly faster than $T^{-1/5}$ (for instance, suppose $h = T^{-1.1/5}h_0$, where h_0 is a constant). Clearly, this means that we have (with sloppy notation)

$$\hat{b}(x) \xrightarrow{d} N \left[b(x), \frac{1}{2\sqrt{\pi}} \frac{\sigma^2(x)}{f(x)} \frac{1}{Th} \right]. \quad (17.8)$$

As a comparison, a linear regression has $b(x) = z'\gamma$ where $z = [1, x]$, so the variance of the fitted value is

$$\text{Var}(z'\hat{\gamma}) = z'V(\hat{\gamma})z, \quad (17.9)$$

where $V(\hat{\gamma})$ is the variance-covariance matrix of $\hat{\gamma}$. Notice that this is different from the variance of a forecast error, since the latter also includes the variance of the residual.

To estimate the density function needed in (17.7), we can use a kernel density estimator of the pdf at some point x

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{h_x \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_t - x}{h_x} \right)^2 \right]. \quad (17.10)$$

Notice that h_x need not be the same as the bandwidth (h) used in the kernel regression. The value $h_x = \text{Std}(x_t)1.06T^{-1/5}$ is sometimes recommended for estimating the density function, since it can be shown to be the optimal choice (in MSE sense) if data is normally distributed and the $N(0, 1)$ kernel is used. See Figure 17.7 for an example where the width

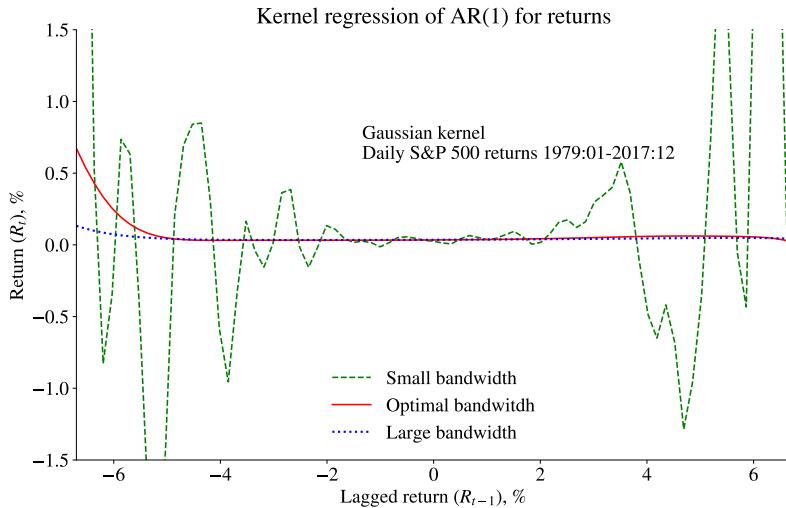


Figure 17.5: Non-parametric regression, importance of bandwidth

of the confidence band varies across x values—mostly because the sample contains few observations close to some x values.

To estimate the function $\sigma^2(x)$ in (17.7), we use a non-parametric regression of the squared fitted residuals on x_t

$$\hat{\varepsilon}_t^2 = \sigma^2(x_t), \text{ where } \hat{\varepsilon}_t = y_t - \hat{b}(x_t), \quad (17.11)$$

where $\hat{b}(x_t)$ are the fitted values from the non-parametric regression (17.1). Notice that this approach allows the variance to depend on the x value.

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.

17.2 Local Linear Regressions*

Notice that (17.2) 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 (17.2), 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.

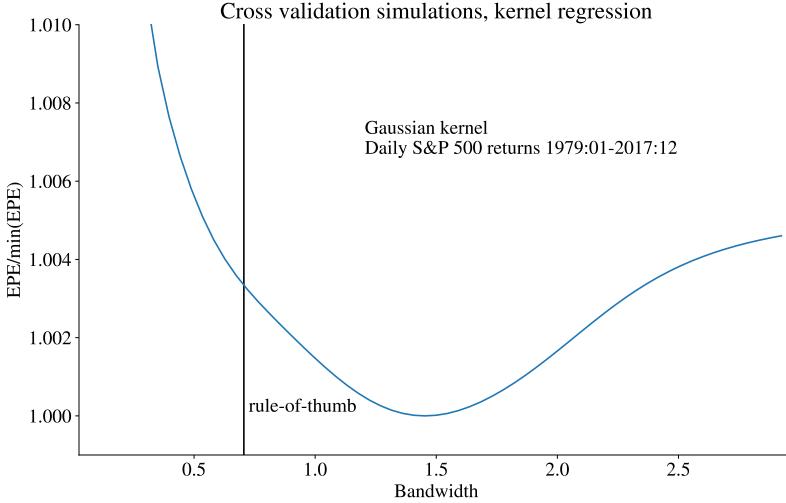


Figure 17.6: Cross-validation

This can be extended to solving a problem like

$$\min_{\alpha_x, \gamma_x} \sum_{t=1}^T w(x_t - x)[y_t - \alpha_x - \beta_x(x_t - x)]^2, \quad (17.12)$$

which defines the *local linear estimator*. (Yes, 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 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

$$\sum_{t=1}^T w(x_t - x) z_t y_t = \sum_{t=1}^T w(x_t - x) z_t z_t' \hat{\theta}_x. \quad (17.13)$$

It is straightforward to solve these, but perhaps even easier if we create $\tilde{z}_t = \sqrt{w(x_t - x)} z_t$ and $\tilde{y}_t = \sqrt{w(x_t - x)} y_t$, because (17.13) is then the same as the first order conditions for a regression of \tilde{y}_t on \tilde{z}_t (without a constant). (An extension to a quadratic or higher function seems straightforward.)

Clearly, solving (17.13)) 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 only shows up in the calculations of the weights, $w(x_t - x)$) can be chosen by a leave-one-out cross validation approach.

It can be shown that the local-linear estimator has the same asymptotic variance as the

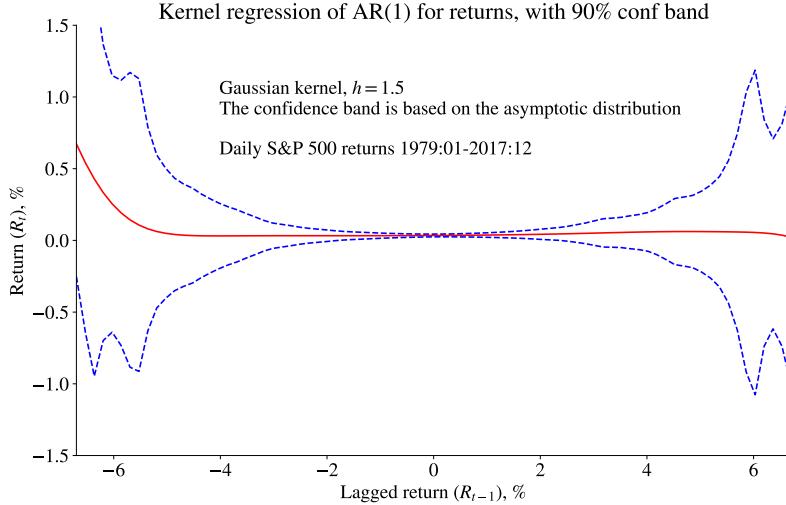


Figure 17.7: Non-parametric regression with confidence bands

kernel regression.

See Figures 17.9 – 17.10 for an empirical illustration.

17.2.1 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \quad \varepsilon_t \text{ is iid and } E \varepsilon_t = 0. \quad (17.14)$$

This makes the estimation problem more data demanding. To see why, suppose we use a uniform density function as weighting function (see in (17.3)). 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.

In any case, the most common way to implement the kernel regressor is to let

$$\hat{b}(x, z) = \frac{\sum_{t=1}^T w(x_t - x)v(z_t - z)y_t}{\sum_{t=1}^T w(x_t - x)v(z_t - z)}, \quad (17.15)$$

where $w(x_t - x)$ and $v(z_t - z)$ are two kernels like in (17.4) 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). 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.

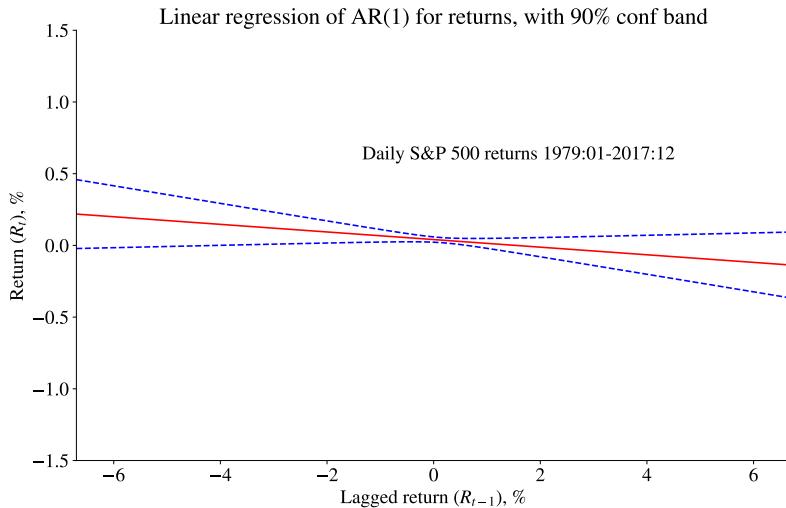


Figure 17.8: Linear regression with confidence bands

See Figure 17.11 for an example.

17.3 Examples of Non-Parametric Estimation

17.3.1 A Model for the Short Interest Rate

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

$$r_t = \rho r_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim N(0, \sigma^2), \text{ so} \quad (17.16)$$

$$r_t - r_{t-1} = (\rho - 1)r_{t-1} + \varepsilon_t, \quad (17.17)$$

and that all term premia are constant. This means that the drift in (17.17) is decreasing in the interest rate, but that the volatility is constant. For instance, if $\rho = 0.95$ (a very persistent interest rate), then (17.17) is

$$r_t - r_{t-1} = -0.05r_{t-1} + \varepsilon_t, \quad (17.18)$$

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 discrete time model in (17.16).) It can then be shown

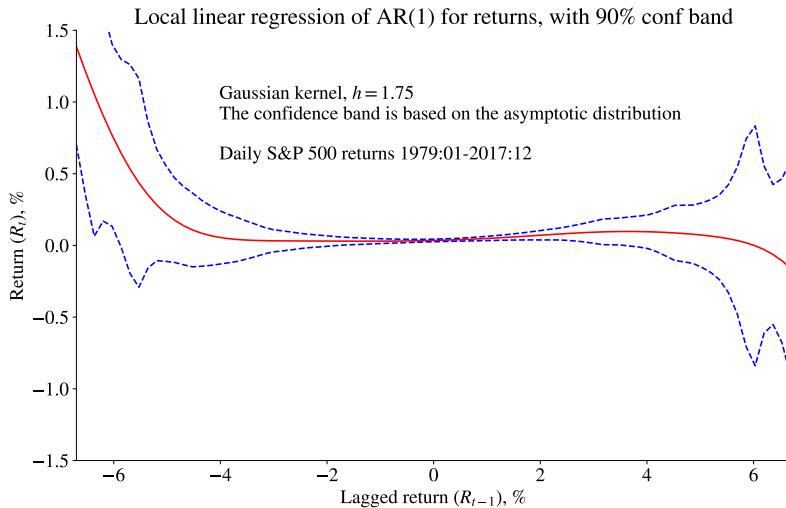


Figure 17.9: Non-parametric local linear regression with confidence bands

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) as in (17.16) 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)). Figure 17.12 gives an example. Note that the volatility is defined as the square of the drift minus expected drift (from the same estimation method).

17.3.2 Non-Parametric Option Pricing

There seems to be systematic deviations from the Black-Scholes model. For instance, implied volatilities are often higher for options far from the current spot (or forward) price—the volatility smile. This is sometimes interpreted as if the beliefs about the future log asset price put larger probabilities on very large movements (“fat tails”) than what is compatible with the normal distribution.

This has spurred many efforts to both describe the distribution of the underlying asset price and to amend the Black-Scholes formula by adding various adjustment terms. One strand of this literature uses non-parametric regressions to fit observed option prices to the variables that also show up in the Black-Scholes formula (spot price of underlying

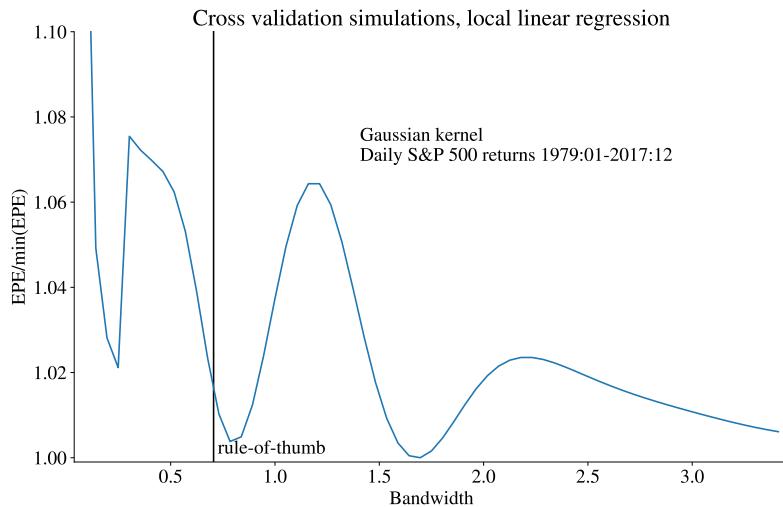


Figure 17.10: Cross-validation

asset, strike price, time to expiry, interest rate, and dividends). For instance, Ait-Sahalia and Lo (1998) apply this to daily data for Jan 1993 to Dec 1993 on S&P 500 index options (14,000 observations). They find interesting patterns of the implied moments (mean, volatility, skewness, and kurtosis) as the time to expiry changes. In particular, the non-parametric estimates suggest that distributions for longer horizons have increasingly larger skewness and kurtosis.

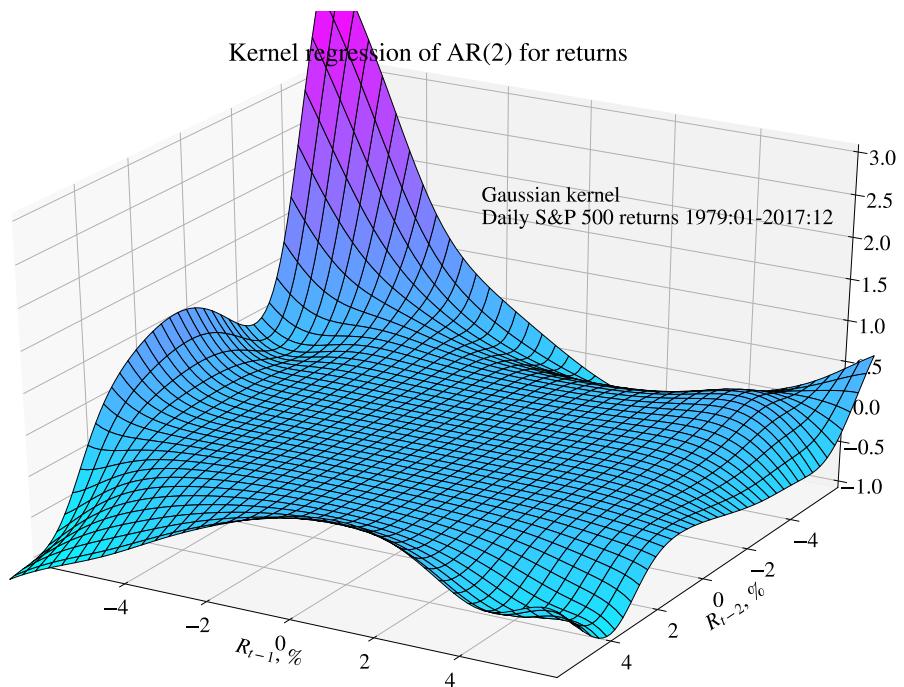


Figure 17.11: Non-parametric regression with two regressors

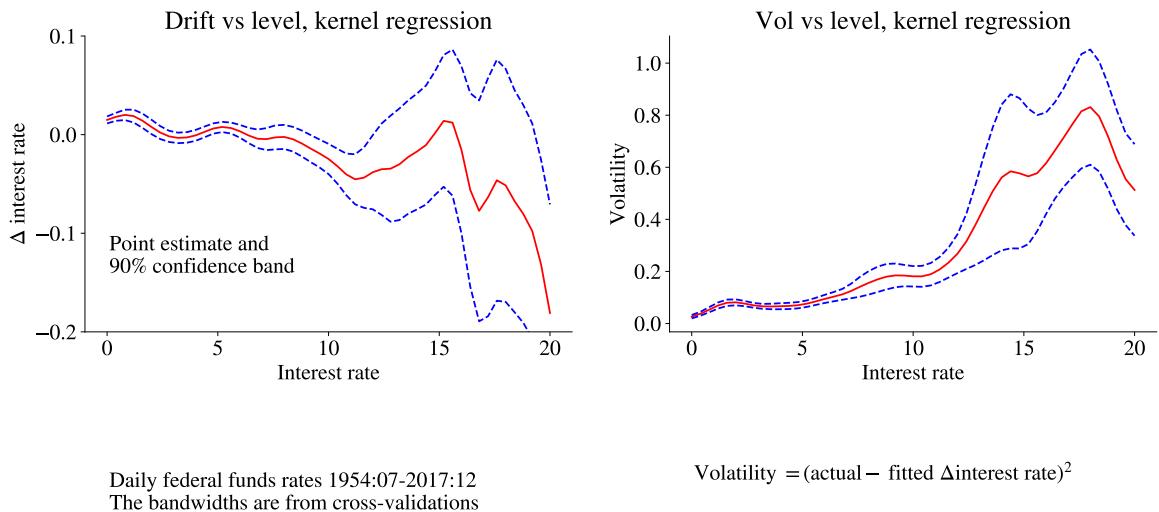


Figure 17.12: Kernel regression, confidence band

Chapter 18

Simulating the Finite Sample Properties*

Reference: Greene (2000) 5.3 and Horowitz (2001)

Additional references: Cochrane (2001) 15.2; Davidson and MacKinnon (1993) 21; Davidson and Hinkley (1997); Efron and Tibshirani (1993) (bootstrapping, chap 9 in particular); and Berkowitz and Kilian (2000) (bootstrapping in time series models)

18.1 Introduction

We know the small sample properties of regression coefficients in linear models with fixed regressors and iid normal error terms. When these conditions are not satisfied, then we may either rely on asymptotic (large sample) results or use Monte Carlo simulations and bootstrapping to understand the small sample properties.

