Lecture Notes in Financial Econometrics (MSc course)

Paul Söderlind[1]

13 June 2013

[1] University of St. Gallen. *Address:* s/bf-HSG, Rosenbergstrasse 52, CH-9000 St. Gallen, Switzerland. *E-mail:* Paul.Soderlind@unisg.ch. Document name: FinEcmtAll.TeX

Contents

1 Review of Statistics

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

1.1 Random Variables and Distributions

1.1.1 Distributions

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

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

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

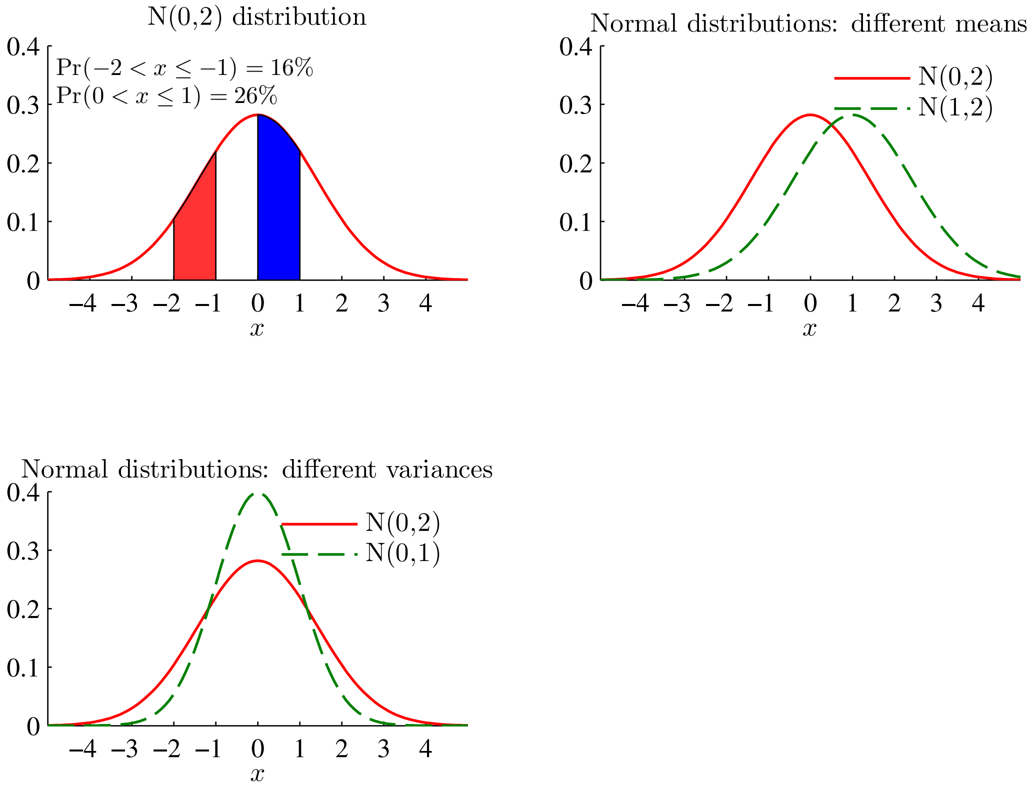
Remark 1.1 *If $x \sim N(\mu, \sigma^2)$, then the probability density function is*

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

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

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

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


N(0,2) distribution

Normal distributions: different means

Normal distributions: different variances

Figure 1.1: A few different normal distributions

A joint normal distributions is completely described by the means and the covariance matrix

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

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. Some alternative notations are used: $\mathrm{E}\,x$ for the mean, $\mathrm{Std}(x)$ for the standard deviation, $\mathrm{Var}(x)$ for the variance and $\mathrm{Cov}(x, y)$ for the covariance.

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. See Figure 1.2 for an example. The correlation of x and y is defined as

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{1.4}$$

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 $k(y)$)

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

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

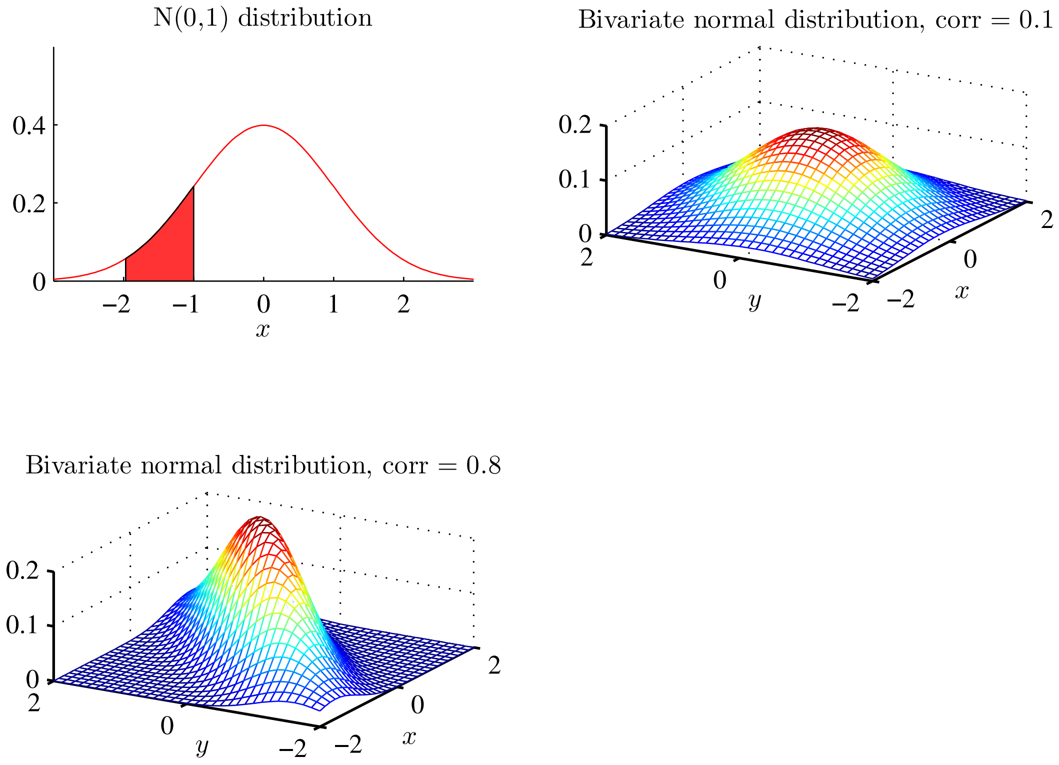


Figure 1.2: Density functions of univariate and bivariate normal distributions

1.1.2 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

$$g(y|x) = h(x, y)/f(x). \tag{1.6}$$

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

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

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 forecast error (using the conditional mean as the forecast). The conditional and marginal distribution coincide if y is uncorrelated with x. (This follows directly from combining (1.5) and (1.6)). Otherwise, the mean of the conditional distribution depends on x, and the variance is smaller than in the marginal distribution (we have more information). See Figure 1.3 for an illustration.

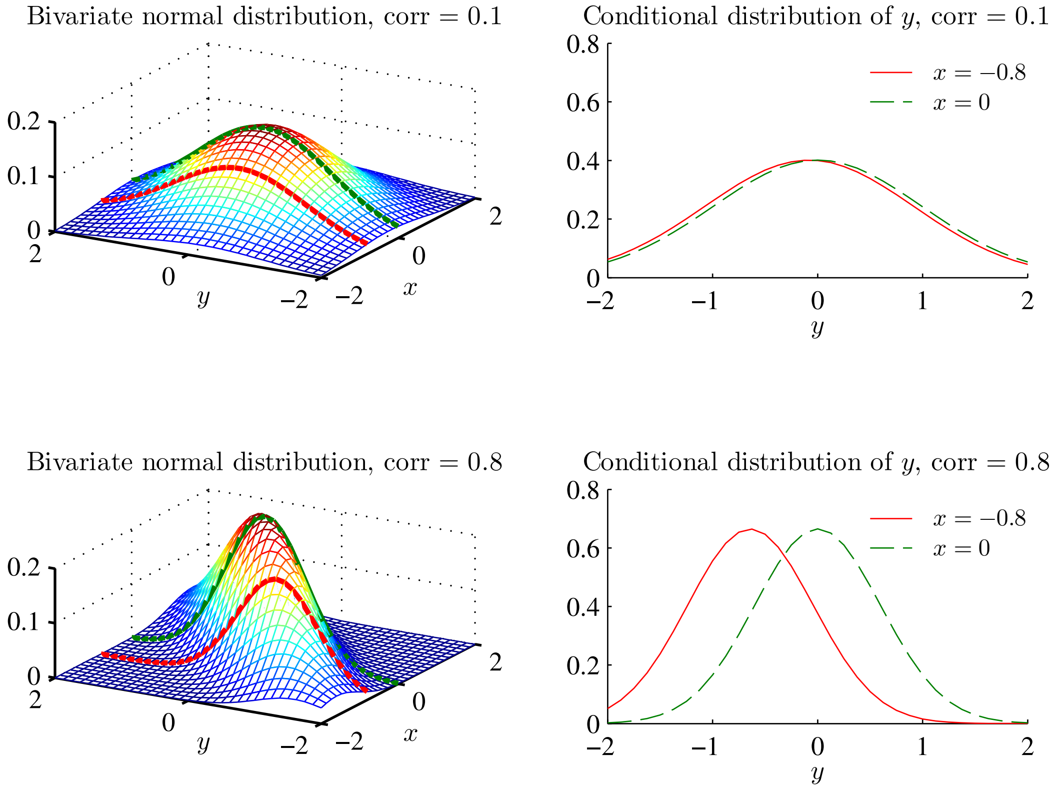


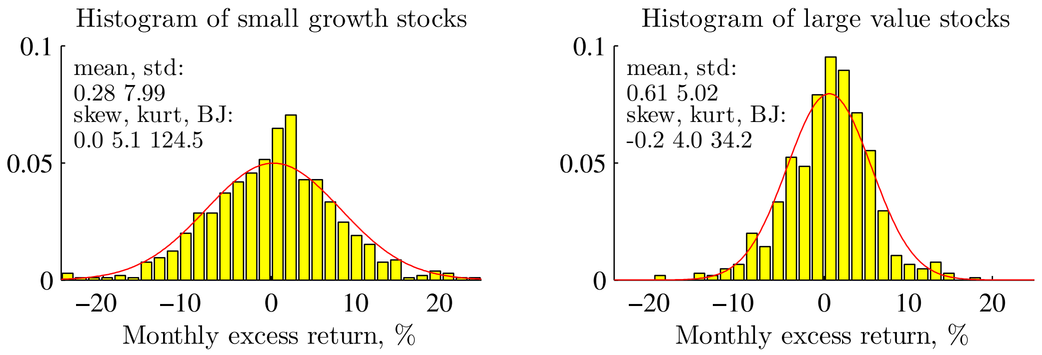
Figure 1.3: Density functions of normal distributions

1.1.3 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 distribution has).

See Figure 1.4 for an illustration.