How such simulations should be implemented depends crucially on the properties of the model and data: if the residuals are autocorrelated, heteroskedastic, or perhaps correlated across regressions equations. These notes summarize a few typical cases.

The need for using Monte Carlos or bootstraps varies across applications and data sets. For a case where it does not matter much, see Table 18.1, and for a case where it matters, compare the traditional and bootstrapped t-stats in Tables 18.2–18.3.

18.2 Monte Carlo Simulations

18.2.1 Monte Carlo Simulations in the Simplest Case

Monte Carlo simulations is essentially a way to generate many artificial (small) samples from a parameterised model and then estimate the statistic (for instance, a slope coefficient) on each of those samples. The distribution of the statistic is then used as the small

	α	t (LS)	t (NW)	t (boot)
A (NoDur)	3.42	2.75	2.55	2.18
B (Durlb)	-1.52	-0.79	-0.80	-0.80
C (Manuf)	0.63	0.70	0.67	0.62
D (Enrgy)	2.69	1.25	1.24	1.30
E (HiTec)	-1.57	-0.94	-0.95	-0.84
F (Telcm)	1.62	1.04	1.00	0.89
G (Shops)	1.23	0.92	0.87	0.84
H (Hlth)	2.28	1.41	1.44	1.37
I (Utils)	2.98	1.76	1.72	1.80
J (Other)	-0.55	-0.56	-0.54	-0.45

Table 18.1: Estimates of CAPM on US industry portfolios 1970:01-2017:12. NW uses 1 lag. The bootstrap samples (y_t, x_t) pairs, in blocks of 10 observations and has 3000 simulations.

	2y	3y	4y	5y
factor	1.00 (6.84)	1.87 (6.96)	2.67 (7.12)	3.45 (7.29)
constant	-0.00 (-0.00)	-0.00 (-0.34)	-0.00 (-0.69)	-0.00 (-1.04)
R^2	0.14	0.15	0.15	0.16
obs	636.00	636.00	636.00	636.00

Table 18.2: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. Numbers in parentheses are t-stats. U.S. data for 1964:01-2017:12.

sample distribution of the estimator.

The following is an example of how Monte Carlo simulations could be done in the special case of a linear model

$$y_t = x_t' \beta + u_t, \quad (18.1)$$

where u_t is iid $N(0, \sigma^2)$ and x_t is stochastic but independent of $u_{t \pm s}$ for all s . This means that x_t cannot include lags of y_t .

Suppose we want to find the small sample distribution of a function of the estimate, $g(\hat{\beta})$. To do a Monte Carlo experiment, we need information on (i) the coefficients β ; (ii) the variance of u_t , σ^2 ; (iii) and a process for x_t .

The process for x_t is typically estimated from the data on x_t (for instance, a VAR system $x_t = A_1 x_{t-1} + A_2 x_{t-2} + e_t$). Alternatively, we could simply use the actual

	2y	3y	4y	5y
factor	1.00 (4.00)	1.87 (4.14)	2.67 (4.29)	3.45 (4.44)
constant	-0.00 (-0.00)	-0.00 (-0.16)	-0.00 (-0.32)	-0.00 (-0.48)
R^2	0.14	0.15	0.15	0.16
obs	636.00	636.00	636.00	636.00

Table 18.3: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. U.S. data for 1964:01-2017:12. Numbers in parentheses are t-stats. Bootstrapped standard errors, with blocks of 10 observations.

sample of x_t and repeat it.

The values of β and σ^2 are often a mix of estimation results and theory. In some case, we simply take the point estimates. In other cases, we adjust the point estimates so that $g(\beta) = 0$ holds, that is, so you *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, you may *simulate the model under an alternative hypothesis* in order to study the power of the test using either critical values from either the asymptotic distribution or from a (perhaps simulated) small sample distribution.

To make it a bit concrete, suppose you want to use these simulations to get a 5% critical value for testing the null hypothesis $g(\beta) = 0$. The Monte Carlo experiment follows these steps.

1. Construct an artificial sample of the regressors (see above), \tilde{x}_t for $t = 1, \dots, T$. Draw random numbers \tilde{u}_t for $t = 1, \dots, T$ from a prespecified distribution (eg. normal) 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

$$\tilde{y}_t = \tilde{x}'_t \beta + \tilde{u}_t, \quad (18.2)$$

by using the prespecified values of the coefficients β (perhaps your point estimates or a tweaked version of the point estimates so $g(\beta) = 0$ holds).

2. Calculate an estimate $\tilde{\beta}$ and record it along with the value of $g(\tilde{\beta})$ and perhaps also the test statistic of the hypothesis that $g(\beta) = 0$.

3. Repeat the previous steps N (3000, say) times. The more times you repeat, the better is the approximation of the small sample distribution.
4. Sort your simulated $\tilde{\beta}$, $g(\tilde{\beta})$, and the test statistic in ascending order. For a one-sided test (for instance, a chi-square test), take the $(0.95N)$ th observations in these 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. You may also record how many times the 5% critical values from the asymptotic distribution would reject a true null hypothesis.
5. You may also want to plot a histogram of $\tilde{\beta}$, $g(\tilde{\beta})$, and the test statistic to see if there is a small sample bias, and how the distribution looks like. Is it close to normal? How wide is it? You could also estimate the variance-covariance matrix of $\tilde{\beta}$ by treating each estimate (from each simulation) as an observation—and then estimate the covariance matrix across these observations.

We have the same basic procedure when y_t is a vector, except that we might have to consider correlations across the elements of the vector of residuals u_t . For instance, we might want to generate the vector \tilde{u}_t from a $N(\mathbf{0}, \Sigma)$ distribution—where Σ is the variance-covariance matrix of u_t .

Remark 18.1 (*Generating $N(\mu, \Sigma)$ random numbers**) Suppose you want to draw an $n \times 1$ vector ε_t of $N(\mu, \Sigma)$ variables. Use the Cholesky decomposition of Σ to calculate the lower triangular P such that $\Sigma = PP'$. Draw u_t from an $N(0, I_n)$ distribution, and define $\varepsilon_t = \mu + Pu_t$. Note that $\text{Cov}(\varepsilon_t) = E P u_t u_t' P' = P I P' = \Sigma$.

It is straightforward to sample the errors from other distributions than the normal, for instance, a student- t distribution. Equipped with uniformly distributed random numbers, you can always (numerically) invert the cumulative distribution function (cdf) of any distribution to generate random variables from any distribution by using the probability transformation method. See *Figure 18.1* for an example.

Remark 18.2 (*The probability transformation method**) A random variable Y has the cdf $u = \Pr(Y \leq y) = F(y)$, where y is a number. Clearly, u is a probability and thus between 0 and 1. Draw random numbers u_i from a uniform distribution over $(0, 1)$. Then, calculate $y_i = F^{-1}(u_i)$, where $F^{-1}()$ is the inverse of $F()$. A sample of y_i will have the cdf F .

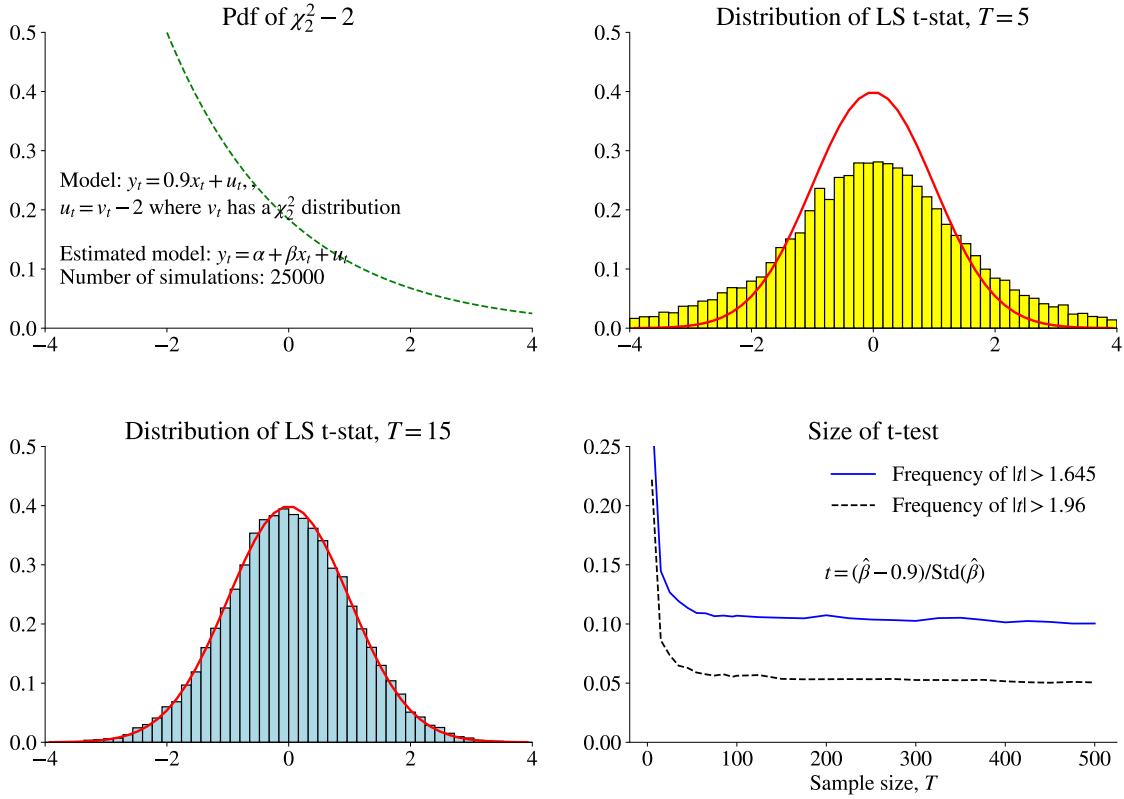


Figure 18.1: Results from a Monte Carlo experiment with thick-tailed errors.

Example 18.3 (*The probability transformation method**) *The exponential cdf is $u = 1 - \exp(-\theta y)$ with inverse $y = -\ln(1-u)/\theta$. Draw u_i from $U(0, 1)$ and transform to y_i to get an exponentially distributed variable.*

18.2.2 *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 so that temporal link is preserved in every artificial sample which we create. For instance, suppose x_t includes y_{t-1} and another vector z_t of variables which are independent of $u_{t \pm s}$ for all s

$$\begin{aligned} y_t &= x_t' \beta + u_t \\ &= \gamma y_{t-1} + \phi' z_t + u_t \end{aligned} \tag{18.3}$$

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

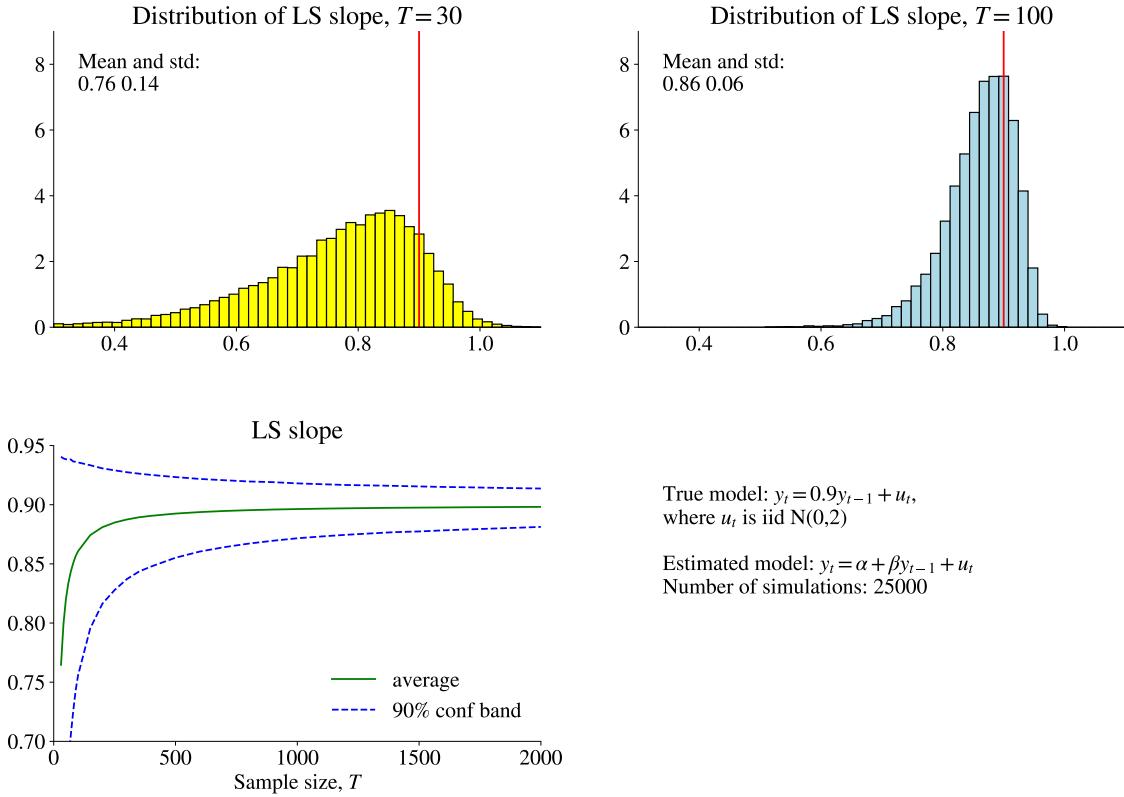


Figure 18.2: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

sample itself. Second, observation t of is generated recursively as

$$\tilde{y}_t = \gamma \tilde{y}_{t-1} + \phi' z_t + \tilde{u}_t \text{ for } t = 1, \dots, T \quad (18.4)$$

We clearly need the initial value \tilde{y}_0 to start up the artificial sample—and then the rest of the sample ($t = 1, 2, \dots$) is calculated recursively. See Figures 18.2–18.2 for an example.

Example 18.4 (Simulating VAR models) A VAR model is similar to (18.3), but includes no z_t variable but may have more lags. For instance, a VAR(2)

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t,$$

where y_t could be a vector of variables. The simulation procedure is straightforward. First, estimate the model on data and record the estimates ($A_1, A_2, \text{Var}(u_t)$). Second, draw a new time series of residuals, \tilde{u}_t for $t = 1, \dots, T$ and construct an artificial

sample recursively (first $t = 1$, then $t = 2$ and so forth) as

$$\tilde{y}_t = A_1 \tilde{y}_{t-1} + A_2 \tilde{y}_{t-2} + \tilde{u}_t.$$

(This requires some starting values for y_{-1} and y_0 .) Third, re-estimate the model on the artificial sample, \tilde{y}_t for $t = 1, \dots, T$.

18.2.3 Monte Carlo Simulations with non-iid Errors

It is more difficult to handle non-iid errors, like those with autocorrelation and heteroskedasticity. We then need to model the error process and generate the errors from that model.

If the errors are *autocorrelated*, then we could estimate a time series process from the fitted errors and then generate artificial samples of errors. For instance, with an AR(2) we get

$$\tilde{u}_t = a_1 \tilde{u}_{t-1} + a_2 \tilde{u}_{t-2} + \tilde{\varepsilon}_t. \quad (18.5)$$

See Figure 18.3 for an illustration.

Alternatively, *heteroskedastic errors* can be generated by, for instance, a GARCH(1,1) model

$$u_t \sim N(0, \sigma_t^2), \text{ where } \sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2. \quad (18.6)$$

However, this specification does not account for any link between the volatility and the regressors (squared)—as tested for by White’s test. This would invalidate the usual OLS standard errors and therefore deserves to be taken seriously. A simple, but crude, approach is to generate residuals from a $N(0, \sigma_t^2)$ process, but where σ_t^2 is approximated by the absolute values of fitted values ($\sigma_t^2 = |\hat{\varepsilon}_t^2|$) from

$$\hat{\varepsilon}_t^2 = c' w_t + \eta_t, \quad (18.7)$$

where w_t include the squares and cross product of all the regressors.

18.3 Bootstrapping

18.3.1 Bootstrapping in the Simplest Case

Bootstrapping is another way to do simulations, where we construct artificial samples by sampling from the actual data. The advantage of the bootstrap is then that we do not

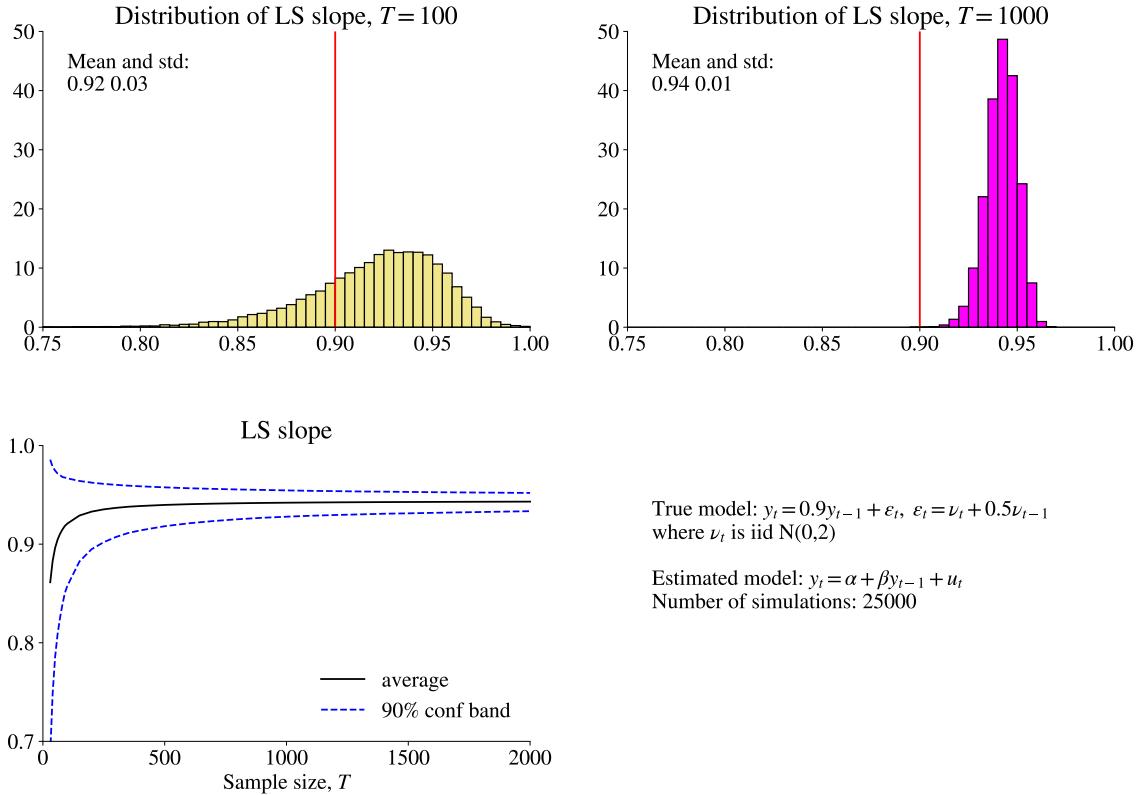


Figure 18.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

have to try to estimate the process of the errors and regressors (as we do in a Monte Carlo experiment). The real benefit of this is that we do not have to make any strong assumption about the distribution of the errors.

The bootstrap approach works particularly well when the errors are iid and independent of x_{t-s} for all s . (This means that x_t cannot include lags of y_t .) We here consider bootstrapping the linear model (18.1), for which we have point estimates (perhaps from LS) and fitted residuals. The procedure is similar to the Monte Carlo approach, except that the artificial sample is generated differently. In particular, Step 1 in the Monte Carlo simulation is replaced by the following:

1. Construct an artificial sample \tilde{y}_t for $t = 1, \dots, T$ by

$$\tilde{y}_t = x'_t \beta + \tilde{u}_t, \quad (18.8)$$

where \tilde{u}_t is drawn (with replacement) from the *fitted residual* and where β is the

point estimate from the original sample.

Example 18.5 With $T = 3$, the artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x'_1\beta + u_2, x_1) \\ (x'_2\beta + u_1, x_2) \\ (x'_3\beta + u_2, x_3) \end{bmatrix}.$$

The approach in (18.8) works also when y_t is a vector of dependent variables—and will then help retain the cross-sectional correlation of the residuals.

18.3.2 *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 in (18.8) since \tilde{u}_t can become correlated with x_t . For instance, if x_t includes y_{t-1} and we happen to sample $\tilde{u}_t = u_{t-1}$, then we get a non-zero correlation. The easiest way to handle this is as in the Monte Carlo simulations in (18.4), but where \tilde{u}_t are drawn (with replacement) from the sample of fitted residuals.

18.3.3 Bootstrapping when Errors Are Heteroskedastic

Suppose now that the errors are heteroskedastic, but serially uncorrelated. If the heteroskedasticity is unrelated to the regressors, then we can still use (18.8).

In contrast, if the heteroskedasticity is related to the regressors, then the traditional LS covariance matrix is not correct (this is the case that White's test for heteroskedasticity tries to identify). It would then be wrong to pair x_t with just any $\tilde{u}_t = u_s$ since that destroys the relation between x_t and the variance of the residual.

An alternative way of bootstrapping can then be used: generate the artificial sample by drawing (with replacement) pairs (y_s, x_s) , that is, we let the artificial pair in t be $(\tilde{y}_t, \tilde{x}_t) = (y_s, x_s) = (x'_s\beta + u_s, x_s)$ for some random draw of s so we are always pairing the residual, u_s , with the contemporaneous regressors, x_s . 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.

Example 18.6 With $T = 3$, the artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (y_2, x_2) \\ (y_3, x_3) \\ (y_3, x_3) \end{bmatrix} = \begin{bmatrix} (x'_2\beta + u_2, x_2) \\ (x'_3\beta + u_3, x_3) \\ (x'_3\beta + u_3, x_3) \end{bmatrix}$$

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.

18.3.4 Bootstrapping when Errors Are Autocorrelated

It is quite hard to handle the case when the errors are serially dependent, since we must sample in such a way that we do not destroy the autocorrelation structure of the data. 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 blocks 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 add 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.

Example 18.7 With $T = 9$ and a block size of 3, the artificial sample could be

$$\underbrace{u_2, u_3, u_4}_{block\ 2}, \underbrace{u_7, u_8, u_9}_{block\ 7}, \underbrace{u_4, u_5, u_6}_{block\ 4}.$$

See Table 18.4 for an illustration.

$\rho :$	0.0	0.75
Simulated	5.8	23.1
OLS formula	5.8	8.7
Newey-West	5.7	16.3
VARHAC	5.7	22.4
Bootstrapped	5.5	19.6

Table 18.4: 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 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

Chapter 19

Financial Panel Data

References: Verbeek (2012) 10 and Baltagi (2008)

19.1 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

$$y_{it} = \alpha_i + x'_{it}\beta_i + u_{it}, \quad (19.1)$$

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

$$\begin{array}{cccc} & \underline{i=1} & \underline{i=2} & \cdots & \underline{i=N} \\ t=1: & y_{11} & y_{21} & & y_{N1} \\ t=2: & y_{12} & y_{22} & & y_{N2} \\ & \vdots & & & \\ t=T: & y_{1T} & y_{2T} & & y_{NT} \end{array} \quad (19.2)$$

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 investors (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 (19.2) 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 3 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) . 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).

19.2 Portfolio Sorts

Reference: Bali, Engle, and Murray (2016)

Portfolio sorts are used to construct portfolios (groups) based on some characteristic, for instance, firm size. Once the portfolios have been defined (to which portfolio does i belong?) we often compute the (possibly weighted) average with each portfolio

$$\bar{y}_{gt} = \sum_{i \in \text{Group}_g} w_{it} y_{it}, \quad (19.3)$$

where w_{it} is the relative portfolio weight of asset i in the portfolio ($\sum w_{it} = 1$). We often use an unweighted average where $w_{it} = 1/(\text{number of members of the group})$.

19.2.1 Univariate Sorts

A simple and commonly applied method for studying how an asset characteristic (x_{it-1}) 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 $x_{i,t-1}$ and then construct three portfolios: (1) for those i whose $x_{i,t-1}$ belong to the top 1/3; (2) those in the mid 1/3 and (3) those in the lowest 1/3. Then, we measure the return on equally weighted portfolios—and perhaps analyse the return of portfolio 1 minus the return on portfolio 3. 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.

Example 19.1 (“Betting against beta” by Frazzini and Pedersen*) This paper creates a long-short portfolio by doing the following. 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, in $t-1$.) This is the low beta group. Second, sort the betas (of this group) in decreasing order and let asset i have the weight $\text{rank}(\beta_{i,t-1})/\text{sum of ranks}$. For instance, with 49 assets in the low beta group, the asset with the lowest beta gets a weight of $49/1225 \approx 0.04$ (which is higher than an equal weighting which would give $1/49 \approx 0.02$). Third, calculate the portfolio return in t . Fourth, repeat for all periods. Fifth, do points 1–4 also for the high beta assets, $\beta_{i,t-1} \geq \text{median}(\beta_{i,t-1})$, but for this group you sort in increasing order. Sixth, form the excess return as the difference between the two portfolios.

19.2.2 Bivariate Sorts

Bivariate sorts can be done in several ways. An *independent bivariate sort* first does a univariate sort of $x_{i,t-1}$ (say, forming 3 groups), then it makes another univariate sort according to another sorting variable $z_{i,t-1}$ (say, forming two groups). Then we find the intersections of the two sorts (think of a matrix). For instance, we find the assets which belong to group 2 according to the x sort and group 1 according to the z -sort.

In *dependent bivariate sort* we first sort according to $x_{i,t-1}$ as before. Then, *within* an x group we sort according to $z_{i,t-1}$. Notice that the ordering matters in the dependent sort: letting x represent book/market and z size will not give the same results as switching the labels. The interpretation of the dependent sort is that you are studying the effect of z (eg. size), but controlling for x (eg. being a value firm).

Remark 19.2 (*The Fama-French factors*) The SMB and HML are created by an independent bivariate sort. First classify firms according to size: small or big, using the median as a cutoff. Second, 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). Create six value weighted portfolios from the intersection of those groups

	<u>Low book/market</u>	<u>Medium book/market</u>	<u>High book/market</u>
Small:	Small Growth (SG)	Small Neutral (SN)	Small Value (SV)
Big:	Big Growth (BG)	Big Neutral (BN)	Big Value (BV)

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 $SMB = 1/3(SG - BG) + 1/3(SN - BN) + 1/3(SV - BV)$, 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).

19.3 Calendar Time Regressions and SURE

Once we have defined portfolios and computed their average excess returns as in (19.3), we may want to regress the portfolio returns on risk factors (the market excess return or the three Fama-French factors) to estimate the portfolio alpha. This section discusses how to do that—in such a way that we can test differences (in alphas) across portfolios.

19.3.1 SURE

Consider a model where each asset (i) has its own intercept and slope coefficients as in (19.1). For convenience, write the model as

$$y_{it} = z_t' \gamma_i + u_{it}, \quad (19.4)$$

where z_t contains both the constant and the other regressors (x_t). Notice that SURE is about the case when all regressors are the same, so we drop the i subscript on z_t .

Previous chapters of the lecture notes have shown that

$$\text{Cov}(\hat{\gamma}_i, \hat{\gamma}_j) = \hat{\sigma}_{ij} S_{zz}^{-1}, \quad (19.5)$$

where $S_{zz} = \Sigma_{t=1}^T z_t z_t'$ and where $\hat{\sigma}_{ij}$ is the sample covariance of the fitted residuals u_{it} and u_{jt} (two different assets). Set j equal to i to get the variance-covariance matrix of $\hat{\gamma}_i$.

Remark 19.3 (*System of OLS regressions, different regressors**) When the regressions have different regressors

$$\begin{aligned} y_{1t} &= x_t' \gamma_1 + u_{1t} \\ y_{2t} &= w_t' \gamma_2 + u_{2t}, \end{aligned}$$

then the variance-covariance matrix of the stacked $(\hat{\gamma}_1, \hat{\gamma}_2)$ vector is

$$\begin{bmatrix} \Sigma_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{ww}^{-1} \end{bmatrix} \begin{bmatrix} \sigma_{11} \Sigma_{xx} & \sigma_{12} \Sigma_{xw} \\ \sigma_{12} \Sigma_{wx} & \sigma_{22} \Sigma_{ww} \end{bmatrix} \begin{bmatrix} \Sigma_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{ww}^{-1} \end{bmatrix}.$$

19.3.2 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 discrete asset groups by a portfolio sort (for instance, construct ten portfolios for different firm size deciles) and calculate their respective excess returns as in (19.3) by letting y_{it} represent the excess return of asset i .

Then, we run a factor model

$$\bar{R}_{gt}^e = z_t' \gamma_g + u_{jt}, \text{ for } g = 1, 2, \dots, M \quad (19.6)$$

where z_t typically includes a constant and various return factors (for instance, the three Fama-French factors). 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 the same for all portfolios.

Example 19.4 (*CalTime with two asset groups*) *With two asset groups, estimate the following SURE system*

$$\begin{aligned}\bar{R}_{1t}^e &= z_t' \gamma_1 + u_{1t}, \\ \bar{R}_{2t}^e &= z_t' \gamma_2 + u_{2t}.\end{aligned}$$

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. We therefore consider panel data models.

See Table 19.1 for results on a ten-year panel of some 60,000 Swedish pension savers from Dahlquist, Martinez, and Söderlind (2016). In this case, the dependent variable is the return of a pension investment portfolio (on day t , individual i). 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).

19.4 An Overview of Different Panel Data Models

A *pooled model* assumes that all individuals have the same coefficients (no subscript on β), so (19.1) becomes

$$y_{it} = \alpha + x_{it}' \beta + u_{it}. \quad (19.7)$$

	Inactive	Active	Higly Active
coef	-0.76	3.08	8.65
t-tstat NW	-0.69	1.77	2.73

Table 19.1: Calendar time regressions. 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).

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

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}. \quad (19.8)$$

An extension of the fixed effects model is to also allow for *time fixed effects*

$$y_{it} = \lambda_t + \alpha_i + x'_{it}\beta + u_{it}. \quad (19.9)$$

Estimation of these models is discussed below.

A *random effects model* is similar to a fixed effects model, except that the individual “mean” α_i now contains a common component (α) and a random individual component (μ_i). We can then write the model as

$$y_{it} = \alpha + x'_{it}\beta + u_{it} \text{ where } u_{it} = \mu_i + \varepsilon_{it}. \quad (19.10)$$

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 (see SURE for the case when the regressor values are common among the assets).

19.5 Pooled OLS

Consider the regression model

$$y_{it} = z'_{it}\gamma + u_{it}, \quad (19.11)$$

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. We assume that u_{jt} is uncorrelated with z_{it} (across all i and j).