Monthly data on two U.S. indices, 1957:1-2012:12
Sample size: 672

Figure 1.4: Histogram of returns, the curve is a normal distribution with the same mean and standard deviation as the return series

1.1.4 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

$$\Pr\left(\mu - c \leq x < \mu + c\right) = 0.9. \tag{1.8}$$

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

$$\begin{aligned} &\Pr(\mu - 1.65\sigma \leq x < \mu + 1.65\sigma) = 0.9 \text{ and} \\ &\Pr(\mu - 1.96\sigma \leq x < \mu + 1.96\sigma) = 0.95. \end{aligned} \tag{1.9}$$

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

$$[\hat{\beta} - 1.65\sigma, \hat{\beta} + 1.65\sigma]. \tag{1.10}$$

In case this band does not include zero, then we would be 90% that the (true) regression coefficient is different from zero.

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

$$[0 - 1.65\sigma, 0 + 1.65\sigma]. \tag{1.11}$$

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 zero. This latter approach is virtually the same as doing a t-test, that, by checking if

$$\left| \frac{\hat{\beta} - 0}{\sigma} \right| > 1.65. \tag{1.12}$$

To see that, notice that if (1.12) holds, then

$$\hat{\beta} < -1.65\sigma \text{ or } \hat{\beta} > 1.65\sigma, \tag{1.13}$$

which is the same as $\hat{\beta}$ being outside the confidence band in (1.11).

1.2 Moments

1.2.1 Mean and Standard Deviation

The mean and variance of a series are estimated as

$$\bar{x} = \textstyle\sum_{t=1}^{T} x_t / T \text{ and } \hat{\sigma}^2 = \textstyle\sum_{t=1}^{T} (x_t - \bar{x})^2 / T. \tag{1.14}$$

The standard deviation (here denoted $\text{Std}(x_t)$), 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 1.4 for an illustration.

A sample mean is normally distributed if x_t is normal distributed, $x_t \sim N(\mu, \sigma^2)$. The basic reason is that a linear combination of normally distributed variables is also normally distributed. However, a sample average is typically 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

$$\text{Var}(\bar{x}) = \sigma^2/T, \text{ if } x_t \text{ is iid.} \tag{1.15}$$

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

$$\mathrm{E}\,\bar{x} = \mathrm{E}\,x_t = \mu. \tag{1.16}$$

Since sample averages are typically normally distributed in large samples (according to the central limit theorem), we thus have

$$\bar{x} \sim N(\mu, \sigma^2/T), \tag{1.17}$$

so we can construct a *t-stat* as

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{T}}, \tag{1.18}$$

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

Proof. (of (1.15)–(1.16)) To prove (1.15), notice that

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

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 (1.16)

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

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

$$\begin{array}{lcll} & & \underline{\text{Test statistic}} & \underline{\text{Distribution}}\\ \text{skewness} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^3 & N\left(0, 6/T\right)\\ \text{kurtosis} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^4 & N\left(3, 24/T\right)\\ \text{Bera-Jarque} & = & \frac{T}{6}\text{skewness}^2 + \frac{T}{24}\left(\text{kurtosis}-3\right)^2 & \chi^2_2. \end{array} \qquad (1.19)$$

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (1.19) 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

$$t = \frac{\text{skewness}}{\sqrt{6/T}} \text{ and } t = \frac{\text{kurtosis} - 3}{\sqrt{24/T}}, \tag{1.20}$$

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

See Figure 1.4 for an illustration.

1.2.3 Covariance and Correlation

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

$$\hat{\sigma}_{xy} = \sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y})/T. \tag{1.21}$$

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

The correlation of two variables is then estimated as

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}, \tag{1.22}$$

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 1.5.

See Figure 1.6 for an empirical illustration.

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

$$\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \sim N(0, 1/T), \tag{1.23}$$

so we can form a t-stat as

$$t = \sqrt{T} \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}, \tag{1.24}$$

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

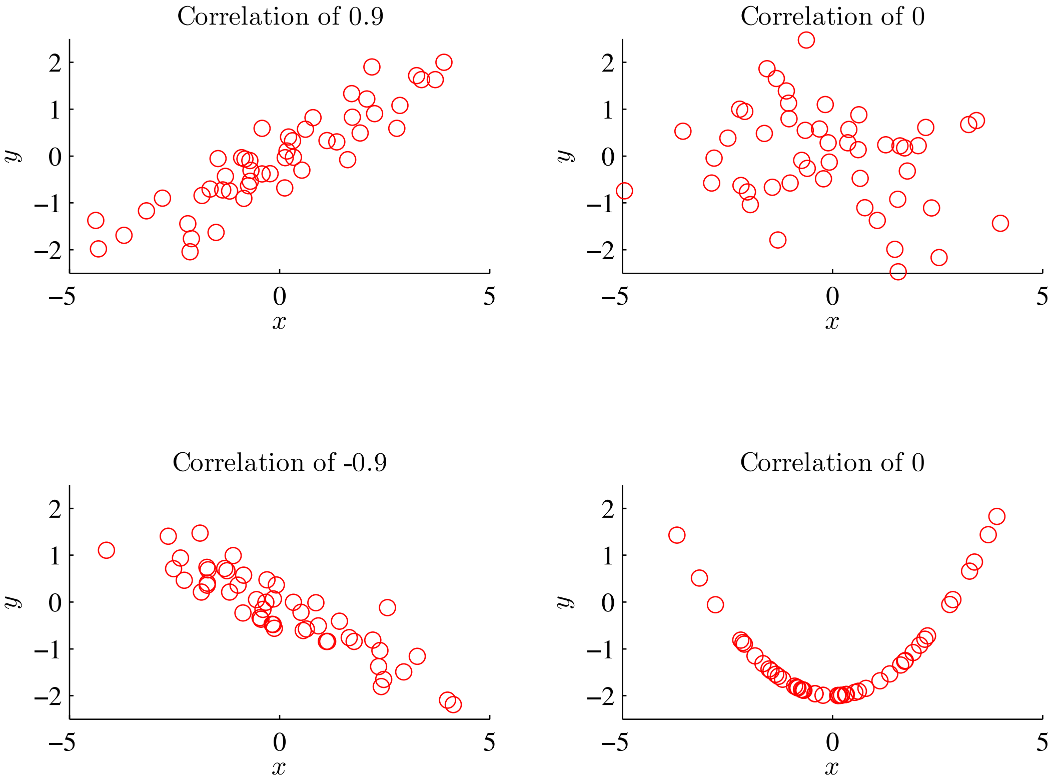


Figure 1.5: Example of correlations.

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 test statistic has a standard normal distribution

$$z = \frac{x - \mu}{\sigma} \sim N(0, 1). \tag{1.25}$$

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

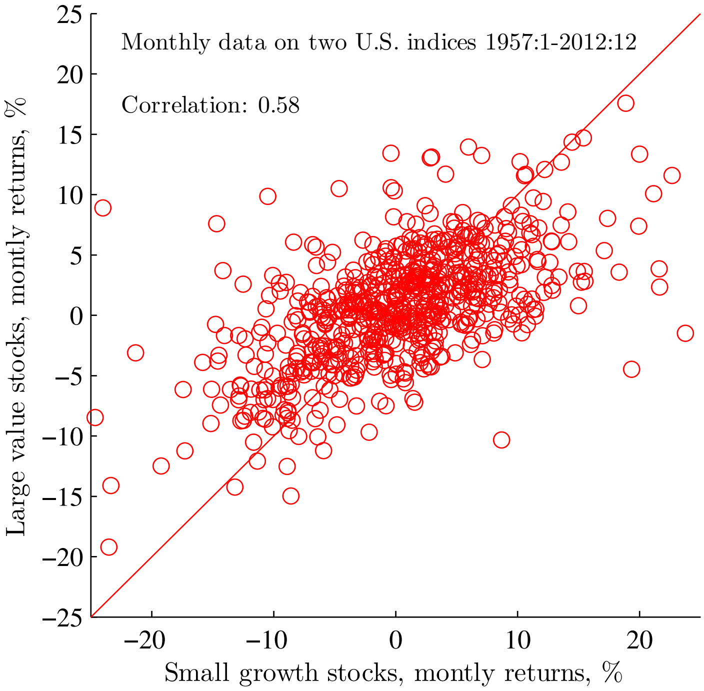


Figure 1.6: Scatter plot of two different portfolio returns

1.3.2 t-distribution

If we instead need to estimate σ to use in (1.25), then the test statistic has t_{df}-distribution

$$t = \frac{x - \mu}{\hat{\sigma}} \sim t_n, \tag{1.26}$$

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: 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 1.7 for a comparison and Table A.1 for critical values.

Example 1.2 *(t-distribution) If $t = 2.0$ and $n = 50$, then this is larger than the10% critical value (but not the 5% critical value) for a 2-sided test in Table A.1.*

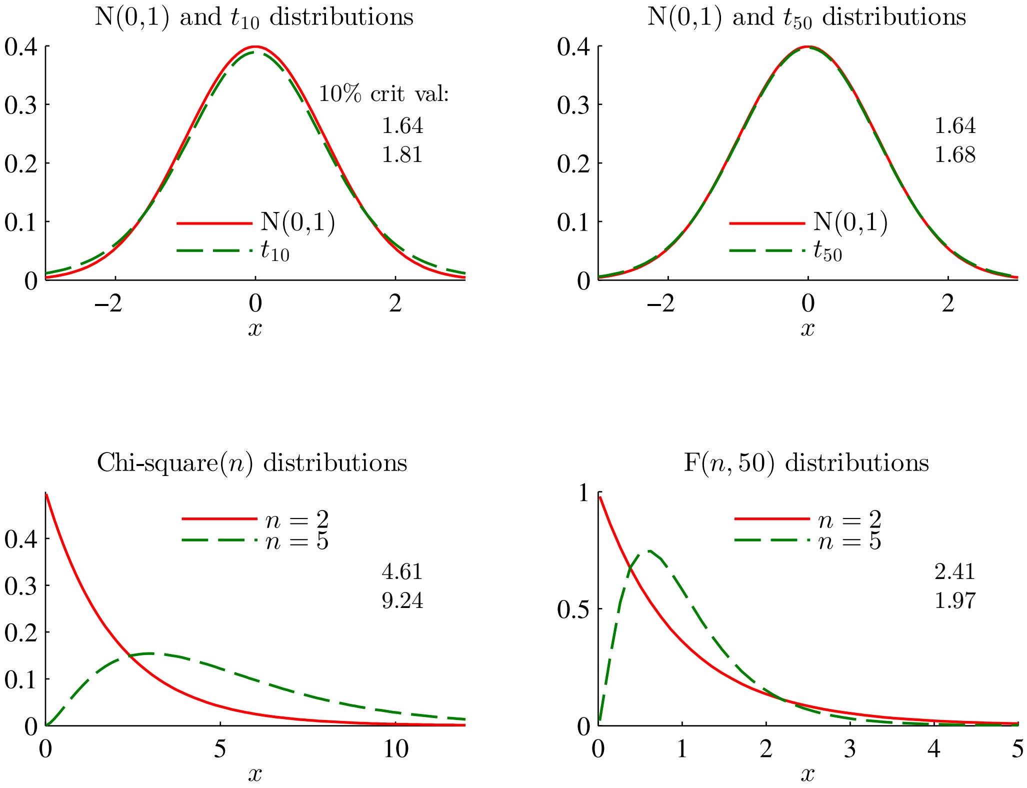


Figure 1.7: Probability density functions

1.3.3 Chi-square Distribution

If $z \sim N(0,1)$, then $z^2 \sim \chi^2_1$, 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

$$v'\Sigma^{-1}v \sim \chi^2_n, \text{ if the } n \times 1 \text{ vector } v \sim N(0, \Sigma). \tag{1.27}$$