Define the matrices

$$S_{zz} = \sum_{t=1}^T \sum_{i=1}^N z_{it}z'_{it} \text{ (a } k \times k \text{ matrix)} \quad (19.12)$$

$$S_{zy} = \sum_{t=1}^T \sum_{i=1}^N z_{it}y_{it} \text{ (a } k \times 1 \text{ vector).} \quad (19.13)$$

The LS estimator (stacking all TN observations) is then

$$\hat{\gamma} = S_{zz}^{-1}S_{zy}. \quad (19.14)$$

In case u_{it} is uncorrelated across time and also across individuals, then the usual expressions for $\text{Std}(\hat{\gamma})$ apply. However, it is often the case that there are *clusters* of individuals (all small firms, say) that have correlated residuals. This would require handling those correlations.

Recall that we can (conceptually) decompose the point estimate $\hat{\gamma}$ by using (19.11) to substitute for y_{it} in S_{zy} (19.13) and then in (19.14). The result is

$$\hat{\gamma} = \gamma + S_{zz}^{-1} \sum_{t=1}^T \sum_{i=1}^N z_{it}u_{it}. \quad (19.15)$$

The variance-covariance matrix can then be written

$$\text{Var}(\hat{\gamma}) = S_{zz}^{-1}SS_{zz}^{-1}, \text{ where} \quad (19.16)$$

$$S = \text{Var}(h_{1,1} + h_{2,1} + \dots + h_{N-1,T} + h_{N,T}), \quad (19.17)$$

where $h_{it} = z_{it}u_{it}$ is used as short hand notation.

When we assume that the residuals in (19.17) are iid and also independent of the regressors, then $\text{Var}(\hat{\gamma})$ simplifies to the usual OLS expression. Instead, with heteroskedasticity (but still no cross-sectional correlations) we can apply the usual White's approach, and with autocorrelation the Newey-West approach. Instead, with cross-sectional corre-

lations we need to make further adjustments (see below for a discussion).

Remark 19.5 (**Panel regression vs average coefficient in the case of common regressors*)
Consider the regression for investor i

$$y_{it} = z_t' \gamma_i + u_{it}, \quad i = 1 \dots N,$$

where the regressors are the same in all regressions—but where the coefficients might be different across investors. Clearly, we have for each i

$$\hat{\gamma}_i = \tilde{S}_{zz}^{-1} \tilde{S}_{zy_i},$$

where $\tilde{S}_{zz} = \sum_{t=1}^T z_t z_t'$ and $\tilde{S}_{zy_i} = \sum_{t=1}^T z_t y_{it}$.

The cross-sectional average of the regression coefficients is therefore

$$\frac{1}{N} \sum_{i=1}^N \gamma_i = (N \tilde{S}_{zz})^{-1} \sum_{i=1}^N \tilde{S}_{zy_i}.$$

Notice that $N \tilde{S}_{zz} = S_{zz}$ in (19.12) and that $\sum \tilde{S}_{zy_i} = S_{zy}$ in (19.13), which shows that $\frac{1}{N} \sum_{i=1}^N \gamma_i = \gamma$, where the latter is from the pooled panel regression (19.14).

19.5.1 Cross Sectional 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 ($\text{Cov}(h_{i,t}, h_{j,t+s}) = 0$ for $s \neq 0$), then we can write 19.17) as

$$S = T \text{Var} \left(\sum_{i=1}^N h_{it} \right). \quad (19.18)$$

Positive correlations in the cross-section (h_{it} is correlated with h_{jt}) increases the right hand side of (19.18). Actually, it is similar to having autocorrelation in a traditional (single-equation) regression model.

Example 19.6 ($N = 4$) To save space, let $V()$ denote a variance and $C(,)$ a covariance. For $N = 4$, the $\text{Var}()$ term in (19.18) can then be written (dropping the time subscripts to save even more space)

$$\begin{aligned} \text{Var}(\sum_{i=1}^4 h_i) &= V(h_1) + V(h_2) + V(h_3) + V(h_4) + \\ &\quad 2 C(h_1, h_2) + 2 C(h_1, h_3) + 2 C(h_1, h_4) + 2 C(h_2, h_3) + 2 C(h_2, h_4) + 2 C(h_3, h_4). \end{aligned}$$

Cross-sectional correlations mean that some $C(,)$ terms are non-zero. The interpretation is easiest when there is a single regressor so h_i is a scalar. (In case there are several regressors, then $V(h_i)$ is a variance-covariance matrix, and $C(h_1, h_2)$ is a covariance matrix.)

A cluster method makes assumptions about which cross-sectional units (i and j) can be correlated—and then uses the time-series data to estimate those.

Example 19.7 (*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,

$$\begin{aligned} \text{Var}(\sum_{i=1}^4 h_i) &= V(h_1) + V(h_2) + V(h_3) + V(h_4) + \\ &\quad 2C(h_1, h_2) + 2\underbrace{C(h_1, h_3)}_0 + 2\underbrace{C(h_1, h_4)}_0 + 2\underbrace{C(h_2, h_3)}_0 + 2\underbrace{C(h_2, h_4)}_0 + 2C(h_3, h_4). \end{aligned}$$

Notice that this can be written

$$\text{Var}(\sum_{i=1}^4 h_i) = \text{Var}(h_1 + h_2) + \text{Var}(h_3 + h_4),$$

which hints at how this is typically estimated.

To estimate the S matrix in (19.18), we first define C clusters ($c = 1, \dots, C$) and then calculate

$$\hat{S} = \sum_{t=1}^T \sum_{c=1}^C h_t^c (h_t^c)' \text{, where } h_t^c = \sum_{i \in \text{cluster } c} h_{it}. \quad (19.19)$$

The iid case is when each i is her/his own cluster. The Driscoll-Kraay approach puts everyone in one cluster.

Remark 19.8 (*The logic of (19.19)**) To make a trivial case, let $N = 1$. According to (19.18) and the previous examples $S = T \text{Var}(h_1)$. The traditional estimate of $\text{Var}(h_1)$ is $\sum_{t=1}^T h_{1t}^2 / T$. Multiply by T to get the estimate of S as $\sum_{t=1}^T h_{1t}^2$.

Remark 19.9 ($T = 1^*$) When we have a pure cross-sectional data set, so $T = 1$, then we can still apply (19.19). However, it is common to replace h_i by $h_i C / (C - 1)$ to improve the small sample properties of S . Also, notice that the Driscoll-Kraay approach (letting everyone be in one big cluster) does not work when $T = 1$ since $\sum_{i=1}^N h_i = 0$.

	coef	t-tstat W	tstat DK
Inactive	-0.76	-56.89	-0.69
Active	3.08	37.48	1.77
Higly Active	8.65	28.73	2.73

Table 19.2: Annualised regression 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).

	coef	t-tstat W	tstat DK
Inactive	-1.10	-1.63	-0.69
Active	3.10	34.61	1.79
Higly 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 19.3: Annualised regression 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).

For an empirical illustration, see Table 19.2 where White's t-stats look massively inflated. In contrast, the Driscoll-Kraay (DK) t-stats are actually the same as in the calendar time approach in Table 19.1.

Table 19.3 extends the panel estimation in Table 19.2, but includes 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.

19.6 The Within Estimator (“Fixed Effects Estimator”)

In the fixed effects model, we allow for different individual intercepts

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}, \quad (19.20)$$

where u_{it} is iid with zero mean and variance σ_u^2 .

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 individual 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.

As an alternative (which gives the same point estimates as pooled OLS with dummies) consider the following approach. First, take average across time (for a given i) of y_{it} and x_{it} in (19.20). That is, think (but do not run any estimation yet...) of forming the cross-sectional regression

$$\bar{y}_i = \alpha_i + \bar{x}'_i \beta + \bar{u}_i, \text{ where} \quad (19.21)$$

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it} \text{ and } \bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}. \quad (19.22)$$

Second, transform the data as

$$y_{it}^* = y_{it} - \bar{y}_i \quad (19.23)$$

$$x_{it}^* = x_{it} - \bar{x}_i. \quad (19.24)$$

Use this to express the difference between (19.20) and (19.21) as

$$y_{it}^* = x_{it}^* \beta + u_{it}^*. \quad (19.25)$$

At this stage, estimate β by running pooled OLS on all observations of (19.25). There is no intercept in this regression. We denote this estimate $\hat{\beta}_{FE}$ (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. Third and finally, get estimates of individual intercepts as

$$\alpha_i = \bar{y}_i - \bar{x}'_i \hat{\beta}_{FE}. \quad (19.26)$$

Clearly, the within estimator wipes out all regressors that are constant across time for a given individual (say, gender and schooling): they are effectively merged with the individual means (μ_i). 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 19.4 for an example (see "south").

We can apply the usual tests on the pooled OLS results from (19.25)—provided the residuals are uncorrelated across time and individuals. Otherwise, we need to apply a cluster method.

Remark 19.10 (*Estimating $\text{Var}(\hat{\beta}_{FE})$) We can estimate the standard error of the residual as

$$\begin{aligned}\sigma_u^2 &= \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N \hat{u}_{it}^{*2}, \text{ where} \\ \hat{u}_{it}^* &= y_{it}^* - x_{it}^{*\prime} \hat{\beta}_{FE},\end{aligned}$$

and the covariance matrix of the slope coefficients as

$$\text{Var}(\hat{\beta}_{FE}) = \sigma^2 S_{xx}^{-1} \text{ where } S_{xx} = \sum_{t=1}^T \sum_{i=1}^N x_{it}^* x_{it}^{*\prime}.$$

Example 19.11 $N = 2, T = 2$. If we stack data for $t = T - 1$ ($i = 1$ and N) first and for $t = T$ second, then we have the following covariance matrix of the residuals u_{it}

$$\text{Cov} \begin{pmatrix} u_{1,T-1} \\ u_{N,T-1} \\ u_{1T} \\ u_{NT} \end{pmatrix} = \begin{bmatrix} \sigma_u^2 & 0 & 0 & 0 \\ 0 & \sigma_u^2 & 0 & 0 \\ 0 & 0 & \sigma_u^2 & 0 \\ 0 & 0 & 0 & \sigma_u^2 \end{bmatrix}.$$

This is a diagonal matrix.

Remark 19.12 (Lagged dependent variable as regressor*) If $y_{i,t-1}$ is among the regressors x_{it} , then the within estimator (19.25) is biased in short samples (that is, when T is small)—and increasing the cross-section (that is, N) does not help. To see the problem, suppose that the lagged dependent variable is the only regressor ($x_{it} = y_{i,t-1}$). The within estimator (19.25) is then

$$y_{it} - \sum_{t=1}^T y_{it}/T = \left(y_{i,t-1} - \sum_{t=2}^T y_{i,t-1}/(T-1) \right) \beta + \left(u_{it} - \sum_{t=1}^T u_{it}/T \right).$$

The problem is that $y_{i,t-1}$ is correlated with $\sum u_{it}$ since the latter contains $u_{i,t-1}$ which affects $y_{i,t-1}$ directly. In addition, $\sum y_{i,t-1}$ contains $y_{i,t}$ which is correlated with u_{it} . It can be shown that this bias can be substantial for panels with small T .

	LS	Fixed eff	Between	GLS
exper/100	7.84 (8.25)	4.11 (6.21)	10.64 (4.05)	4.57 (7.12)
exper ² /100	-0.20 (-5.04)	-0.04 (-1.50)	-0.32 (-2.83)	-0.06 (-2.37)
tenure/100	1.21 (2.47)	1.39 (4.25)	1.25 (0.90)	1.38 (4.32)
tenure ² /100	-0.02 (-0.85)	-0.09 (-4.36)	-0.02 (-0.20)	-0.07 (-3.77)
south	-0.20 (-13.51)	-0.02 (-0.45)	-0.20 (-6.67)	-0.13 (-5.70)
union	0.11 (6.72)	0.06 (4.47)	0.12 (3.09)	0.07 (5.57)

Table 19.4: 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.

19.6.1 The Within Estimator with Time Fixed Effects

When we allow for both time fixed effects and individual fixed effects

$$y_{it} = \lambda_t + \alpha_i + x'_{it}\beta + u_{it}, \quad (19.27)$$

then we could once again introduce dummies (now for both time periods and individuals) and apply pooled OLS.

As before, it is often easier to transform the data before estimating with pooled OLS. In this case, we run the regression on transformed variables

$$y_{it}^* = x'_{it}\beta + u_{it}^*. \quad (19.28)$$

The transformations are

$$y_{it}^* = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} \quad (19.29)$$

$$x_{it}^* = x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}, \quad (19.30)$$

where \bar{x}_i is defined in (19.22) and

$$\begin{aligned} \bar{x}_t &= \sum_{i=1}^N x_{it}/N \text{ and} \\ \bar{x} &= \sum_{t=1}^T \sum_{i=1}^N x_{it}/(TN). \end{aligned}$$

(Similarly for the transformation of y_{it} .) The last terms (\bar{y}, \bar{x}) makes sure that the grand mean of the transformed variable is zero. (If we instead add an intercept to (19.28), then this is not important for the slope coefficients.)

The estimation and testing of (19.28) is the same as for the standard within estimator (see above).

19.7 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)

$$\Delta y_{it} = \Delta \lambda_t + \Delta x'_{it} \beta + u_{it}^*, \quad (19.31)$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and similarly for the regressors. Notice that

$$u_{it}^* = u_{it} - u_{i,t-1}, \quad (19.32)$$

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 (19.32) can be estimated by pooled OLS. However, unadjusted standard errors are likely to overstate the uncertainty. This suggests that using the unadjusted standard errors is a conservative approach (harder to reject the null hypothesis).

Example 19.13 $N = 2, T = 2$. Stack the data for individual $i = 1$ first and those for individual $i = N$ second. If $u_{i,t}$ are iid, then

$$\text{Cov} \begin{pmatrix} u_{1,T-1}^* \\ u_{1T}^* \\ u_{N,T-1}^* \\ u_{NT}^* \end{pmatrix} = \text{Cov} \begin{pmatrix} u_{1,T-1} - u_{1,T-2} \\ u_{1,T} - u_{1,T-1} \\ u_{N,T-1} - u_{N,T-2} \\ u_{N,T} - u_{N,T-1} \end{pmatrix} = \begin{bmatrix} 2\sigma_u^2 & -\sigma_u^2 & 0 & 0 \\ -\sigma_u^2 & 2\sigma_u^2 & 0 & 0 \\ 0 & 0 & 2\sigma_u^2 & -\sigma_u^2 \\ 0 & 0 & -\sigma_u^2 & 2\sigma_u^2 \end{bmatrix}.$$

Remark 19.14 (*Lagged dependent variable as regressor**) If $y_{i,t-1}$ is among the regressors x_{it} , then the first-difference method (19.31) 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).

19.8 Differences-in-Differences Estimator

Consider the model (19.31) when one of the regressors is a dummy variable indicating whether individual 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. *If there are no other regressors*, then (19.31) can be written

$$\Delta y_{it} = \Delta \lambda_t + \beta Q_{it} + u_{it}^*, \quad (19.33)$$

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 ($\Delta \bar{y}_{A2}$)

$$\hat{\beta} = \Delta \bar{y}_{B2} - \Delta \bar{y}_{A2}. \quad (19.34)$$

(Notice that the change of the average is the same as the average of the change.)

19.9 Random Effects Model*

The random effects model allows for *random* individual “intercepts” (μ_i)

$$y_{it} = \alpha + x'_{it} \beta + \mu_i + \varepsilon_{it}, \text{ where} \quad (19.35)$$

$$\varepsilon_{it} \text{ is iid } N(0, \sigma_\varepsilon^2) \text{ and } \mu_i \text{ is iid } N(0, \sigma_\mu^2). \quad (19.36)$$

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

$$y_{it} = \alpha + x'_{it}\beta + u_{it}, \text{ where } u_{it} = \mu_i + \varepsilon_{it}, \quad (19.37)$$

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 (19.22))

$$\bar{y}_i = \alpha + \bar{x}'_i\beta + \underbrace{\mu_i + \bar{\varepsilon}_i}_{\text{residual}_i}, \quad (19.38)$$

is also consistent, but discards all time-series information. Third, LS on

$$y_{it} = \alpha + x'_{it}\beta + \underbrace{\mu_i + \varepsilon_{it}}_{\text{residual}_{it}} \quad (19.39)$$

is consistent (but not really efficient). However, in this case we may need to adjust $\text{Cov}(\hat{\beta})$ 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 19.15 $N = 2, T = 2$. If we stack the data for individual $i = 1$ first and those for individual $i = N$ second

$$\text{Cov} \begin{pmatrix} u_{1,T-1} \\ u_{1T} \\ u_{N,T-1} \\ u_{NT} \end{pmatrix} = \begin{bmatrix} \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 & 0 & 0 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 \\ 0 & 0 & \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 \end{bmatrix},$$

which has elements off the main diagonal.