See Figure 1.7 for an illustration and Table A.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 $\sim \chi^2_2$ *distribution. Notice that 6.1 is higher than the 5% critical value (but not the 1% critical value) in Table A.2.*

1.3.4 F-distribution

If we instead need to estimate Σ in (1.27) and let n_1 be the number of elements in v (previously called just n), then

$$v'\hat{\Sigma}^{-1}v/n_1 \sim F_{n_1,n_2} \tag{1.28}$$

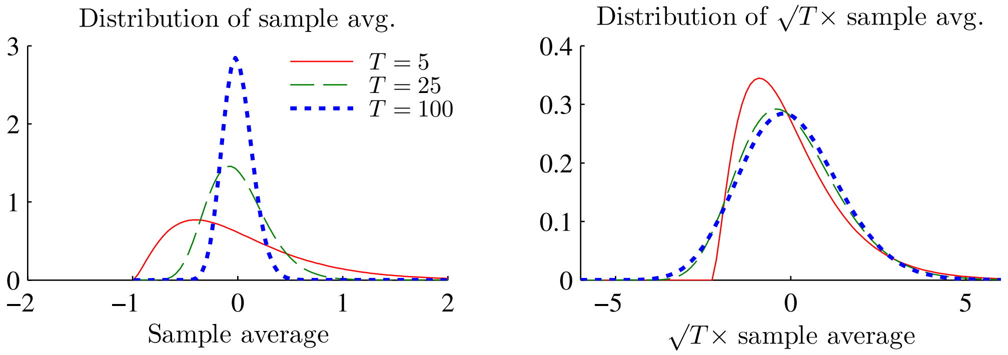
where F_{n_1,n_2} denotes an F-distribution with (n_1, n_2) degrees of freedom. Similar to the t-distribution, n_2 is the number of observations minus the number of estimated parameters. See Figure 1.7 for an illustration and Tables A.3–A.4 for critical values.

1.4 Normal Distribution of the Sample Mean as an Approximation

In many cases, it is unreasonable to just assume that the variable is normally distributed. The nice thing with a sample mean (or sample average) is that it will still be normally distributed—at least approximately (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 (1.16)) and that the variance goes to zero as the sample size goes to infinity (as in (1.15)). (This is also called convergence in mean square.) To see the LLN in action, see Figure 1.8.

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 1.8 for an illustration. This also holds for a large class of random variables—and it is a very useful result since it allows



Sample average of $z_t - 1$ where z_t has a χ^2_1 distribution

Figure 1.8: Sampling distributions

us to test hypothesis. Most estimators (including least squares and other methods) are effectively some kind of sample average, so the CLT can be applied.

A Statistical Tables

n	Critical values		
	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 A.1: Critical values (two-sided test) of t distribution (different degrees of freedom) and normal distribution.

n	Critical values		
	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 A.2: Critical values of chisquare distribution (different degrees of freedom, n).

$n1$			$n2$			$\chi^2_{n1}/n1$
	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 A.3: 5% Critical values of $F_{n1,n2}$ distribution (different degrees of freedom).

$n1$			$n2$			$\chi^2_{n1}/n1$
	10	30	50	100	300	
1	3.29	2.88	2.81	2.76	2.72	2.71
2	2.92	2.49	2.41	2.36	2.32	2.30
3	2.73	2.28	2.20	2.14	2.10	2.08
4	2.61	2.14	2.06	2.00	1.96	1.94
5	2.52	2.05	1.97	1.91	1.87	1.85
6	2.46	1.98	1.90	1.83	1.79	1.77
7	2.41	1.93	1.84	1.78	1.74	1.72
8	2.38	1.88	1.80	1.73	1.69	1.67
9	2.35	1.85	1.76	1.69	1.65	1.63
10	2.32	1.82	1.73	1.66	1.62	1.60

Table A.4: 10% Critical values of $F_{n1,n2}$ distribution (different degrees of freedom).

2 Least Squares Estimation

Reference: Verbeek (2008) 2 and 4
More advanced material is denoted by a star (*). It is not required reading.

2.1 Least Squares

2.1.1 Simple Regression: Constant and One Regressor

The simplest regression model is

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

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).

Note 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. If the regressor summarizes all the useful information we have in order to describe y_t, then the assumptions imply that we have no way of making a more intelligent guess of u_t (even after having observed x_t) than that it will be zero.

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

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

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 Figure 2.1 for an illustration.

Remark 2.1 *(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*

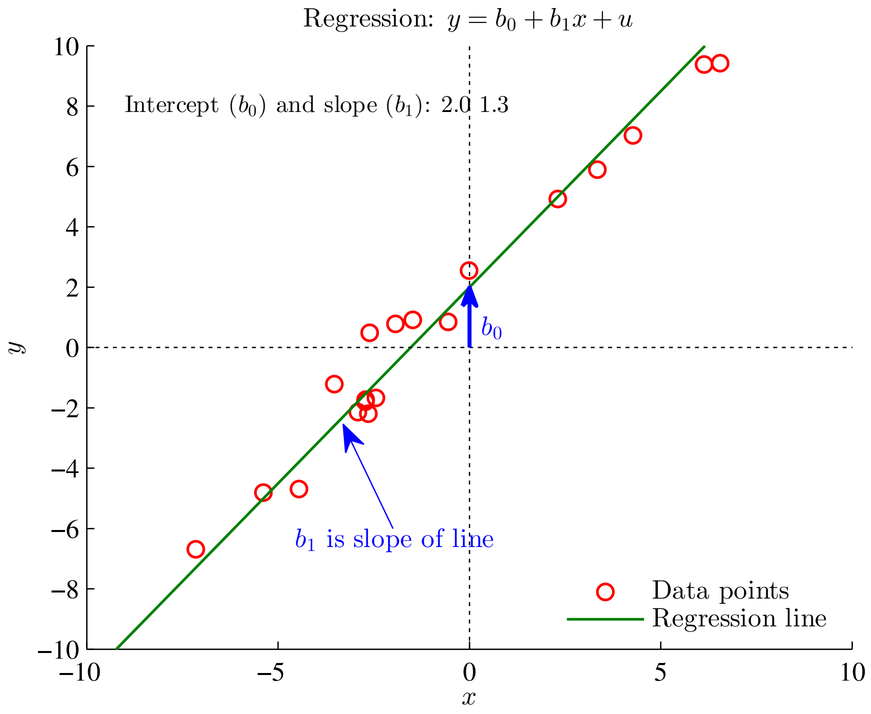


Figure 2.1: Example of OLS

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 2.2.

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

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

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

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

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

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

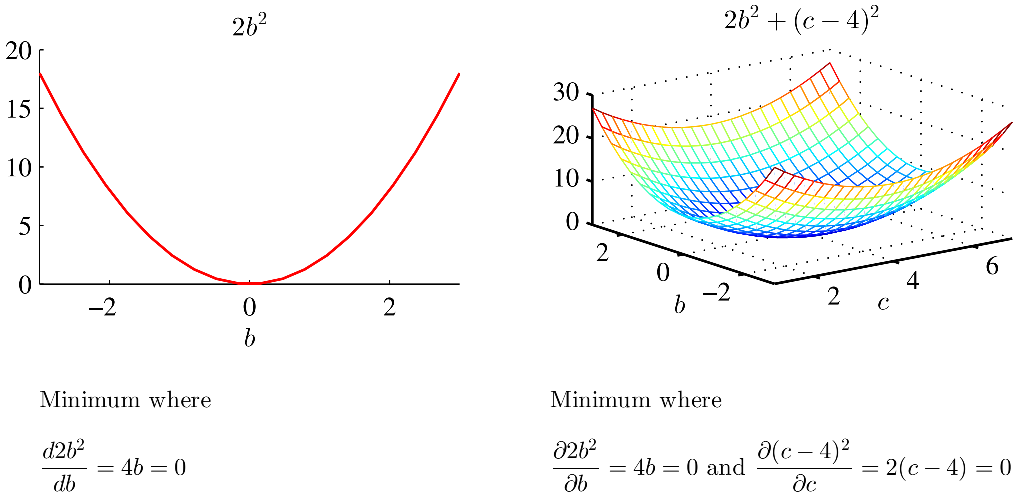


Figure 2.2: Quadratic loss function. Subfigure a: 1 coefficient; Subfigure b: 2 coefficients

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 $\mathrm{E}\, u_t = 0$ and $\mathrm{Cov}(x_t, u_t) = 0$. 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 should also be zero, which implies that the sample covariance is zero as well since $\hat{u}_t$ has a zero sample mean.

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.*

Remark 2.3 *Least squares is only one of many possible ways to estimate regression coefficients. We will discuss other methods later on.*

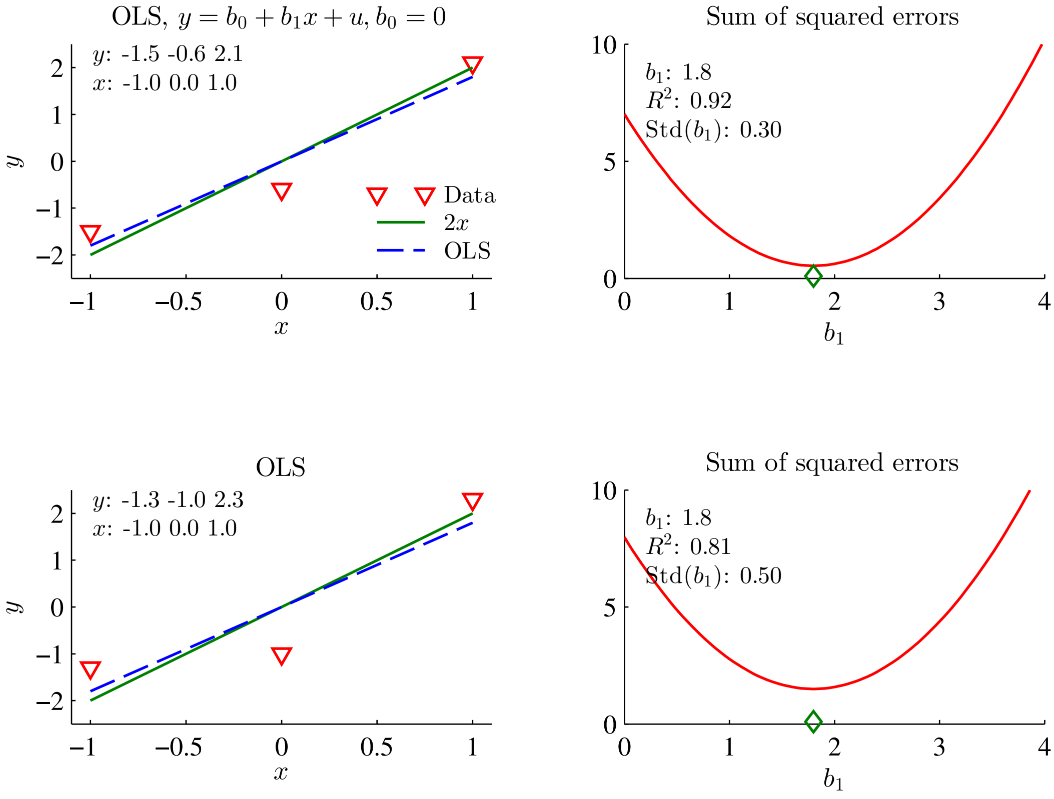


Figure 2.3: Example of OLS estimation

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

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

When $x = y$, then we get $\operatorname{Var}(x) = \mathrm{E}\,x^2 - (\mathrm{E}\,x)^2$. *These results hold for sample moments too.*

When the means of y and x are zero, then we can disregard the constant. In this case, (2.6) with $\hat{\beta}_0 = 0$ immediately gives

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

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).