Remark 19.16 (*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 (19.25) 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” (19.38). 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 (19.25). 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.

19.10 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

$$y_{it} = \lambda_t + x_{it}' \beta_t + u_{it} \quad (19.40)$$

period by period (using the cross section $i = 1 - N$). The FMB has the nice properties of easily handling unbalanced data sets (the cross-sectional regressions (19.40) are run the available cross section for each time period).

Second, estimate the time averages

$$\hat{\beta} = \frac{1}{T} \sum_{t=1}^T \hat{\beta}_t. \quad (19.41)$$

Remark 19.17 (*Step 0**) The FMB can also be used to test CAPM (or other linear factor models). In this case, y_{it} in (19.40) are the excess returns on asset i in period t (R_{it}^e) and x_{it} are the loadings (γ_{it}) of the excess return on the market excess return (or other

factors) according to the regression $R_{it}^e = \alpha + f_t' \gamma_{it} + \varepsilon_{it}$. In many cases, the γ_{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 γ_{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

$$\text{Var}(\hat{\beta}_t) = \frac{1}{T} \sum_{t=1}^T (\hat{\beta}_t - \hat{\beta})^2. \quad (19.42)$$

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,

$$\text{Var}(\hat{\beta}) = \frac{1}{T} \text{Var}(\hat{\beta}_t) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\beta}_t - \hat{\beta})^2. \quad (19.43)$$

When x_{it} are constant across time, then FMB and pooled OLS give the same point estimates (provided (19.40) is estimated without an intercept, effectively setting $\lambda_t = 0$). However, FMB's $\text{Var}(\hat{\beta})$ automatically handles the cross sectional correlations between residuals, while the pooled OLS would require applying a cluster method.

It can be noticed that when x_{it} is time-varying, then the FMB approach is not the same as OLS on pooled data. In fact, FMB is focused on the average cross-sectional effect, not on the time-series effect. For instance, regressions where all fixed effects have been taken out by demeaning are the same in FMB and pooled OLS.

Table 19.5 shows estimates of the factor risk premia from several methods based on the 25 FF portfolios.

	Data	CR	FMB1	FMB2
Rm	6.56 (1.92)	6.23 (2.02)	6.23 (1.94)	-8.01 (3.53)
SMB	2.33 (1.31)	2.22 (1.39)	2.22 (1.35)	1.88 (1.35)
HML	4.09 (1.21)	4.64 (1.34)	4.64 (1.25)	4.20 (1.24)

Table 19.5: Different estimates of factor risk premia, annualized %. Numbers in (parentheses) are standard deviations. The 25 FF portfolios 1957:01-2017:12. 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.

Chapter 20

Binary Choice and Truncated Models*

Reference: Verbeek (2012) 7

20.1 Binary Choice Model

20.1.1 Basic Model Setup

Consider the binary variable

$$y_i = \begin{cases} 0 & \text{firm } i \text{ doesn't pay dividends} \\ 1 & \text{firm } i \text{ pays dividends} \end{cases} \quad (20.1)$$

and suppose we know a few things about firm i : x_i (industry, size, profitability...).

We are interested in the probability that firm i pays dividends—and think it is some function of x_i

$$\Pr(y_i = 1|x_i) = F(x'_i \beta). \quad (20.2)$$

We want to study how x_i affects the probability that a firm pays dividends ($y_i = 1$), that is, we want to estimate β .

What function $F()$ should we use in (20.2)? Mostly a matter of convenience. A *probit model* assumes that $F()$ is a standard normal cumulative distribution function, see Figure 20.1. 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 20.2 for an illustration.

The results are often interpreted by looking at the marginal effects. For instance, the marginal effect of changing regressor k is

$$\frac{\partial F(x'_i \beta)}{\partial x_{k,i}} = f(x'_i \beta) \beta_k, \quad (20.3)$$

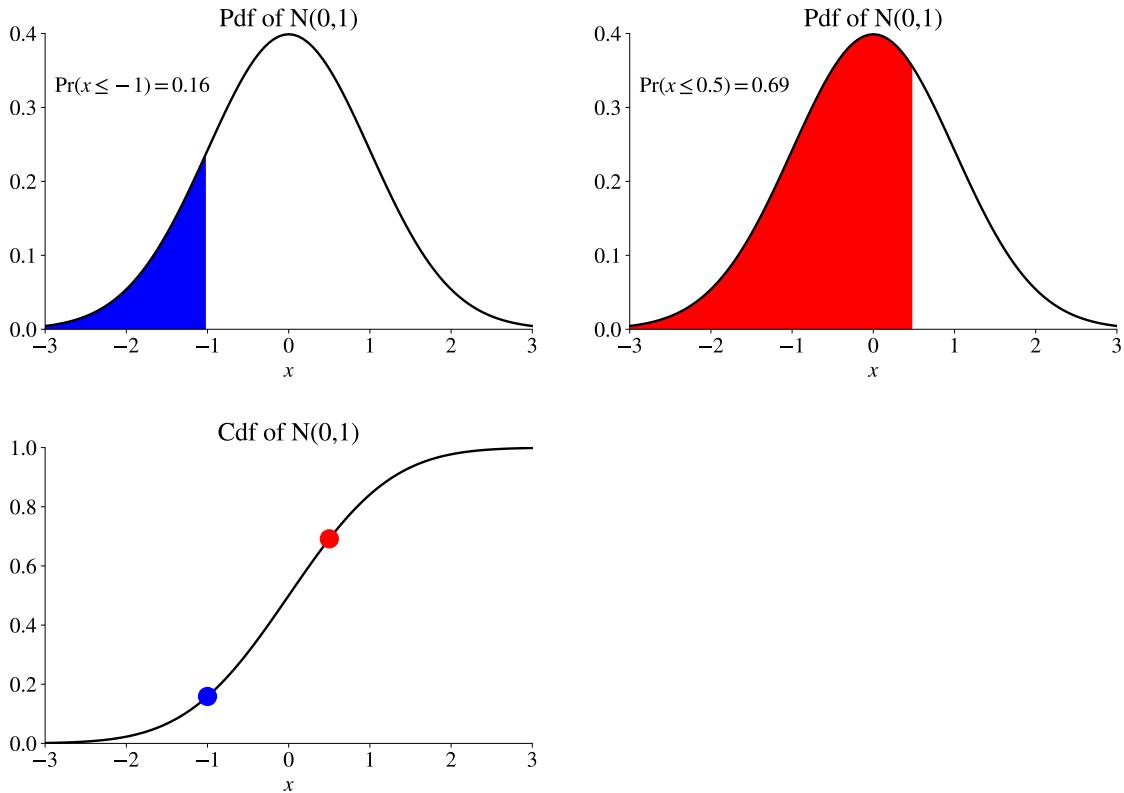


Figure 20.1: Pdf and cdf of $N(0,1)$

where $f()$ is the derivative of $F()$. (The notation $x_{k,i}$ is meant to indicate regressor k for firm i and β_k is the coefficient on regressor k) This is calculated at some typical value x_t (for instance, at the sample average of the regressors). This could, for instance, answer the question: how does the probability of having dividends change when profits change? Notice that if the $F()$ function is increasing (all three alternatives mentioned above), then the derivative (20.3) has the same sign as β_k (since $f() > 0$).

Example 20.1 Constant plus two more regressors (w and z): $x_i' \beta = \beta_0 + \beta_1 w_i + \beta_2 z_i$, then

$$\frac{\partial F(x_i' \beta)}{\partial w} = f(\beta_0 + \beta_1 w_i + \beta_2 z_i) \beta_1,$$

where $f()$ is the derivative of $F()$. This is calculated at some typical values of (w_i, z_i) .

Example 20.2 If a regressor is a dummy variable, then use a simple difference instead of attempting a derivative. For instance, if z_i is either 0 or 1, then we can use

$$F(\beta_0 + \beta_1 w_i + \beta_2) - F(\beta_0 + \beta_1 w_i).$$

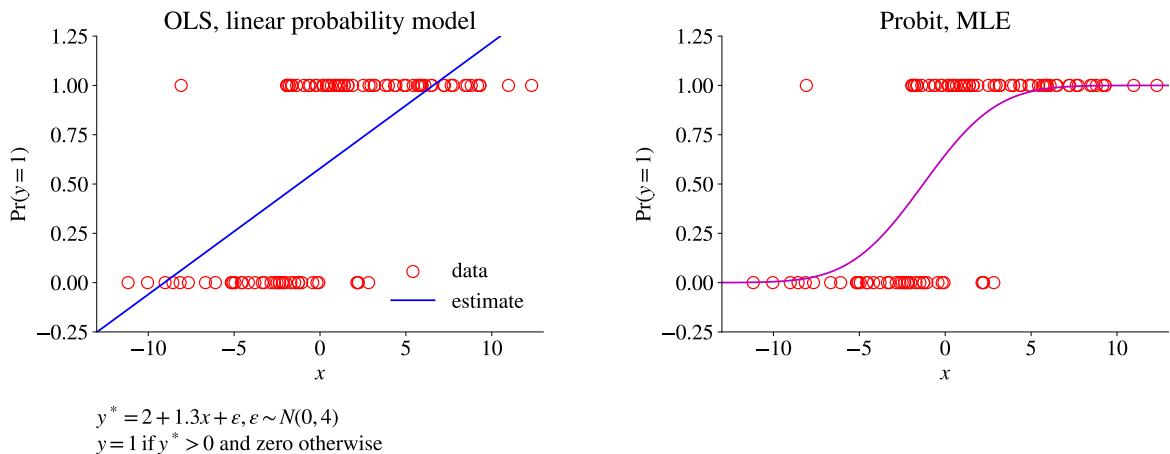


Figure 20.2: Example of probit model

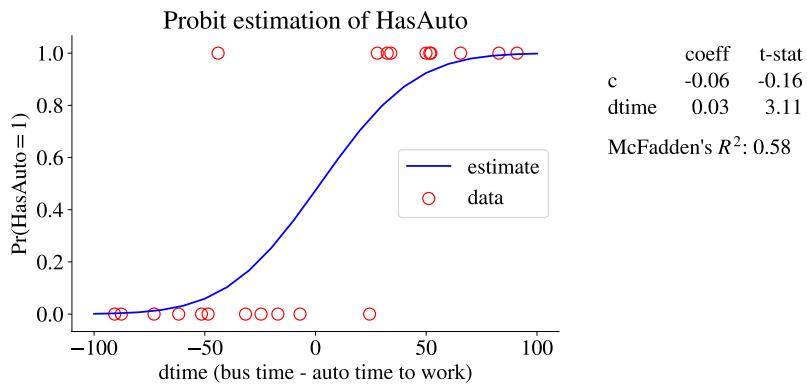


Figure 20.3: Example of probit model, Hill et al (2008), Table 16.1

This is calculated at some typical value of w_i .

Notice from (20.3) that the ratio of two coefficients equals the ratio of their marginal effect on the probability

$$\beta_k / \beta_m = \frac{\partial F(x'_i \beta)}{\partial x_{k,i}} / \frac{\partial F(x'_i \beta)}{\partial x_{m,i}}.$$

20.1.2 Estimation

The model is typically estimated with MLE. To do that we need to construct the likelihood function

$$\ln L = \sum_{i=1}^N y_i \ln F(x'_i \beta) + (1 - y_i) \ln [1 - F(x'_i \beta)]. \quad (20.4)$$

We find the ML estimate by maximizing this log likelihood function with respect to the parameters β . See Figure 20.3 for an empirical example.

Proof. (of (20.4)*) Recall that a Bernoulli distribution is specified as $\Pr(y_i = 1) = p_i$, $\Pr(y_i = 0) = 1 - p_i$. Assume independent observations (here, firm 1 and 2).

Then, the probabilities (likelihoods) for the different outcomes are

$$\begin{aligned}\Pr(y_1 = 1 \text{ and } y_2 = 1) &= p_1 p_2 \\ \Pr(y_1 = 1 \text{ and } y_2 = 0) &= p_1(1 - p_2) \\ \Pr(y_1 = 0 \text{ and } y_2 = 1) &= (1 - p_1) p_2 \\ \Pr(y_1 = 0 \text{ and } y_2 = 0) &= (1 - p_1)(1 - p_2).\end{aligned}\tag{20.5}$$

At this point p_i are just symbols for the probabilities. Soon, we use replace them with the specification in (20.2). This will be long (and messy to program) when there are many observations (firms). We therefore use an alternative way of writing the same thing. Notice that

$$p_i^{y_i} (1 - p_i)^{1-y_i} = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{if } y_i = 0. \end{cases}$$

For the sample with two data points (firm 1 and 2), the probability (likelihood) can be written

$$L = p_1^{y_1} (1 - p_1)^{1-y_1} \times p_2^{y_2} (1 - p_2)^{1-y_2}.$$

Let $p_i = F(x'_i \beta)$ from (20.2) and use in the likelihood function

$$L = F(x'_1 \beta)^{y_1} [1 - F(x'_1 \beta)]^{1-y_1} \times F(x'_2 \beta)^{y_2} [1 - F(x'_2 \beta)]^{1-y_2}.$$

or as log (after slight rearranging)

$$\begin{aligned}\ln L &= y_1 \ln F(x'_1 \beta) + y_2 \ln F(x'_2 \beta) \\ &\quad + (1 - y_1) \ln [1 - F(x'_1 \beta)] + (1 - y_2) \ln [1 - F(x'_2 \beta)].\end{aligned}$$

The extension to N data points is straightforward. ■

20.1.3 Goodness of Fit

To measure the fit, we use several different approaches—since a traditional R^2 is not appropriate for a non-linear model.

First, McFadden's R^2 is a commonly applied measure that has many features in com-

mon with a traditional R^2 . It is

$$\text{McFadden's } R^2 = 1 - \frac{\log \text{likelihood value (at max)}}{\log \text{likelihood value (all coeffs=0, except constant)}}. \quad (20.6)$$

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 (20.6) is therefore between 0 (as bad as a model with only a constant) and 1 (a perfect model).

Example 20.3 If $\ln L = \ln 0.9$ (at max) and the model with only a constant has $\ln L = \ln 0.5$

$$\text{McFadden's } R^2 = 1 - \frac{\ln 0.9}{\ln 0.5} \approx 0.84$$

If instead, the model has $\ln L = \ln 0.8$ (at max), then

$$\text{McFadden's } R^2 = 1 - \frac{\ln 0.8}{\ln 0.5} \approx 0.68$$

An alternative measure of the goodness of fit is an “ R^2 ” for the predicted probabilities. To compare predictions to data, let the “predictions” be

$$\hat{y}_i = \begin{cases} 1 & \text{if } F(x'_i \hat{\beta}) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (20.7)$$

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. We now cross-tabulate 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

There are $n_{01} + n_{10}$ incorrect predictions: n_{01} are the number of cases when $y_i = 0$ but $\hat{y}_i = 1$, and n_{10} for the opposite case ($y_i = 1$ but $\hat{y}_i = 0$). Define an “ R^2_{pred} ” for the prediction as

$$\text{“}R^2_{pred}\text{”} = 1 - \frac{\text{number of incorrect predictions}}{\text{number of incorrect predictions, constant probabilities}}. \quad (20.9)$$

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 (constant probability).

The constant probability is just the fraction of data where the binary variable equals one

$$\begin{aligned}\hat{p} &= \text{Fraction of } (y_i = 1) \\ &= N_1/N,\end{aligned}\tag{20.10}$$

where $N_1 = n_{10} + n_{11}$ are the number of cases when $y_i = 1$ and N is the total number of data points.

When $\hat{p} \leq 0.5$, the (naive) constant probability model always predicts $y_i = 0$, so the number of incorrect predictions (denominator of (20.9)) is N_1 . Otherwise it is N_0 . For the estimated model, the number of incorrect predictions (when $\hat{y}_i \neq y_i$) is $n_{10} + n_{01}$. This gives the “ R^2_{pred} ” in (20.9) as

$$\text{“}R^2_{pred}\text{”} = \begin{cases} 1 - \frac{n_{01} + n_{10}}{N_1} & \text{if } \hat{p} \leq 0.5 \\ 1 - \frac{n_{01} + n_{10}}{N_0} & \text{if } \hat{p} > 0.5. \end{cases}$$

Example 20.4 Let x_i be a scalar. Suppose we have the following data

$$\begin{aligned}\begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 1 & 1 \end{bmatrix} \text{ and} \\ \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} &= \begin{bmatrix} 1.5 & -1.2 & 0.5 & -0.7 \end{bmatrix}\end{aligned}$$

See Figure 20.4

Suppose $\beta = 0$, then we get the following values

$$F(x'_i \beta) = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}$$

$$\begin{aligned}y_i \log F(x'_i \beta) + (1 - y_i) \log [1 - F(x'_i \beta)] \\ \approx \begin{bmatrix} -0.69 & -0.69 & -0.69 & -0.69 \end{bmatrix}\end{aligned}$$

$$\log L \approx -2.77$$

Now, suppose instead that $\beta = 1$

$$F(x'_i \beta) \approx \begin{bmatrix} 0.93 & 0.12 & 0.69 & 0.24 \end{bmatrix}$$

$$\begin{aligned}
& y_i \log F(x'_i \beta) + (1 - y_i) \log [1 - F(x'_i \beta)] \\
& \approx \begin{bmatrix} -0.07 & -0.12 & -0.37 & -1.42 \end{bmatrix} \\
& \log L \approx -1.98,
\end{aligned}$$

which is higher than at $\beta = 0$. If $\beta = 1$ happened to maximize the likelihood function (it almost does...), then

$$\text{McFadden's } R^2 = 1 - \frac{-1.98}{-2.77} \approx 0.29$$

and the predicted would be

$$\hat{y}_i \approx \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}.$$

Cross-tabulation of actual (y_i) and predicted (\hat{y}_i) values

	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Total
$y_i = 0:$	1	0	1
$y_i = 1:$	1	2	3
Total:	2	2	4

Since the constant probability is

$$\hat{p} = 3/4,$$

the constant probability model always predicts $y_i = 1$. We therefore get

$$\text{"}R^2_{pred}\text{"} = 1 - \frac{1}{1+0} = 0.$$

20.1.4 Related Models

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).

Models for count data are useful for answer questions like: “how many visits to the supermarket this week?” They are like a standard model, but y_i can only take on integer values (0, 1, 2, 3, ..).

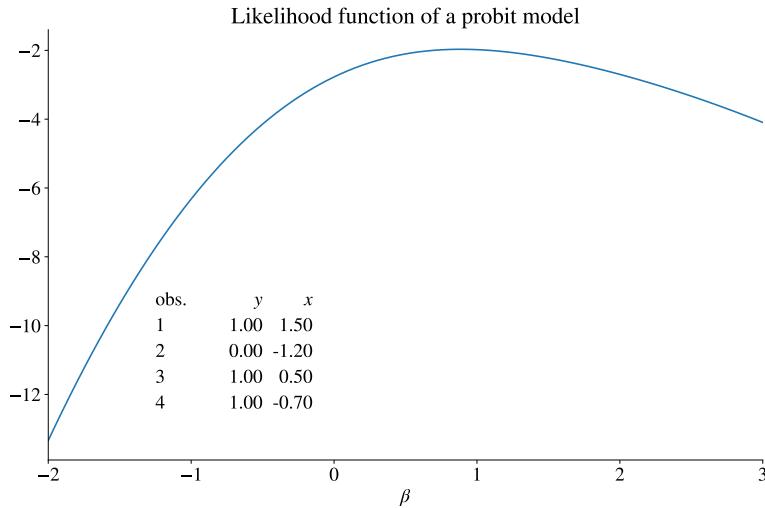


Figure 20.4: Example of ML estimation of probit model

20.2 Truncated Regression Model

20.2.1 Basic Model Setup

Suppose the correct model is linear

$$y_i^* = x_i' \beta + \varepsilon_i, \quad \varepsilon_i \sim \text{iid}N(0, \sigma^2), \quad (20.11)$$

but that data (also regressors) is completely missing if $y_i^* \leq 0$

$$\begin{aligned} y_i &= y_i^* && \text{if } y_i^* > 0 \\ (y_i, x_i) &\text{ not observed} && \text{otherwise.} \end{aligned} \quad (20.12)$$

The cutoff at zero is just a normalisation that could easily be changed.

The problem with this 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. See Figure 20.7 for an illustration. In fact, running OLS to estimate

$$y_i = x_i' \beta + \varepsilon_i \quad (20.13)$$

on the available data will give biased (and inconsistent) estimates.

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

$$\varepsilon_i > -x'_i \beta, \quad (20.14)$$

since otherwise (20.13) becomes negative (and those are the observations that do not enter the sample). This which ε_i realizations that enter the sample is not random any more—and it depends on the x_i value (when $x'_i \beta < 0$, then only positive u_i realizations enter the sample and vice versa). This correlation of u_i and x_i is a classical reason for why OLS estimates of the regression coefficients are inconsistent.

Remark 20.5 (*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

$$E(y_i | y_i > 0, x_i) = x'_i \beta + E(\varepsilon_i | y_i^* > 0) \quad (20.15)$$

$$= x'_i \beta + E(\varepsilon_i | \varepsilon_i > -x'_i \beta). \quad (20.16)$$

The second line follows from the fact that $y_i^* > 0$ happens when $x'_i \beta + \varepsilon_i > 0$ (see (20.12)) which happens when $\varepsilon_i > -x'_i \beta$. The key result is that the last term is positive and is correlated with x_i . (It is positive since $E(\varepsilon_i | \varepsilon_i > -\infty) = 0$, and here we are conditioning on $\varepsilon_i > -x'_i \beta$ which depends on x_i .)

20.2.2 Examples of the Effect of Truncation

This section presents a useful result on truncated normal distributions —and then uses that in a few simple examples of how the truncation affects the OLS estimates.

Remark 20.6 (*Truncated normal distribution*) Let $\varepsilon \sim N(\mu, \sigma^2)$, then

$$E(\varepsilon | \varepsilon > a) = \mu + \sigma \frac{\phi(a_0)}{1 - \Phi(a_0)} \text{ and } a_0 = (a - \mu)/\sigma$$

See Figure 20.5.

Example 20.7 As a trivial example, suppose the model is $y_i^* = 0 + \varepsilon_i$ with $\varepsilon_i \sim iidN(0, 1)$. Then

$$\begin{aligned} E(y_i | y_i > 0, x_i) &= 0 + E(\varepsilon_i | \varepsilon_i > 0) \\ &= 0 + \frac{\phi(0)}{1 - \Phi(0)} = \sqrt{2/\pi} \approx 0.80, \end{aligned}$$

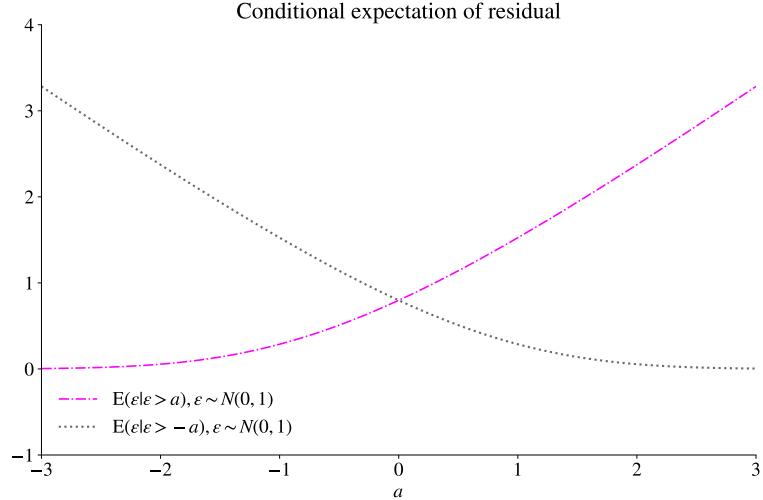


Figure 20.5: Expectations of a truncated variable

which is far from the true mean (0). OLS will therefore estimate an intercept of around 0.8 instead of 0.

Example 20.8 Suppose the model is $y_i^* = 2x_i + \varepsilon_i$ with $\varepsilon_i \sim iidN(0, 1)$ and where x_i a scalar random variable. Then

$$\begin{aligned} E(y_i | y_i > 0, x_i) &= 2x_i + E(\varepsilon_i | \varepsilon_i > -2x_i) \\ &= 2x_i + \frac{\phi(-2x_i)}{1 - \Phi(-2x_i)} \end{aligned}$$

For some selected values of x_i we have

$$\begin{aligned} E(y_i | y_i > 0, x_i) &= \\ &= 2x_i + E(\varepsilon_i | \varepsilon_i > -2x_i) \\ &= \begin{cases} 2 \times (-1) + E(\varepsilon_i | \varepsilon_i > 2) & x = -1 \\ 2 \times 0 + E(\varepsilon_i | \varepsilon_i > 0) & x_i = 0 \\ 2 \times 1 + E(\varepsilon_i | \varepsilon_i > -2) & x_i = 1 \end{cases} \\ &= \begin{cases} 2 \times (-1) + 2.37 = 0.37 & x = -1 \\ 2 \times 0 + 0.8 = 0.80 & x_i = 0 \\ 2 \times 1 + 0.06 = 2.06 & x_i = 1 \end{cases} \end{aligned}$$

so the slope is lower than 2: OLS will therefore fit a slope coefficient that is lower than 2. See Figure 20.6. The basic point is that $E(\varepsilon_i | \varepsilon_i > -2x_i)$ is much higher for low than

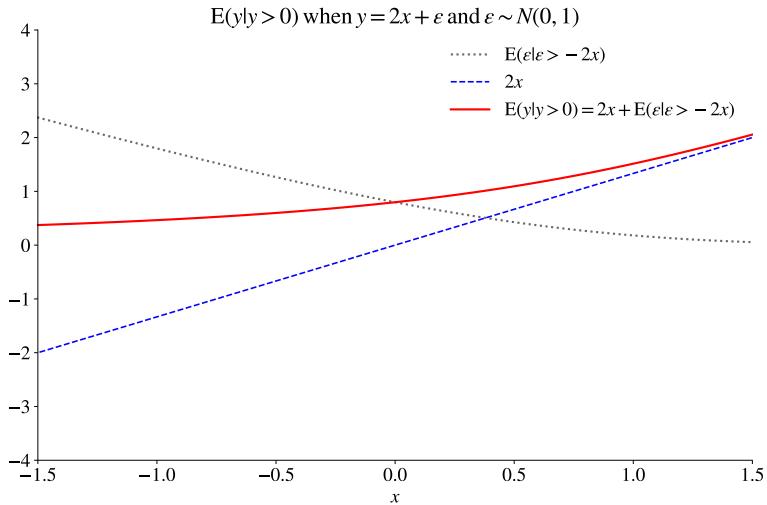


Figure 20.6: Expectations of a truncated variable

for high values of x_i (compare $x_i = -1$ and $x_i = 1$), making the regression line look flatter. (Notice that $\frac{\phi(-2x_i)}{1-\Phi(-2x_i)}$ can also be written $\frac{\phi(2x_i)}{\Phi(2x_i)}$ since the $N(0,1)$ distribution is symmetric around zero.)

20.2.3 Estimation

Remark 20.9 (*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 $\text{pdf}(\varepsilon|a < \varepsilon \leq b) = \text{pdf}(\varepsilon)/\Pr(a < \varepsilon \leq b)$.

The log likelihood function is

$$\begin{aligned}\ln L &= \sum_{i=1}^N \ln L_i, \text{ where} \\ L_i &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - x'_i \beta)^2}{\sigma^2}\right) / \Phi(x'_i \beta / \sigma).\end{aligned}\tag{20.17}$$

We maximize this likelihood function with respect to β and σ^2 (numerical optimization). Notice that $\Phi(x'_i \beta / \sigma)$ is the new part compared with OLS. See Figure 20.7 for an illustration.

Proof. (of (20.17)*) Here, we need the density function of y_i conditional on $y_i > 0$

(or equivalently of ε_i), conditional on $y_i^* = x_i' \beta + \varepsilon_i > 0$ (so $\varepsilon_i > -x_i' \beta$). This is

$$\text{pdf}(\varepsilon_i | \varepsilon_i > -x_i' \beta) = \frac{\text{pdf}(\varepsilon_i)}{\Pr(\varepsilon_i > -x_i' \beta)}.$$

To specify this, we notice the following. *First*, if $\varepsilon_i \sim N(0, \sigma^2)$, the denominator is

$$\begin{aligned}\Pr(\varepsilon_i > -x_i' \beta) &= \Pr\left(\varepsilon_i/\sigma > -x_i' \beta/\sigma\right) \\ &= 1 - \Phi\left(-x_i' \beta/\sigma\right) \\ &= \Phi\left(x_i' \beta/\sigma\right).\end{aligned}$$

The last line follows from $N(0, 1)$ being symmetric around 0, so $\Phi(z) = 1 - \Phi(-z)$. *Second*, the numerator (in the first equation of this proof) is the pdf of an $N(0, \sigma^2)$ variable. *Third*, combining and replacing ε_i by $y_i - x_i' \beta$, taking logs gives (20.17). ■

20.3 Censored Regression Model (Tobit Model)

The censored regression model is similar to truncated model, but we are here fortunate to always observe the regressors. x_i . We have a bit *more information* than in truncated case, and we should try to use it. In short, the model and data are

$$y_i^* = x_i' \beta + \varepsilon_i, \varepsilon_i \sim \text{iid}N(0, \sigma^2) \quad (20.18)$$

$$\text{Data: } (y_i, x_i) = \begin{cases} (y_i^*, x_i) & \text{if } y_i^* > 0 \\ (0, x_i) & \text{otherwise.} \end{cases}$$

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*. As an example: y_i^* is investment into stocks by a household, x_i is household income—and households with low income are assigned a common value (normalized to $y_i = 0$) in the survey.

If we estimate

$$y_i = x_i' \beta + u_i \quad (20.19)$$

with LS, using all data with $y_i > 0$, then we are in same situation as in truncated model: LS is not consistent. See Figure 20.7.