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

$$\left[\begin{array}{ccc} y_1 & y_2 & y_3 \end{array}\right] = \left[\begin{array}{ccc} -1.5 & -0.6 & 2.1 \end{array}\right] \textit{and} \left[\begin{array}{ccc} x_1 & x_2 & x_3 \end{array}\right] = \left[\begin{array}{ccc} -1 & 0 & 1 \end{array}\right]$$

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

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

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

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

See Figure 2.3 for an illustration.

See Table 2.1 and Figure 2.4 for illustrations.

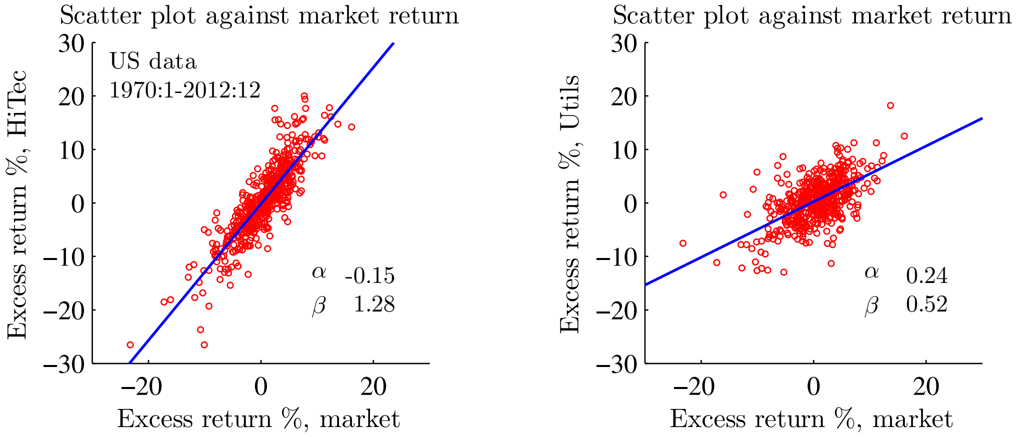


Figure 2.4: Scatter plot against market return

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

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

where y_t and u_t are scalars, x_t a $k\times1$ vector, and β is a $k\times1$ vector of the true coefficients (see Appendix A for a summary of matrix algebra). Least squares minimizes the sum of the squared fitted residuals

$$\textstyle\sum_{t=1}^{T}\hat{u}_t^2 = \sum_{t=1}^{T}(y_t - x_t'\hat{\beta})^2, \tag{2.9}$$

by choosing the vector β. The first order conditions are

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

which can be solved as

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

	HiTec	Utils
constant	−0.15	0.24
	(−1.00)	(1.58)
market return	1.28	0.52
	(33.58)	(12.77)
R2	0.75	0.34
obs	516.00	516.00
Autocorr (t)	−0.73	0.86
White	6.19	20.42
All slopes	386.67	176.89

Table 2.1: CAPM regressions, monthly returns, %, US data 1970:1-2012:12. Numbers in parentheses are t-stats. Autocorr is a N(0,1) test statistic (autocorrelation); White is a chi-square test statistic (heteroskedasticity), df = K(K+1)/2 - 1; All slopes is a chi-square test statistic (of all slope coeffs), df = K-1

Example 2.6 *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 2.7 *(Regression with an intercept and slope) Suppose we have the following data:*

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

This is clearly the same as in Example 2.5, except that we allow for an intercept—which turns out to be zero. The notation we need to solve this problem is the same as for a

	HiTec	Utils
constant	0.12	0.03
	(0.95)	(0.19)
market return	1.11	0.65
	(31.08)	(16.67)
SMB	0.23	−0.19
	(4.37)	(−3.61)
HML	−0.58	0.45
	(−9.74)	(7.01)
R2	0.83	0.47
obs	516.00	516.00
Autocorr (t)	0.47	1.18
White	70.79	49.06
All slopes	425.95	242.74

Table 2.2: Fama-French regressions, monthly returns, %, US data 1970:1-2012:12. Numbers in parentheses are t-stats. Autocorr is a N(0,1) test statistic (autocorrelation); White is a chi-square test statistic (heteroskedasticity), df = K(K+1)/2 - 1; All slopes is a chi-square test statistic (of all slope coeffs), df = K-1

general multiple regression. Therefore, calculate the following:

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

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

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

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

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

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

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

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

Notice that we must have constant in regression for R^2 to make sense—unless all variables have zero means.

Example 2.8 *(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 2.3.

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

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

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

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

■

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 noting 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 2.5 for an example.

Remark 2.9 *(R^2 from simple regression) Suppose $\hat{y}_t = \beta_0 + \beta_1 x_t$, so (2.13) becomes*

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

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.

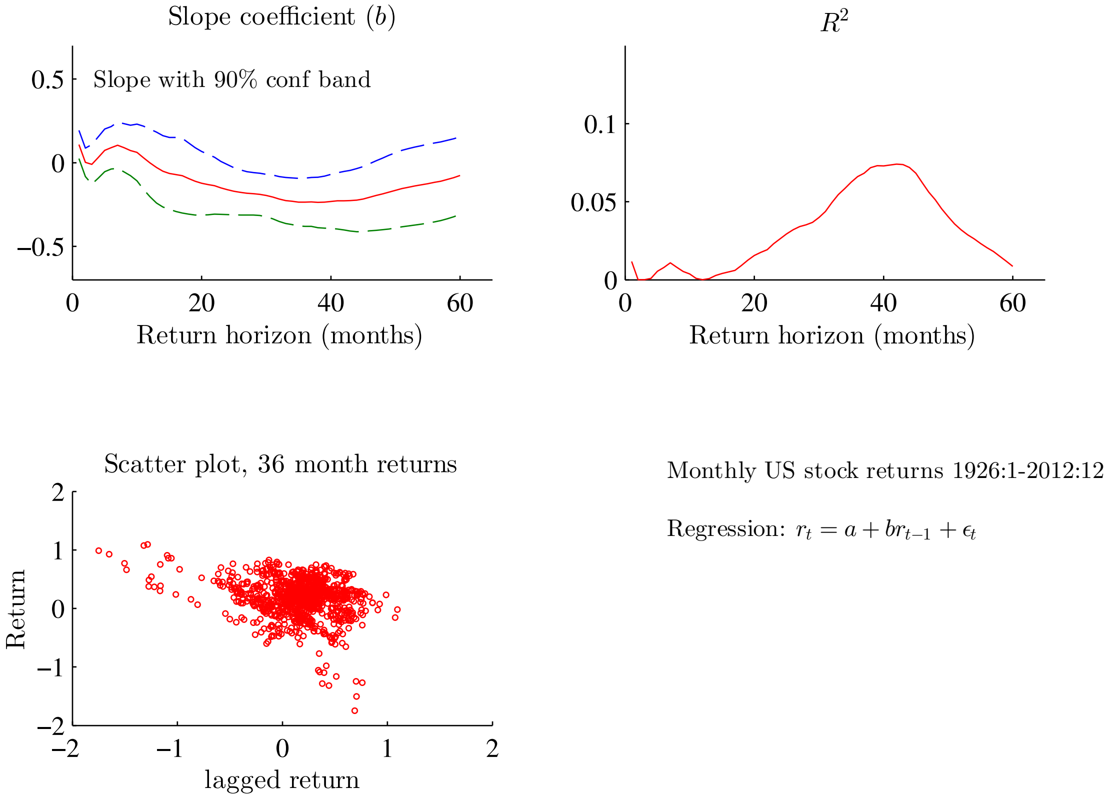


Figure 2.5: Prediction equations for US stock returns

2.1.4 Least Squares: Outliers*

Since the loss function in (2.2) is quadratic, a few outliers can easily have a very large influence on the estimated coefficients. For instance, suppose the true model is $y_t =$

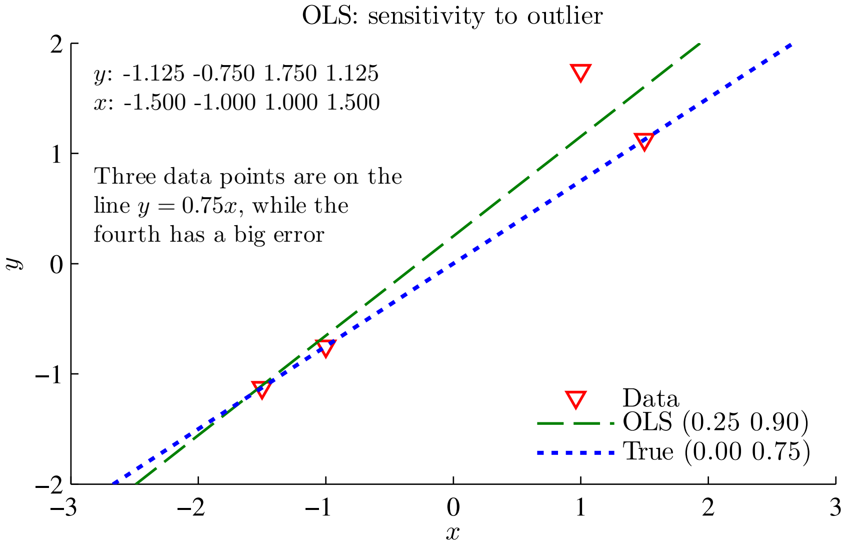


Figure 2.6: Data and regression line from OLS

$0.75x_t + u_t$, and that the residual is very large for some time period s. If the regression coefficient happened to be 0.75 (the true value, actually), the loss function value would be large due to the u_t^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).

There is of course nothing sacred about the quadratic loss function. Instead of (2.2) one could, for instance, use a loss function in terms of the absolute value of the error $\Sigma_{t=1}^{T} |y_t - \beta_0 - \beta_1 x_t|$. This would produce the Least Absolute Deviation (LAD) estimator. It is typically less sensitive to outliers. This is illustrated in Figure 2.7. 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.)