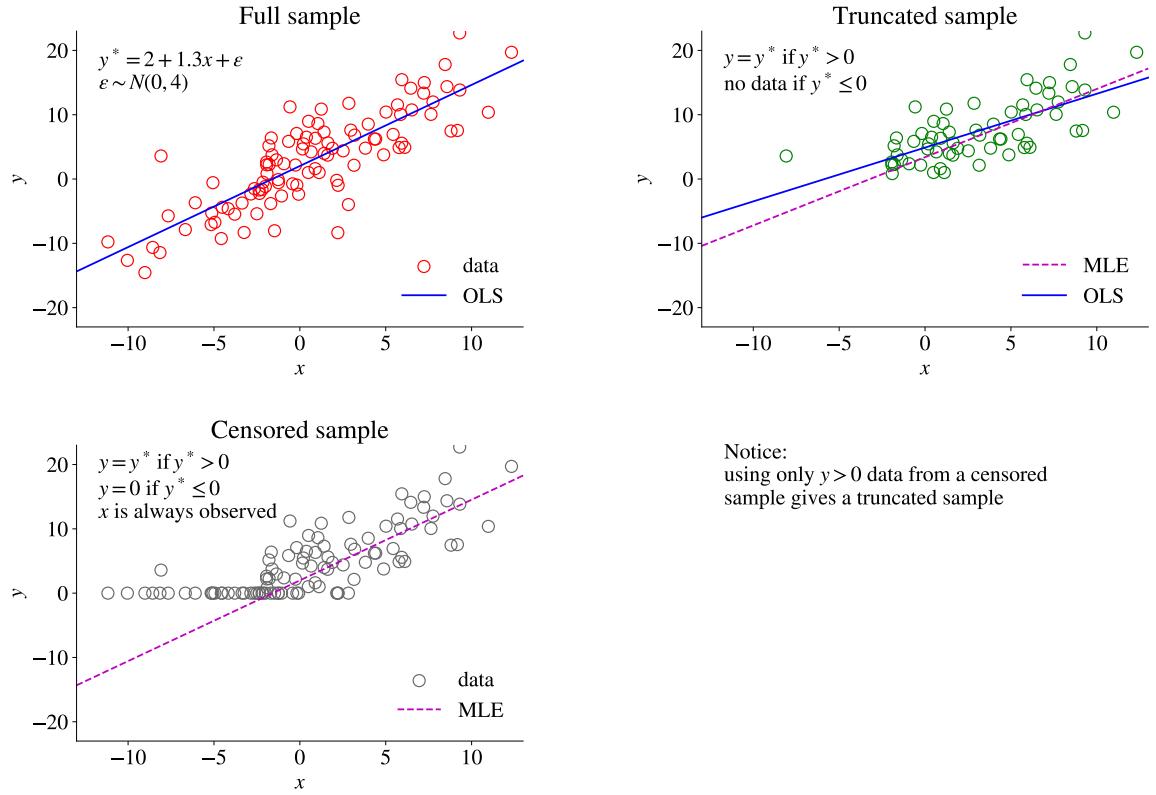


Figure 20.7: Estimation on full, truncated and censored sample

20.3.1 Estimation of Censored Regression Model

Remark 20.10 (*Likelihood function with different “states”*) The likelihood contribution of observation i is $\text{pdf}(y_i)$ which can also be written $\text{pdf}(y_i|\text{state } K) \times \Pr(\text{state } K)$. 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 y_i and x_i). This gives the likelihood function

$$L = \sum_{i: \text{no data on } y_i} \Phi(-x_i' \beta / \sigma) + \sum_{i: \text{with data}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - x_i' \beta)^2}{\sigma^2}\right). \quad (20.20)$$

Maximize with respect to β and σ^2 (numerical optimization). 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 (20.20)*) State $y_i^* \leq 0$ happens when $y_i^* = x_i' \beta + \varepsilon_i \leq 0$, that is, when

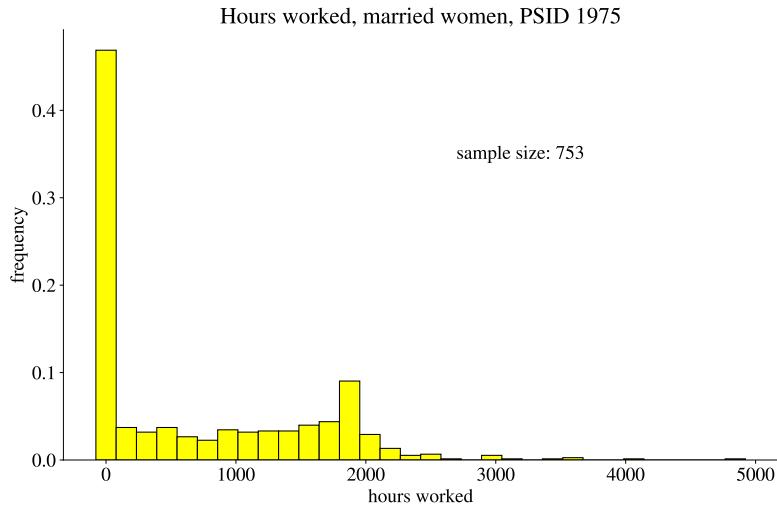


Figure 20.8: Example of probit model, Hill et al (2008), Table 16.1

$\varepsilon_i \leq -x'_i \beta$. The probability of this is

$$\begin{aligned} \Pr(\varepsilon_i \leq -x'_i \beta) &= \Pr(\varepsilon_i/\sigma \leq -x'_i \beta/\sigma) \\ &= \Phi(-x'_i \beta/\sigma). \end{aligned}$$

The conditional density function in this state has the constant value of one, so the likelihood contribution (see Remark 20.10) 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 (20.19), but the difference here is that the contribution to the likelihood function (again, see Remark 20.10) is

$$\begin{aligned} L_i(\text{if } y_i^* > 0) &= \text{pdf}(\varepsilon_i | \varepsilon_i > -x'_i \beta) \times \Pr(\varepsilon_i > -x'_i \beta) \\ &= \text{pdf}(\varepsilon_i), \end{aligned}$$

where $\text{pdf}(\varepsilon_i)$ is the pdf of $N(0, \sigma^2)$. ■

	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 20.1: 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).

20.3.2 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 derivative provides an answer

$$\frac{\partial \Pr(y_i = 0)}{\partial x_{k,i}} = -\phi(x_i'\beta/\sigma)\beta_k/\sigma. \quad (20.21)$$

This derivative is high (in absolute value) when $x_i'\beta \approx 0$, since a small change in x_k can then tip the balance towards $y_i = 0$. (Formally, this follows from the fact that the pdf $\phi()$ peaks when its argument is zero.) 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 (as we are already safely in $y_i = 0$ or $y_i = 1$ territory). *Second*, how is the expected value of y_i affected by a change in regressor k ? Once again, we can calculate a derivative

$$\frac{\partial \mathbb{E} y_i}{\partial x_{k,i}} = \Phi(x_i'\beta/\sigma)\beta_k. \quad (20.22)$$

This derivative depends on $x_i'\beta$. For low values of $x_i'\beta$, the derivative 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 . (To derive these two results, recall that $\mathbb{E} y_i = (1 - Q)0 + Q \mathbb{E}(y_i | y_i^* > 0)$, where Q is the probability of $y_i^* > 0$. Then use the remark on expected values from truncated normal distributions.)

20.4 Heckit: 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

$$w_i^* = x_{1i}' \beta_1 + \varepsilon_{1i} \quad (20.23)$$

$$h_i^* = x_{2i}' \beta_2 + \varepsilon_{2i}. \quad (20.24)$$

For instance, 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 0/1 (doesn't work/works). In short,

$$\text{Data: } (w_i, h_i) = \begin{cases} w_i = w_i^*, h_i = 1 & \text{if } h_i^* > 0 \\ w_i \text{ not observed, } h_i = 0 & \text{otherwise.} \end{cases} \quad (20.25)$$

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 denoted w_i^*). This requires that (20.24) is the same (or at least proportional) to (20.23)

It is typical to assume that the residuals in the two equations could be correlated

$$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix} \right). \quad (20.26)$$

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

$$E(w_i | h_i = 1) = x_{1i}' \beta_1 + \sigma_{12} \lambda_i, \text{ where } \lambda_i = \frac{\phi(x_{2i}' \beta_2)}{\Phi(x_{2i}' \beta_2)}, \quad (20.27)$$

where $\phi()$ and $\Phi()$ are the standard normal pdf and cdf (λ_i is called the inverse Mill's ratio or Heckman's lambda). The point of (20.27) is that the covariance of the residuals in the two equations (20.23)–(20.24) is crucial. In fact, when $\sigma_{12} = 0$, then we can estimate

(20.23) with OLS. Otherwise, it is biased (and inconsistent).

Proof. (of (20.27))

$$\begin{aligned} E(w_i | h_i = 1) &= x'_{1i} \beta_1 + \underbrace{E(\varepsilon_{1i} | h_i = 1)}_{E(\varepsilon_{1i} | \varepsilon_{2i} > -x'_{2i} \beta_2)} \\ &= x'_{1i} \beta_1 + E(\varepsilon_{1i} | \varepsilon_{2i} > -x'_{2i} \beta_2), \end{aligned}$$

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 (20.26), then it is a standard result that

$$E(\varepsilon_{1i} | \varepsilon_{2i} > -q) = \sigma_{12} \lambda, \text{ where } \lambda = \frac{\phi(q)}{\Phi(q)}.$$

(Showing this is straightforward, but a bit tedious.) ■

Another way to see the problem highlighted by (20.27) is the following. Consider the observable data (when $h_i = 1$)

$$w_i = x'_{1i} \beta_1 + \varepsilon_{1i} \quad (20.28)$$

and ask if $E(x_{1i} \varepsilon_{1i}) = 0$ for this data? To keep it simple, suppose x_{2i} includes just a constant: w_i 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 $\sigma_{12} \neq 0$)? Well, then OLS of (20.28) 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}). As an extreme case, consider $x_{1i} = x_{2i}$.

20.4.1 Estimation

Use MLE or Heckman's 2-step approach, which is as follows:

1. Estimate (20.24) with the probit method (recall $h_i = 0$ or 1) using the likelihood function in (20.20). Extract $x'_{2i} \hat{\beta}_2$ and create $\hat{\lambda}_i$ as in (20.27). See Table 20.2 for an

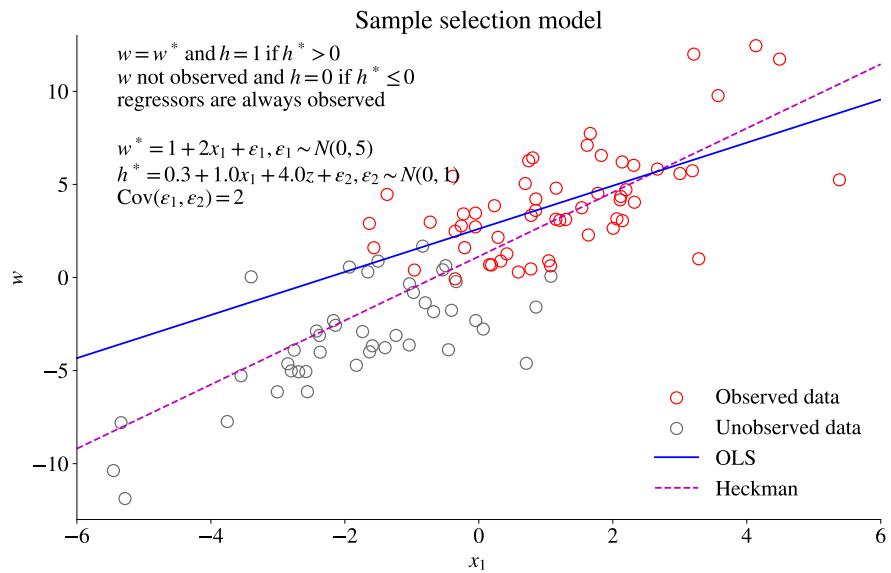


Figure 20.9: Sample selection model

example.

2. Estimate $(\beta_1$ and $\sigma_{12})$ with LS

$$w_i = x'_{1i}\beta_1 + \sigma_{12}\hat{\lambda}_i + \eta_i \quad (20.29)$$

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 $\sigma_{12} = 0$).

c	-4.16
	(-2.96)
age	0.19
	(2.78)
age2	-0.00
	(-3.10)
faminc	0.00
	(1.00)
kids	-0.45
	(-3.53)
educ	0.10
	(4.35)

Table 20.2: 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	-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)
	0	

Table 20.3: 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.

Chapter 21

LAD and Quantile Regressions*

21.1 LAD

Reference: Amemiya (1985) 4.6, Greene (2012) 7.3, Wooldridge (2010) 12.10

The least absolute deviations (LAD) estimator minimizes the sum of absolute residuals (rather than the squared residuals)

$$\hat{b}_{LAD} = \arg \min_b \sum_{t=1}^T |y_t - x'_t b| \quad (21.1)$$

The optimization is a non-linear problem, but a simple iteration works nicely (see below). The estimator is typically less sensitive to outliers than OLS. (There are also other ways to estimate robust regression coefficients.) This is illustrated in Figure 21.1.

See Figure 21.2 for an empirical example.

If we assume that the median of the true residual, u_t , is zero, then (under strict assumptions, discussed below) we have

$$\begin{aligned} \sqrt{T}(\hat{b}_{LAD} - b_0) &\xrightarrow{d} N[0, f(0)^{-2} \Sigma_{xx}^{-1}/4], \text{ where} \\ \Sigma_{xx} &= \text{plim} \sum_{t=1}^T x_t x'_t / T, \end{aligned} \quad (21.2)$$

where $f(0)$ is the value of the pdf of the residual at zero. Unless we know this density function (or else we would probably have used MLE instead of LAD), we need to estimate it—for instance with a kernel density method. However, to arrive at the result in (21.2) we must assume that the residual is independent of the regressors. (This is discussed in some detail below, see quantile regressions).

Example 21.1 ($N(0, \sigma^2)$) When $u_t \sim N(0, \sigma^2)$, then $f(0) = 1/\sqrt{2\pi\sigma^2}$ (since $e^0 = 1$),

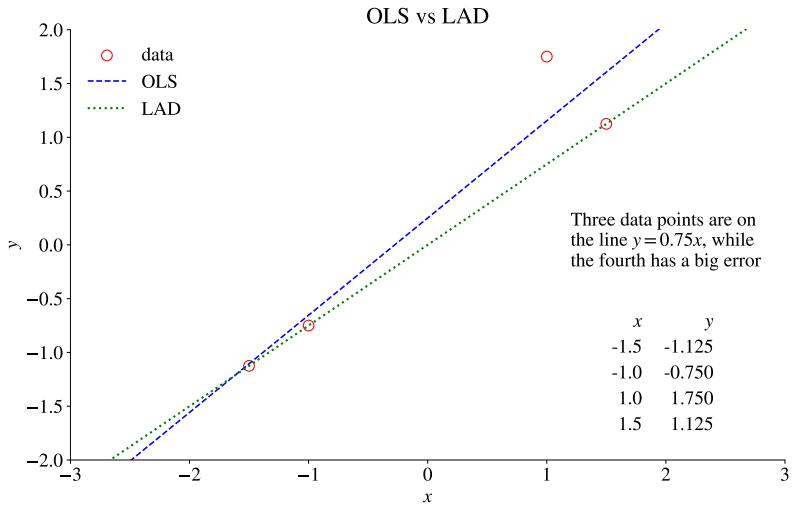


Figure 21.1: Data and regression line from OLS and LAD

so the covariance matrix in (21.2) becomes $\pi\sigma^2\Sigma_{xx}^{-1}/2$. This is $\pi/2$ times larger than when using LS.

Remark 21.2 (Algorithm for LAD) The LAD estimator can be written

$$\hat{b}_{LAD} = \arg \min_b \sum_{t=1}^T w_t \hat{u}_t(b)^2, \text{ where}$$

$$w_t = 1/|\hat{u}_t(b)|, \text{ with}$$

$$\hat{u}_t(\hat{b}) = y_t - x'_t \hat{b}$$

so it is a weighted least squares where both y_t and x_t are multiplied by $1/|\hat{u}_t(\hat{b})|$. It can be shown that iterating on LS with the weights given by $1/|\hat{u}_t(\hat{b})|$, where the residuals are from the previous iteration, converges very quickly to the LAD estimator.

21.1.1 Reinterpreting the LAD

Consider a linear regression

$$y_t = x'_t b + u_t. \quad (21.3)$$

In the OLS context we typically assume $E u_t = 0$ and $\text{Cov}(x_t, u_t) = 0$. The latter is the same as $E(u_t|x_t) = 0$ which means that

$$E(y_t|x_t) = x'_t b. \quad (21.4)$$

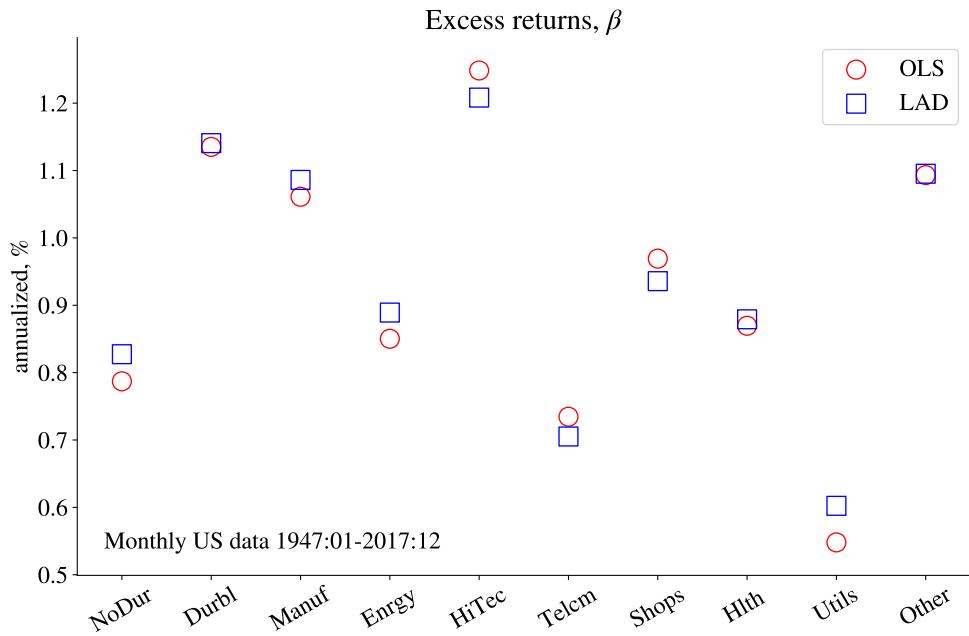


Figure 21.2: Betas of US industry portfolios

We can interpret the LAD estimator as an alternative way of getting good estimates of b , 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.