2.1.5 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

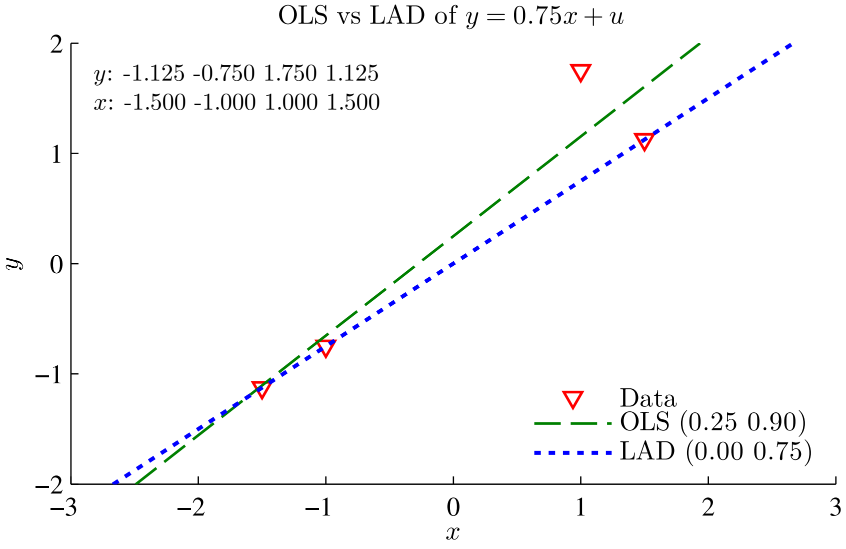


Figure 2.7: Data and regression line from OLS and LAD

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 in (2.7). Use (2.1) to substitute for y_t (recall $\beta_0 = 0$)

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

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. 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. Since u_t is a random variable, $\hat{\beta}_1$ is too. 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 u_t, x_t, and y_t. Since the true value, β, is a fixed constant, this distribution describes the uncertainty we should have about the true value after having obtained a specific estimated value.

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 simulation); *(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 Figure 2.11 for an example.

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

$$\hat{\beta} \sim N\left[\beta, \left(\textstyle\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sigma^2\right], \tag{2.15}$$

which allows for x_t to be a vector with k elements. Clearly, with $k = 1$, $x_t' = x_t$. See Figure 2.8 for an example.

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

$$\sqrt{T}(\hat{\beta} - \beta) \sim N\left[0, \left(\textstyle\sum_{t=1}^{T} x_t x_t'/T\right)^{-1} \sigma^2\right]. \tag{2.16}$$

This is the same as (2.15). (To see that, consider dividing the LHS of (2.16) by $\sqrt{T}$. Then, the variance on the RHS must be divided by T, which gives the same variance as in (2.15). Then, add β to the LHS, which changes the mean on the RHS to β. We then have (2.15).)

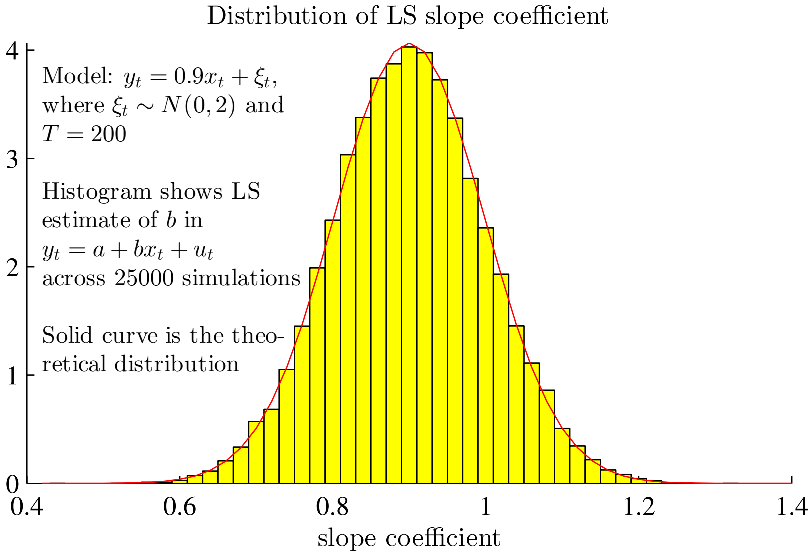


Figure 2.8: Distribution of OLS estimate, from simulation and theory

Example 2.10 *(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 2.11 *(Covariance matrix of b_1 and b_2) From Example 2.7*

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

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

The standard deviations (also called standard errors) are therefore

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

2.1.6 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*: the residuals have zero means, have constant variances and 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 is 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).

Write (2.14) as

$$\hat{\beta}_1 = \beta_1 + \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(x_1 u_1 + x_2 u_2 + \dots x_T u_T \right). \tag{2.17}$$

Since x_t are assumed to be constants (not random), the expected value of this expression is

$$\operatorname{E} \hat{\beta}_1 = \beta_1 + \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(x_1 \operatorname{E} u_1 + x_2 \operatorname{E} u_2 + \dots x_T \operatorname{E} u_T \right) = \beta_1 \tag{2.18}$$

since we always assume that the residuals have zero means (see (2.1)). The interpretation is that we can expected 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_1). Clearly, this is something we want from an estimation method (a method that was systematically wrong would not be very attractive).

Remark 2.12 *(Linear combination of normally distributed variables.) If the random variables z_t and v_t are normally distributed, then $a + bz_t + cv_t$ is too. To be precise, $a + bz_t + cv_t \sim N\left(a + b\mu_z + c\mu_v, b^2\sigma_z^2 + c^2\sigma_v^2 + 2bc\sigma_{zv}\right)$.*

Suppose $u_t \sim N\left(0, \sigma^2\right)$, then (2.17) shows that $\hat{\beta}_1$ is normally distributed. The reason is that $\hat{\beta}_1$ is just a constant (β_1) plus a linear combination of normally distributed residuals (with fixed regressors $x_t / \sum_{t=1}^{T} x_t x_t$ can be treated as constant). It is straightforward to see that the mean of this normal distribution is β_1 (the true value), since the rest is a linear combination of the residuals—and they all have a zero mean. Finding the variance of $\hat{\beta}_1$ 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 (independently distributed) and have the same variances (identically distributed). The variance of (2.17) is then

$$\begin{aligned}
\operatorname{Var}(\hat{\beta}_1) &= \frac{1}{\sum_{t=1}^{T} x_t x_t} \operatorname{Var}\left(x_1 u_1 + x_2 u_2 + \ldots x_T u_t\right) \frac{1}{\sum_{t=1}^{T} x_t x_t} \\
&= \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(x_1^2 \sigma_1^2 + x_2^2 \sigma_2^2 + \ldots x_T^2 \sigma_T^2\right) \frac{1}{\sum_{t=1}^{T} x_t x_t} \\
&= \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(x_1^2 \sigma^2 + x_2^2 \sigma^2 + \ldots x_T^2 \sigma^2\right) \frac{1}{\sum_{t=1}^{T} x_t x_t} \\
&= \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(\sum_{t=1}^{T} x_t x_t\right) \sigma^2 \frac{1}{\sum_{t=1}^{T} x_t x_t} \\
&= \frac{1}{\sum_{t=1}^{T} x_t x_t} \sigma^2. \qquad (2.19)
\end{aligned}$$

The first line follows directly from (2.17), since β_1 is a constant. Notice that the two $\sum_{t=1}^{T} x_t x_t$ terms are kept separate in order to facilitate the comparison with the case of several regressors. The second line follows from assuming that the residuals are uncorrelated with each other ($\operatorname{Cov}(u_i, u_j) = 0$ if $i \neq j$), so all cross terms ($x_i x_j \operatorname{Cov}(u_i, u_j)$) 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.

Notice that the denominator increases with the sample size while the numerator stays constant: a larger sample gives a smaller uncertainty about the estimate. Similarly, a lower volatility of the residuals (lower σ^2) also gives a lower uncertainty about the estimate. See Figure 2.9.

Example 2.13 *When the regressor is just a constant (equal to one) $x_t = 1$, then we have*

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

(This is the classical expression for the variance of a sample mean.)

Example 2.14 *When the regressor is a zero mean variable, then we have*

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

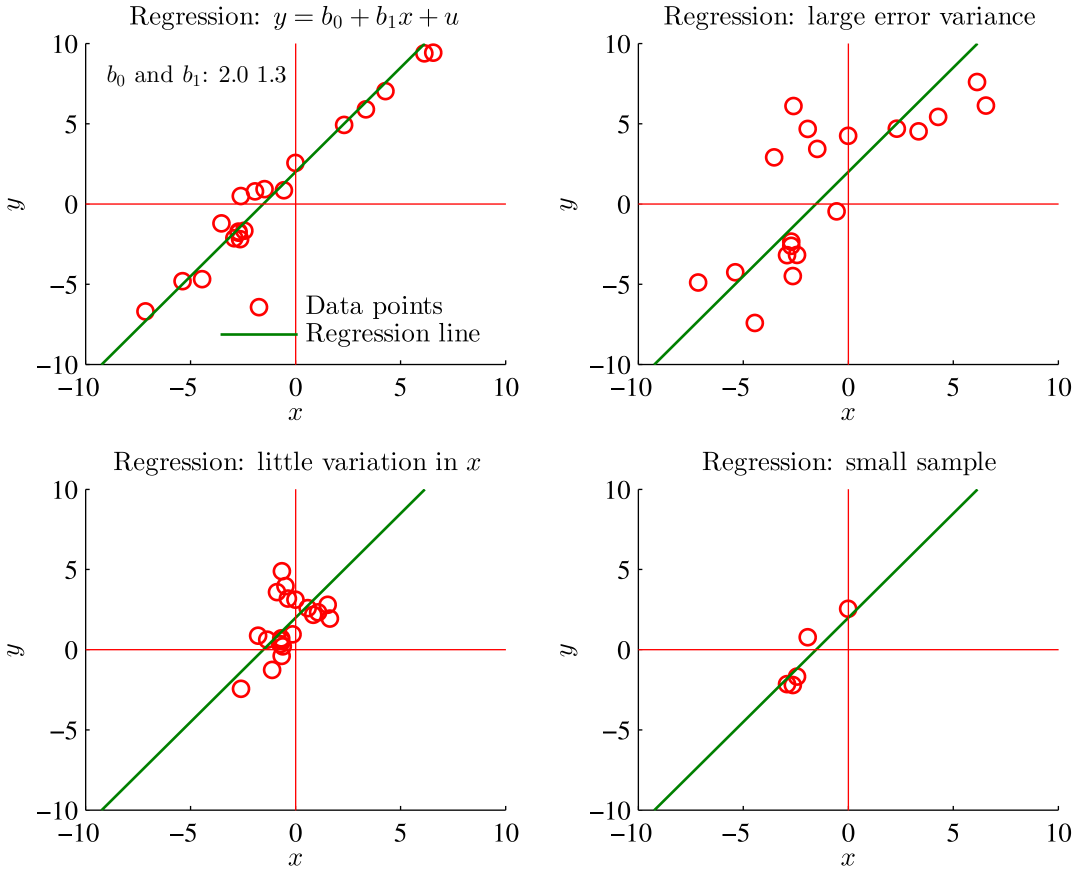


Figure 2.9: Regressions: importance of error variance and variation of regressor

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

Example 2.15 *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*

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

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

A combination of the two previous examples.