Remark 21.3 ($E(u|x) = 0$ or $(E u_t = 0, \text{Cov}(x_t, u_t) = 0)^*$) For any random variables u and x ,

$$\text{Cov}(x, u) = \text{Cov}[x, E(u|x)].$$

The condition $E(u|x) = 0$ therefore implies $\text{Cov}(x, u) = 0$. It also implies $E u = 0$ since $E u = E_x[E(u|x)] = E_x[0] = 0$.

Remark 21.4 (Mean and median as solutions to minimization problems) If u is a random variable, then the mean, μ , is the solution to $\min_{\mu} E(u - \mu)^2$, while the median, m , is the solution to $\min_m E |u - m|$. (There are some restrictions on u for this to be true, but we disregard that here.)

The previous remark shows that the LAD estimator (21.1) amounts to finding the b

coefficients (in a linear model) so that

$$\text{Median}(u_t|x_t) = 0, \text{ which implies} \quad (21.5)$$

$$\text{Median}(y_t|x_t) = x_t'b. \quad (21.6)$$

This is the alternative interpretation of the LAD: it tries to set the median of the residuals, at a given x_t vector, equal to zero. In contrast, OLS tries to set the mean of the residuals, at a given x_t vector, to zero.

21.2 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 (21.5), it rather states that the q th quantile (conditional on x_t) of the residual is zero

$$Q(u_t|x_t; q) = 0, \text{ which implies} \quad (21.7)$$

$$Q(y_t|x_t; q) = x_t'b^{(q)}. \quad (21.8)$$

Here $Q(u_t|x_t; q)$ denotes the q th quantile of u_t at a particular value of x_t and we also index the coefficients $b^{(q)}$ to remember that this refers to the q th quantile. Clearly, the LAD is the special case when $q = 0.5$.

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 21.3 for an example where the slopes differ across the quantiles and Figure 21.4 where they do not. In particular, in Figure 21.3 the data follows a *location and scale model*

$$y_t = x_t'\beta + u_t \text{ where } u_t = x_t'\gamma\varepsilon_t, \text{ and } \varepsilon_t \text{ is iid.} \quad (21.9)$$

This is basically a linear model ($y_t = x_t'\beta$), 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 (0.9, say) do not necessarily describe the relation between y_t and x_t at generally (unconditionally) high y_t (or x_t) values—see Figure

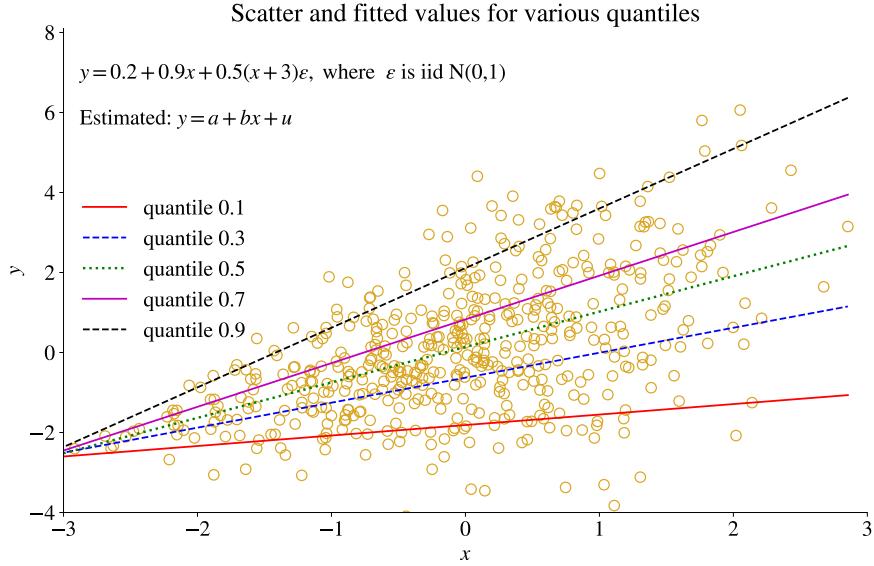


Figure 21.3: Example of quantile regression

21.3. Rather, the slopes describes the relation between y_t and x_t at high ε_t values. This is perhaps best seen by using the location and scale model in (21.9) which implies

$$Q(y_t|x_t; q) = x_t' \beta + Q(u_t|x_t; q) \quad (21.10)$$

$$= x_t' \beta + x_t' \gamma Q(\varepsilon_t; q) \quad (21.11)$$

$$= x_t' [\beta + \gamma Q(\varepsilon_t; q)]. \quad (21.12)$$

(In the second line, $Q(\varepsilon_t; q)$ need not be conditioned on x_t since ε_t is independent of x_t .) Comparing with (21.8) shows that

$$b^{(q)} = \beta + \gamma Q(\varepsilon_t; q). \quad (21.13)$$

For instance, if $\gamma > 0$, then $b^{(q)}$ is increasing with q (since $Q(\varepsilon_t; q)$ is). This is the case illustrated in Figure 21.3—where the higher slopes at high quantiles basically capture heteroskedasticity. In contrast, Figure 21.4 shows the case where the γ coefficient on the non-constant regressors are all zero: the $b^{(q)}$ coefficients (except for the constants) are the same across quantiles.

Figure 21.5 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 neg-

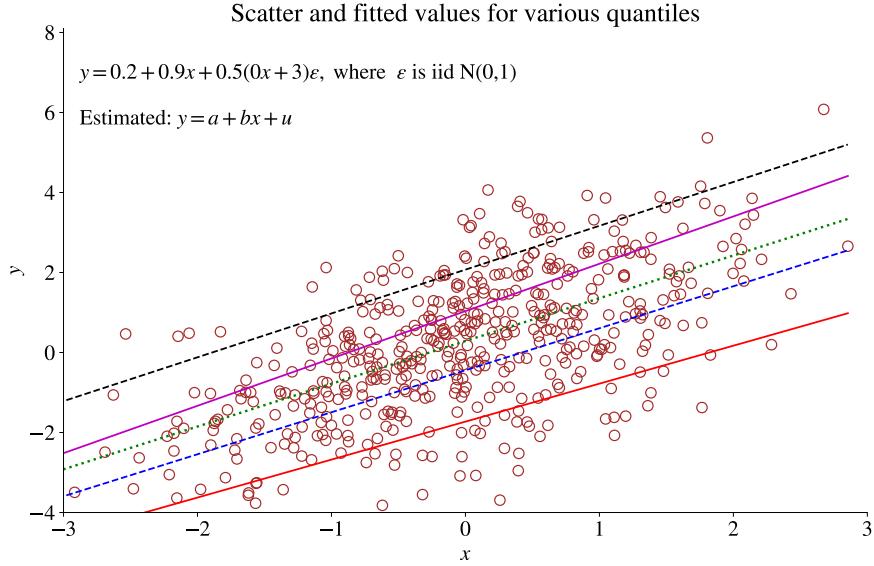


Figure 21.4: Example of quantile regression

ative return to be followed by positive returns and vice versa). More pronounced is the smaller 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, $b^{(q)}$, solves the following problem

$$\min_{b^{(q)}} \sum_{t:u_t \geq 0} q|u_t| + \sum_{t:u_t < 0} (1-q)|u_t|, \text{ where } u_t = y_t - x'_t b^{(q)}. \quad (21.14)$$

This is a highly non-linear problem (and the objective function does not have continuous derivatives), which can be solved by either linear programming method or a derivative-free minimization algorithm. As a special case, $q = 0.5$ gives the LAD where (21.14) becomes

$$\min_{b^{(0.5)}} 0.5 \sum_{t=1}^T |u_t|, \text{ where } u_t = y_t - x'_t b^{(q)}, \quad (21.15)$$

which is clearly the same as (21.1).

Remark 21.5 (Alternative way of writing (21.14)) Suppose $u_1 \geq 0$ and $u_2 < 0$, then the sum in (21.14) can be written $qu_1 + (q-1)u_2$. This suggests that if we define a dummy d_t to be 1 if $u_t < 0$ and zero otherwise, then we can write the sum as $(q-d_1)u_1 + (q-d_2)u_2$.

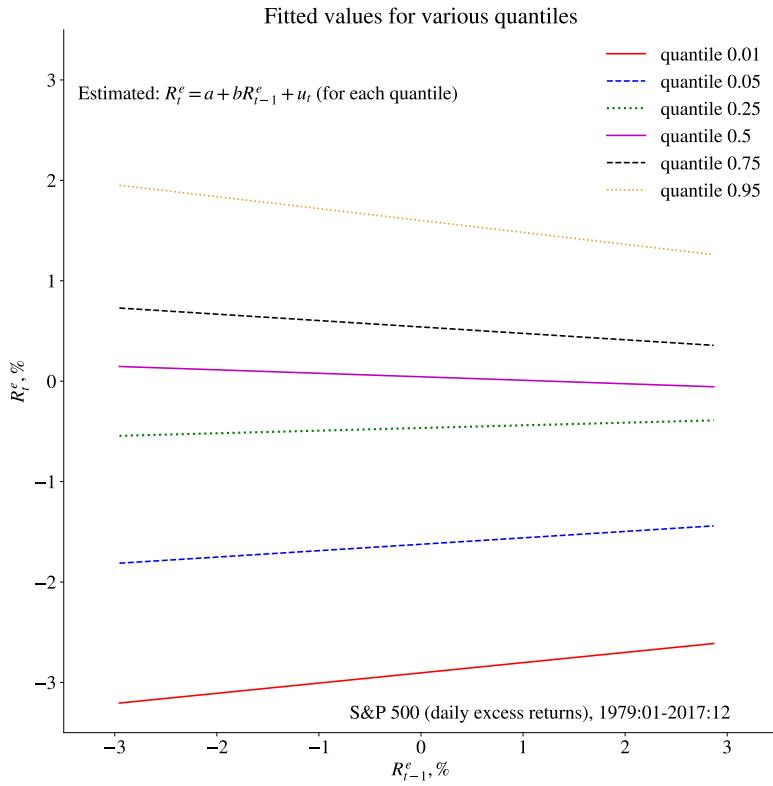


Figure 21.5: Quantile regression of AR(1) of daily returns

In general, the minimisation problem can then be written

$$\min_{b^{(q)}} \sum_{t=1}^T (q - d_t) u_t.$$

The term $(q - d_t) u_t$ is sometimes written as a function $\rho_q(u_t)$, so the sum becomes $\sum_{t=1}^T \rho_q(u_t)$. As a function of u_t , the $\rho_q(u_t)$ function has a (skewed) v-shape around $u_t = 0$. For $u_t < 0$ the function is $(q - 1)u_t$ so it is linear with a (negative) slope of $q - 1$, and for u_t it is also linear but with a slope of q .

The asymptotic distribution of the estimates is typically

$$\begin{aligned} \sqrt{T}(\hat{b}^{(q)} - b_0^{(q)}) &\xrightarrow{d} N[0, q(1-q)C^{-1}\Sigma_{xx}C^{-1}], \text{ where} \\ \Sigma_{xx} &= \text{plim } \sum_{t=1}^T x_t x_t' / T \text{ and} \\ C &= \text{plim } \sum_{t=1}^T f(0|x_t) x_t x_t' / T, \end{aligned} \quad (21.16)$$

where $f(0|x_t)$ is the value of the pdf of the residual, conditional on the regressor value,

at a zero residual. If x_t 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 (21.2) once we set $q = 0.5$.

One way of obtaining a consistent estimate of C is via a kernel density estimate

$$C = \sum_{t=1}^T w_t x_t x_t' / T, \text{ with} \quad (21.17)$$

$$w_t = \frac{1}{h\sqrt{2\pi}} \exp \left[-(\hat{u}_t/h)^2 / 2 \right],$$

and where \hat{u}_t is the fitted residual and h could be chosen as $h = 1.06 \text{Std}(\hat{u}_t)T^{-1/5}$ (see Silverman (1986)). 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. Notice that the product of w_t and $x_t x_t'$ in (21.17) is aimed at capturing the fact that $f(0|x_t)$ and $x_t x_t'$ are related—something that was ruled out in (21.2).

Bibliography

- Ait-Sahalia, Y., 1996, “Testing continuous-time models of the spot interest rate,” *Review of Financial Studies*, 9, 385–426.
- Ait-Sahalia, Y., and A. W. Lo, 1998, “Nonparametric estimation of state-price densities implicit in financial asset prices,” *Journal of Finance*, 53, 499–547.
- Alexander, C., 2008a, *Market Risk Analysis: Practical Financial Econometrics*, Wiley.
- Alexander, C., 2008b, *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, A., and J. Chen, 2002, “Asymmetric correlations of equity portfolios,” *Journal of Financial Economics*, 63, 443–494.
- Ang, J. S., and S. J. Ciccone, 2001, “International differences in analyst forecast properties,” mimeo, Florida State University.
- 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.
- Bawa, V. S., and E. B. Lindenberg, 1977, “Capital market equilibrium in a mean-lower partial moment framework,” *Journal of Financial Economics*, 5, 189–200.

- 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.
- Bodie, Z., A. Kane, and A. J. Marcus, 2005, *Investments*, McGraw-Hill, Boston, 6th 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., 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., 2001, *Asset pricing*, Princeton University Press, Princeton, New Jersey.
- Copeland, T. E., J. F. Weston, and K. Shastri, 2005, *Financial theory and corporate policy*, Pearson Education, 4 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, 2016, “Individual Investor Activity and Performance,” forthcoming in *The Review of Financial Studies*.
- 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.
- Diebold, F. X., 2001, *Elements of forecasting*, South-Western, 2nd edn.
- Diebold, F. X., and R. S. Mariano, 1995, “Comparing predcitve accuracy,” *Journal of Business and Economic Statistics*, 13, 253–265.
- 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, 2003, *Modern portfolio theory and investment analysis*, John Wiley and Sons, 6th edn.

- Elton, E. J., M. J. Gruber, S. J. Brown, and W. N. Goetzmann, 2010, *Modern portfolio theory and investment analysis*, John Wiley and Sons, 8th edn.
- Enders, W., 2004, *Applied econometric time series*, John Wiley and Sons, New York, 2nd edn.
- 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.
- Gourioux, 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., 2000, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.
- Greene, W. H., 2003, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 5th edn.

- Greene, W. H., 2012, *Econometric analysis*, Pearson Education Ltd, Harlow, Essex, 7th edn.
- 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.
- Hentschel, L., 1995, “All in the family: nesting symmetric and asymmetric GARCH models,” *Journal of Financial Economics*, 39, 71–104.
- Heston, S. L., and S. Nandi, 2000, “A closed-form GARCH option valuation model,” *Review of Financial Studies*, 13, 585–625.
- Horowitz, J. L., 2001, “The Bootstrap,” in J.J. Heckman, and E. Leamer (ed.), *Handbook of Econometrics* . , vol. 5, Elsevier.
- Huberman, G., and S. Kandel, 1987, “Mean-variance spanning,” *Journal of Finance*, 42, 873–888.
- Hull, J. C., 2006, *Options, futures, and other derivatives*, Prentice-Hall, Upper Saddle River, NJ, 6th edn.
- Jondeau, E., S.-H. Poon, and M. Rockinger, 2007, *Financial Modeling under Non-Gaussian Distributions*, Springer.
- 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., 2006, *Derivatives markets*, Addison-Wesley, 2nd 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.
- Murphy, J. J., 1999, *Technical analysis of the financial markets*, New York Institute of Finance.
- Nantell, T. J., and B. Price, 1979, “An analytical comparison of variance and semivariance capital market theories,” *Journal of Financial and Quantitative Analysis*, 14, 221–242.
- Neely, C. J., 1997, “Technical analysis in the foreign exchange market: a layman’s guide,” *Federal Reserve Bank of St. Louis Review*.
- Nelson, D. B., 1991, “Conditional heteroskedasticity in asset returns,” *Econometrica*, 59, 347–370.
- Newbold, P., 1995, *Statistics for business and economics*, Prentice-Hall, 4th edn.
- Pagan, A., and A. Ullah, 1999, *Nonparametric econometrics*, Cambridge University Press.
- 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.
- Silverman, B. W., 1986, *Density estimation for statistics and data analysis*, Chapman and Hall, London.
- 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.

The Economist, 1993, “Frontiers of finance,” pp. 5–20.

Verbeek, M., 2012, *A guide to modern econometrics*, Wiley, 4th edn.

Wooldridge, J. M., 2002, *Econometric analysis of cross section and panel data*, MIT Press.

Wooldridge, J. M., 2010, *Econometric analysis of cross section and panel data*, MIT Press, 2nd edn.