2.1.7 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.

As a simple example, consider the regression

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t, \tag{2.20}$$

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

$$\text{Var}(\hat{\beta}_2) = \frac{1}{1 - \text{Corr}(x_{1t}, x_{2t})^2} \frac{1}{\text{Var}(x_{2t})} \frac{\sigma^2}{T}, \tag{2.21}$$

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 variables have an effect on y_t, but it is hard to tell if that effect is from regressor one or two.

Proof. (of 2.21). 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 (2.20) we get

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

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

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

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

which is the same as (2.21). ■

2.1.8 The Distribution of $\hat{\beta}$: When the Assumptions Are Wrong

The results on the distribution $\hat{\beta}$ 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 assumptions—and unlikely to be relevant for financial data. This forces us to rely on asymptotic ("large sample") theory (or do simulations). The main results from asymptotic theory (see below) is that the main result (2.15) is a good approximation in large samples, provided the Gauss-Markov assumptions are correct (if not, see later sections on heteroskedasticity and autocorrelation). However, things are more complicated in small samples. Only simulations can help us there.

Example 2.16 *(When OLS is biased, at least in a finite sample) OLS on an AR(1), $y_t = \beta_1 + \beta_2 y_{t-1} + u_t$, is not unbiased. In this case, the regressors are not fixed and u_t is not independent of the regressor for all periods: u_t is correlated with y_t (obviously)—which is the regressor in $t+1$.* See Figure 2.10.

2.1.9 The Distribution of $\hat{\beta}$: A Bit of Asymptotic Theory*

A law of large numbers would (in most cases) say that both $\sum_{t=1}^{T} x_t^2/T$ and $\sum_{t=1}^{T} x_t u_t/T$ in (2.14) converges to their expected values as $T \to \infty$. The reason is that both are sample

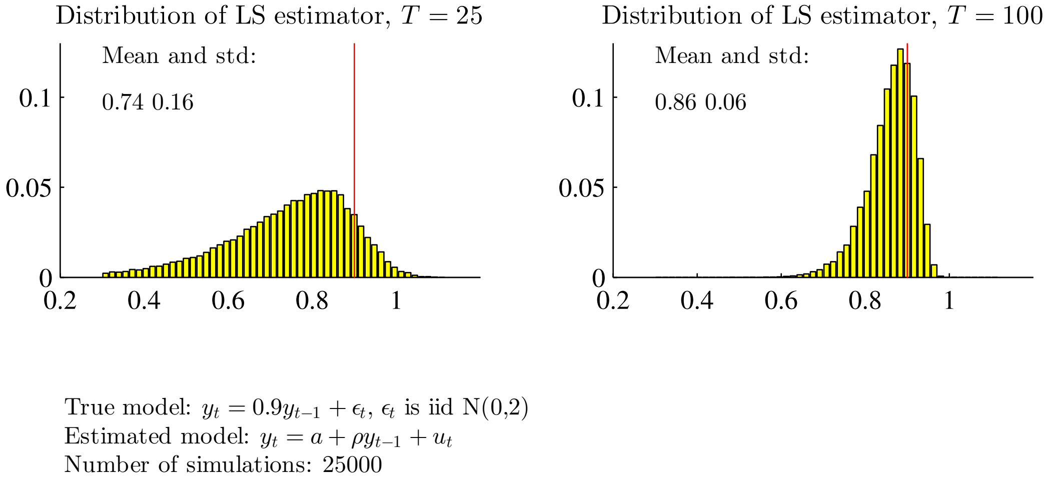


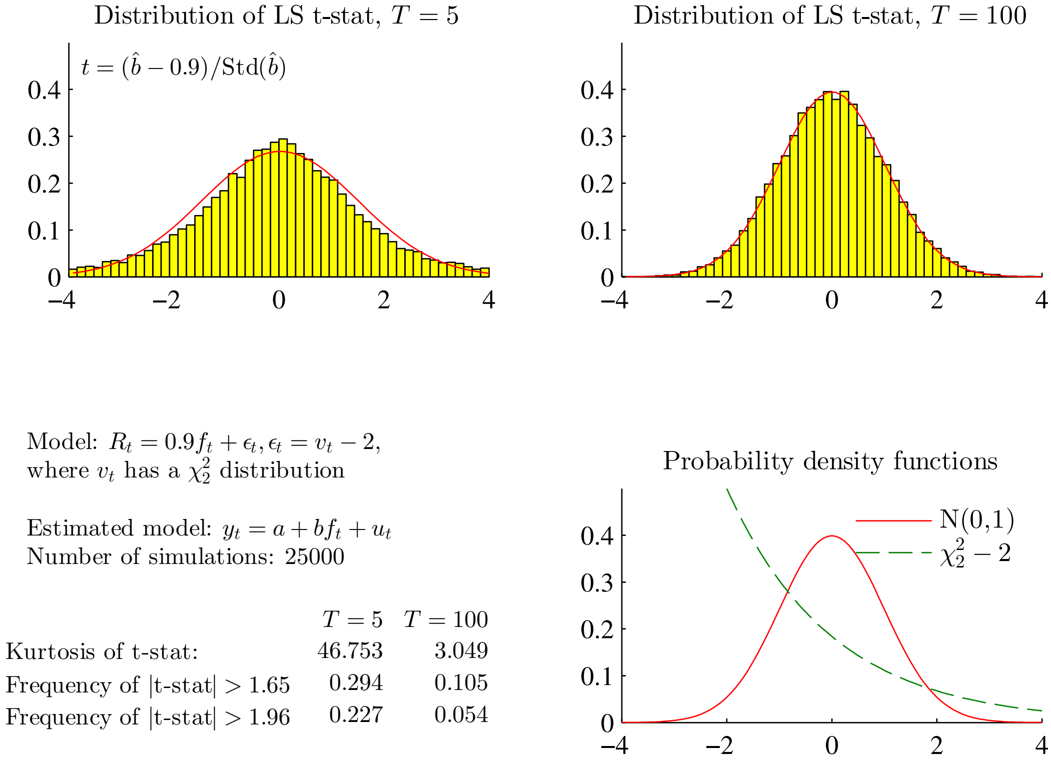
Figure 2.10: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

averages of random variables (clearly, both x_t^2 and $x_t u_t$ are random variables). These expected values are Var (x_t) and Cov (x_t, u_t), respectively (recall both x_t and u_t have zero means). The key to show that $\hat{\beta}$ is *consistent* is that Cov $(x_t, u_t) = 0$. This highlights the importance of using good theory to derive not only the systematic part of (2.1), but also in understanding the properties of the errors. For instance, when economic theory tells us that y_t and x_t affect each other (as prices and quantities typically do), then the errors are likely to be correlated with the regressors—and LS is inconsistent. One common way to get around that is to use an instrumental variables technique. Consistency is a feature we want from most estimators, since it says that we would at least get it right if we had enough data.

Suppose that $\hat{\beta}$ is consistent. Can we say anything more about the asymptotic distribution. Well, the distribution of $\hat{\beta}$ converges to a spike with all the mass at β, but the distribution of $\sqrt{T}(\hat{\beta} - \beta)$, will typically converge to a non-trivial normal distribution. To see why, note from (2.14) that we can write

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\sum_{t=1}^{T} x_t^2 / T\right)^{-1} \sqrt{T} \sum_{t=1}^{T} x_t u_t / T. \tag{2.22}$$

The first term on the right hand side will typically converge to the inverse of Var (x_t), as discussed earlier. The second term is $\sqrt{T}$ times a sample average (of the random variable



	$T = 5$	$T = 100$
Kurtosis of t-stat:	46.753	3.049
Frequency of $\lvert$t-stat$\rvert > 1.65$	0.294	0.105
Frequency of $\lvert$t-stat$\rvert > 1.96$	0.227	0.054

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

$x_t u_t$) with a zero expected value, since we assumed that $\hat{\beta}$ is consistent. Under weak conditions, a central limit theorem applies so $\sqrt{T}$ times a sample average converges to a normal distribution. This shows that $\sqrt{T}\hat{\beta}$ has an *asymptotic normal distribution*. It turns out that this is a property of many estimators, 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.

2.2 Hypothesis Testing

2.2.1 Testing a Single Coefficient

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

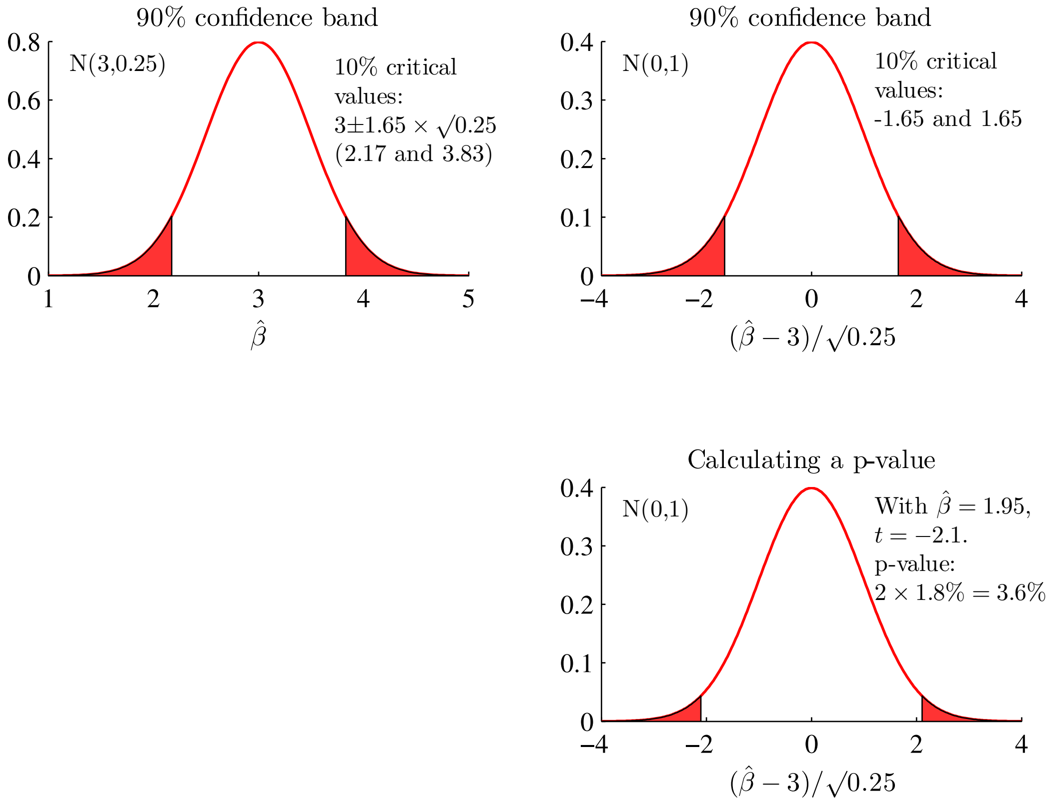


Figure 2.12: Confidence band around the null hypothesis

We assume that the estimates are normally distributed. 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 $\mathrm{E}\,\hat{\beta}$ equals the true value) and by further dividing by the standard error (deviation) of $\hat{\beta}$, we should have

$$t=\frac{\hat{\beta}-q}{\mathrm{Std}(\hat{\beta})}\sim N(0,1) \tag{2.23}$$

In case $|t|$ is very large (say, 1.65 or larger), then $\hat{\beta}$ is a very unlikely outcome if $\mathrm{E}\,\hat{\beta}$ (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.

The logic this hypothesis test is perhaps best described by a 90% confidence band around the null hypothesis

$$\begin{aligned}\Pr(\hat{\beta}\text{ is in ConfBand}) &= 90\%,\text{ where} \\ \text{ConfBand} &= q \pm 1.65\,\text{Std}(\hat{\beta}).\end{aligned} \tag{2.24}$$

See Figure 2.12 for an illustration. The idea is that if the true value of the coefficient is q, then the estimate $\hat{\beta}$ should be inside this band in 90% of the cases (that is, different samples). Hence, a $\hat{\beta}$ outside the band is unlikely to happen if the true coefficient is q: we will interpret that situation as if the true value is not q.

If the point estimate is outside this confidence band band, then this is the same as rejecting the null hypothesis. To see that, notice that for $\hat{\beta}$ to be outside the band we must have

$$|t| > 1.65, \tag{2.25}$$

See Figure 2.12 for an illustration.

Proof. (of (2.25)) For $\hat{\beta}$ to be outside the band we must have

$$\hat{\beta} < q - 1.65\,\text{Std}(\hat{\beta}) \text{ or } \hat{\beta} > q + 1.65\,\text{Std}(\hat{\beta}).$$

Rearrange this by subtracting q from both sides of the inequalities and then divide both sides by $\text{Std}(\hat{\beta})$

$$\frac{\hat{\beta}-q}{\text{Std}(\hat{\beta})} < -1.65 \text{ or } \frac{\hat{\beta}-q}{\text{Std}(\hat{\beta})} > 1.65.$$

This is the same as (2.25). ■

Example 2.17 *(t-test) With* $\text{Std}(\hat{\beta}) = \sqrt{0.25}$ *and* $q = 3$*, the 90% confidence band is* $3 \pm 1.65 \times \sqrt{0.25}$*, 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* $[-1.65, 1.65]$.

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 hypotheis), then you may change from the *critical value* 1.65 to 1.96. This gives a 95% confidence band, so the significance level is 5%. See Figure 2.13 for an illustration and Tables 2.1 and Table 2.2 for examples.

The *p-value* is a related concept. It is the lowest significance level at which we can reject the null hypothesis. See Figure 2.12 for an illustration.

Example 2.18 *(p-value) With* $\operatorname{Std}(\hat{\beta}) = \sqrt{0.25}$*,* $\hat{\beta} = 1.95$ *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 1.65 for the standard normal distribution). However, for samples of more than 30–40 data points, the difference is trivial—see Table A.1.

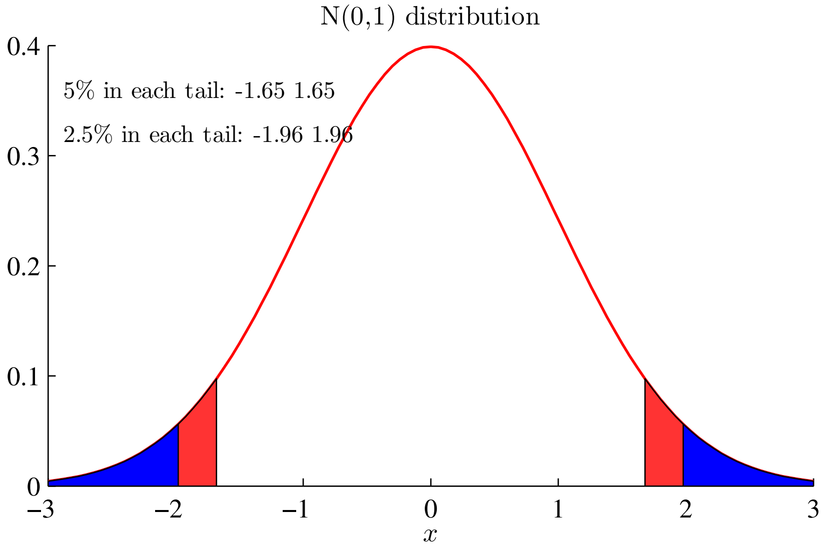


Figure 2.13: Density function of a standard normal distribution

2.2.2 Confidence Band around the Point Estimate

Andther way to construct a confidence band is to center the band on the point estimate

$$\hat{\beta} \pm 1.65 \operatorname{Std}(\hat{\beta}). \tag{2.26}$$

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 1.65 to 1.96 to get a 95% confidence band.)

2.2.3 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 $(-1.65, 1.65)$, the size is 10%. (The size is sometime called the type I error.)

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 we 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 -1.65 and 1.65, the power would be

$$\text{power} = \Pr(t \leq -1.65) + \Pr(t \geq 1.65). \tag{2.27}$$

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

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

$$t = \frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} \sim N\left[\frac{\beta - q}{\text{Std}(\hat{\beta})}, 1\right]. \tag{2.28}$$

We can then calculate the power as the probability that $t \leq -1.65$ or $t \geq 1.65$, when t has the distribution on the RHS in (2.28).

Example 2.19 *If* $\beta = 1.6$*,* $q = 1$ *and* $\text{Std}(\hat{\beta}) = 1/3$*, then the power is 0.56. See Figure 2.14.*

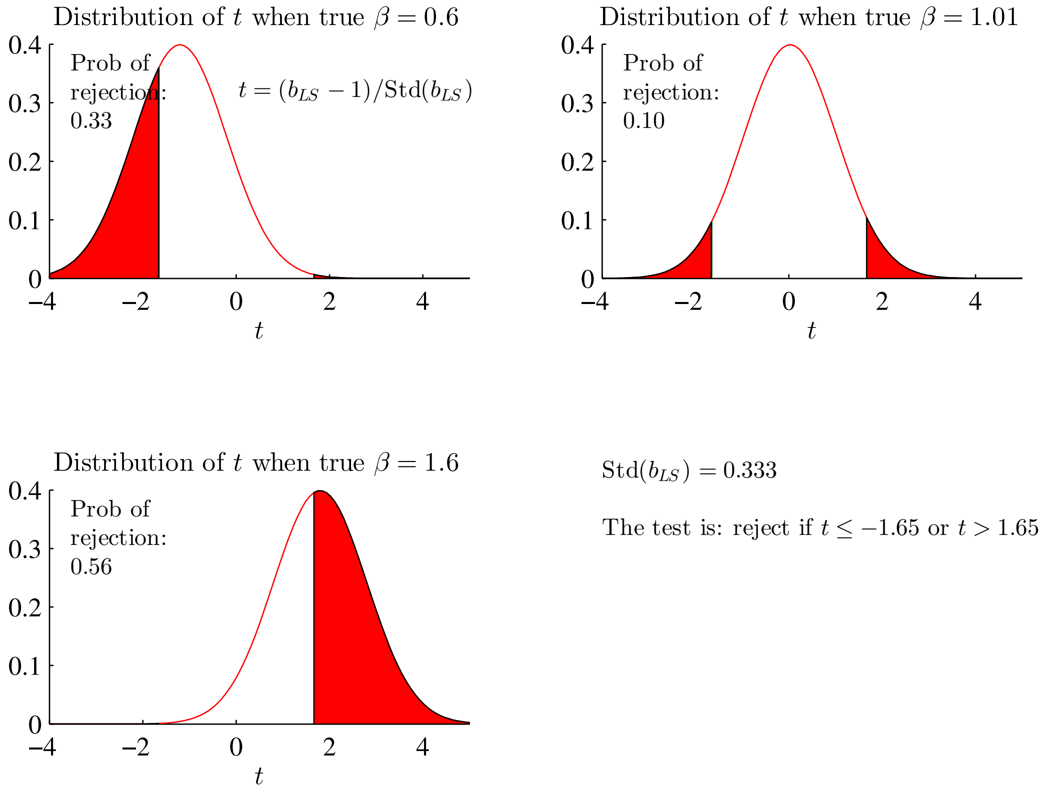


Figure 2.14: Power of t-test, assuming different true parameter values

2.2.4 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 exploiting the power of repeated sampling as illustrated by the following example. My null hypothesis might be that I am a better tennis player than my friend. After playing (and losing) once, I cannot reject the null—since pure randomness (wind, humidity,...) might have caused the result. The same is true for the second game (and loss)—if I treat the games as completely unrelated events. However, considering both games, the evidence against the null hypothesis is (unfortunately) much stronger.

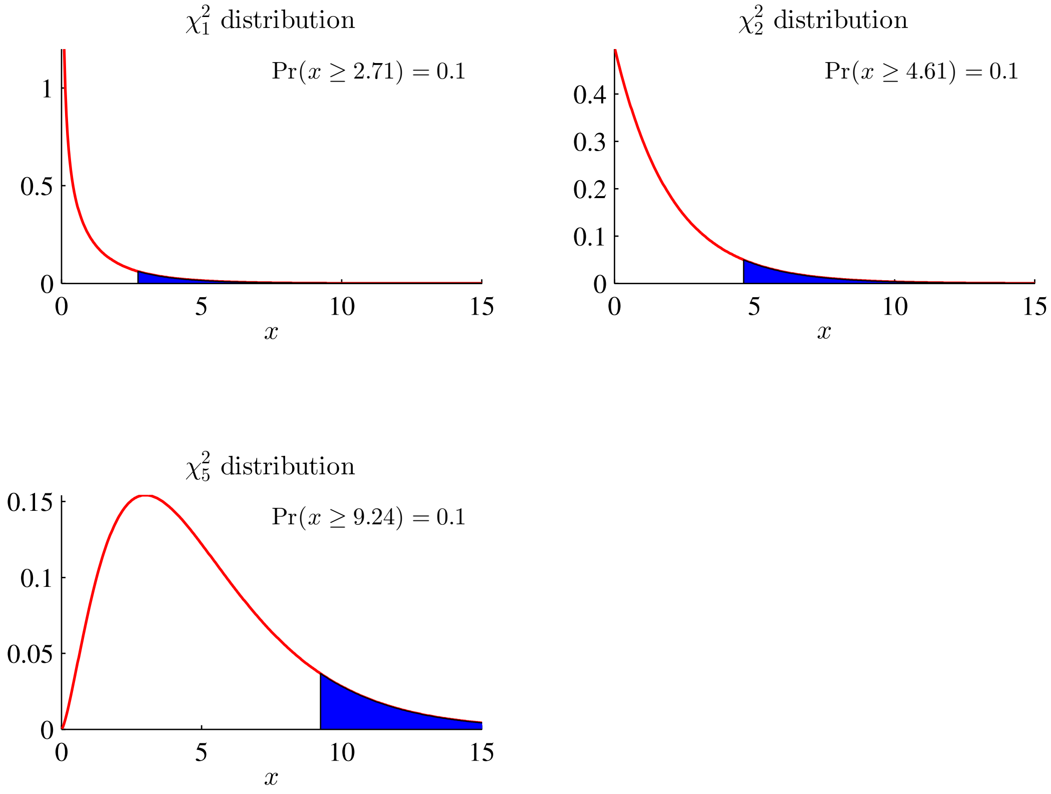


Figure 2.15: Density functions of χ^2 distributions with different degrees of freedom

A joint test makes use of the following remark.

Remark 2.20 *(Chi-square distribution) If v is a zero mean normally distributed vector, then we have*

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

As a special case, suppose the vector only has one element. Then, the quadratic form can be written $[v/\operatorname{Std}(v)]^2$, which is the square of a t-statistic.

Example 2.21 *(Quadratic form with a chi-square distribution) If the 2×1 vector v has the following normal distribution*

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

then the quadratic form

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

has a χ_2^2 distribution.

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

$$H_0 : \beta_1 = 1 \text{ and } \beta_3 = 0. \tag{2.29}$$

It is convenient to write this on matrix form as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ or more generally} \tag{2.30}$$

$$R\beta = q, \tag{2.31}$$

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

$$\text{Var}(R\hat{\beta}) = RV(\hat{\beta})R', \tag{2.32}$$

where $V(\hat{\beta})$ denotes the covariance matrix of the coefficients, for instance, from (2.19). Putting together these results we have the test static (a scalar)

$$(R\hat{\beta} - q)'[RV(\hat{\beta})R']^{-1}(R\hat{\beta} - q) \sim \chi_J^2. \tag{2.33}$$

This test statistic is compared to the critical values of a χ_J^2 distribution—see Table A.2. (Alternatively, it can 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)

$$TR^2/(1 - R^2) \sim \chi_{k-1}^2. \tag{2.34}$$

See Tables 2.1 and 2.2 for examples of this test.

Example 2.22 *(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*

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

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

Then, (2.33) 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).

Proof. (of (2.34)) 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 $V(\hat{\beta}) = [T\,\text{Var}(x_t)]^{-1}\sigma^2$ (see (2.15)), 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 (2.33) with $R = I$ and $q = \mathbf{0}$ and J equal to the number of slope coefficients. ■

2.2.5 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 (2.33) by J and replace $V(\hat{\beta})$ by the estimated covariance matrix $\hat{V}(\hat{\beta})$. This is, for instance, from (2.19) $\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}$, 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

$$\frac{\left(R\hat{\beta}-q\right)'\left[R\hat{V}(\hat{\beta})R'\right]^{-1}\left(R\hat{\beta}-q\right)}{J} \sim F_{J,T-k}, \text{ where} \quad (2.35)$$
$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum\nolimits_{i=1}^{N} x_i x_i'\right)^{-1}.$$

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)

$$\frac{R^2/(k-1)}{(1-R^2)/(T-k)} \sim F_{k-1,T-k}. \quad (2.36)$$

Proof. (of (2.35)) Equation (2.35) can also be written

$$\frac{\left(R\hat{\beta}-q\right)'\left[R\sigma^2\left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} R'\right]^{-1}\left(R\hat{\beta}-q\right)/J}{\hat{\sigma}^2/\sigma^2}.$$

The numerator is a χ^2_J variable divided by J. If the residuals are normally distributed, then it can be shown that the denominator is a χ^2_{T-k} variable divided by $T-k$. 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 2.23 *(Joint F test) Continuing Example 2.22, and assuming that $\hat{V}(\hat{\beta}) = V(\hat{\beta})$, 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.*

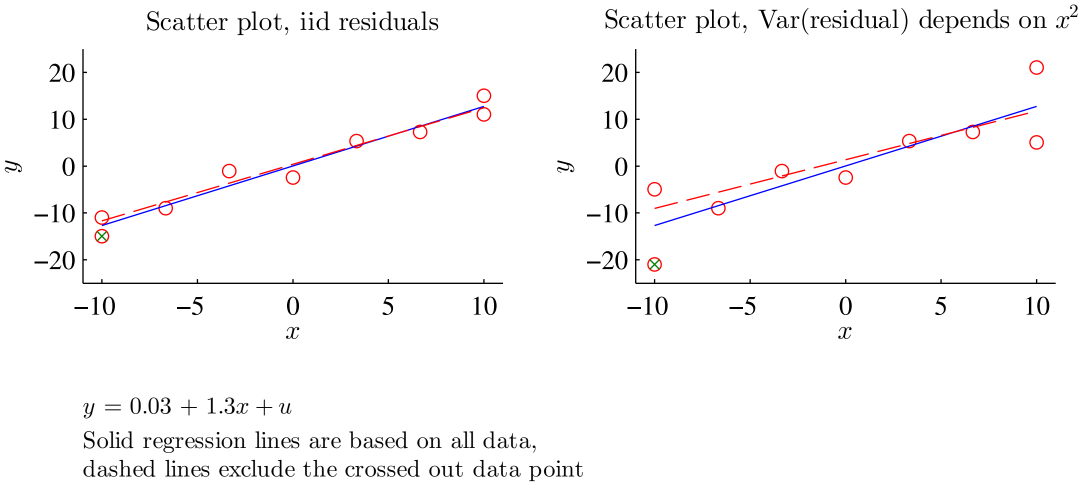


Figure 2.16: Effect of heteroskedasticity on uncertainty about regression line

2.3 Heteroskedasticity

Suppose we have a regression model

$$y_t = x_t' b + u_t, \text{ where } \mathrm{E}\, u_t = 0 \text{ and } \mathrm{Cov}(x_{it}, u_t) = 0. \tag{2.37}$$

In the standard case we assume that u_t is iid (independently and identically distributed), which rules out 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)—and it is reasonably efficient (in terms of the variance of the estimates). However, the standard expression for the standard errors (of the coefficients) is (except in a special case, see below) not correct. This is illustrated in Figure 2.17.

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 residual is uncorrelated with these squared regressors, then the usual LS covariance matrix applies—even if the residuals have some other sort of heteroskedasticity (this is the special case mentioned before).

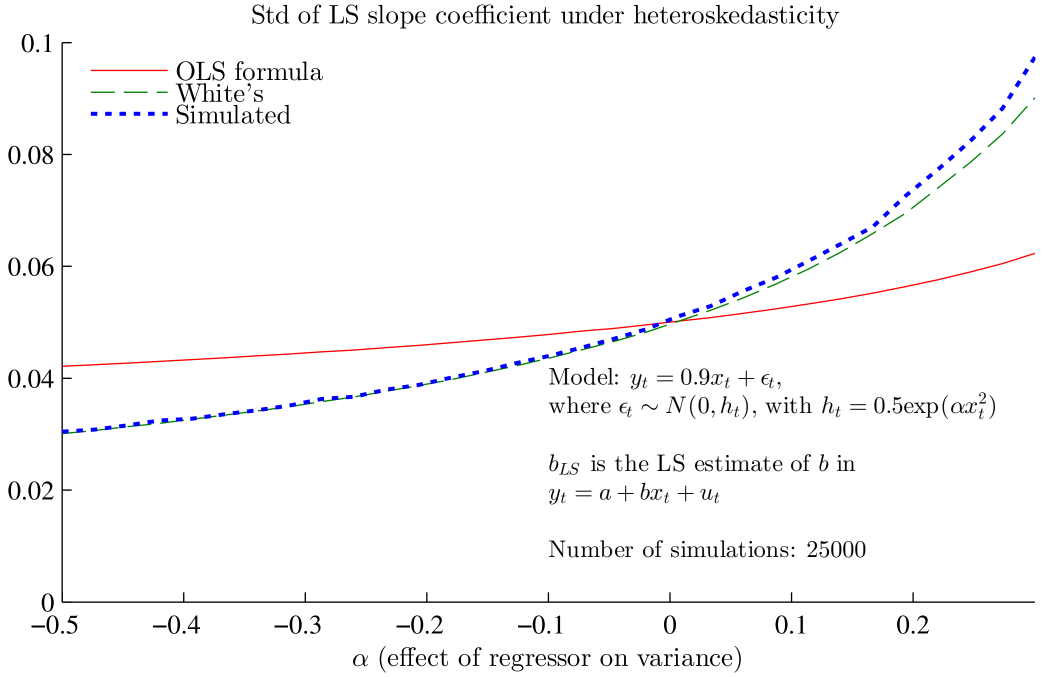


Figure 2.17: Variance of OLS estimator, heteroskedastic errors

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

$$\hat{u}_t^2 = w_t'\gamma + v_t, \tag{2.38}$$

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

Example 2.24 *(White's test) If the regressors include* $(1, x_{1t}, x_{2t})$ *then* w_t *in (2.38) is the vector* $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$*. (Notice that the cross product of* $(1, x_{1t}, x_{2t})$ *with* 1 *gives us the regressors in levels, not squares.)*

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 (2.37) with an ARCH structure of the residuals—and estimate the whole thing with maximum likelihood (MLE) is one way. 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.

To understand the construction of White's covariance matrix, recall that the variance of $\hat{\beta}_1$ is found from

$$\hat{\beta}_1 = \beta_1 + \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(x_1 u_1 + x_2 u_2 + \dots x_T u_T \right). \tag{2.39}$$

This gives

$$\begin{aligned}
\operatorname{Var}(\hat{\beta}_1) &= \frac{1}{\sum_{t=1}^{T} x_t x_t} \operatorname{Var}\left(x_1 u_1 + x_2 u_2 + \dots x_T u_t \right) \frac{1}{\sum_{t=1}^{T} x_t x_t} \\
&= \frac{1}{\sum_{t=1}^{T} x_t x_t} \left(x_1^2 \operatorname{Var}(u_1) + x_2^2 \operatorname{Var}(u_2) + \dots x_T^2 \operatorname{Var}(u_T) \right) \frac{1}{\sum_{t=1}^{T} x_t x_t} \\
&= \frac{1}{\sum_{t=1}^{T} x_t x_t} \sum_{t=1}^{T} x_t^2 \sigma_t^2 \frac{1}{\sum_{t=1}^{T} x_t x_t}, \qquad (2.40)
\end{aligned}$$

where the second line assumes that the residuals are uncorrelated. 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 $\sum_{t=1}^{T} x_t^2 \sigma_t^2$ by $\sum_{t=1}^{T} x_t x_t' \hat{u}_t^2$ (which 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).

Remark 2.25 *(Standard OLS vs White's variance) If x_t^2 is not related to σ_t^2, then we could write the last term in (2.40) as*

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

where $\overline{\sigma^2}$ 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. 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 $\frac{1}{T} \sum_{t=1}^{T} \sigma_t^2 \sum_{t=1}^{T} x_t^2 =$ $\frac{1}{2}(5 + 2)(10 + 1) = 38.5$.

2.4 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 $\hat{\beta}_1$. In this case, LS is (typically) still consistent (exceptions: when the lagged dependent variable is a regressor), but the variances are (again) wrong. In particular, not even the the first line of (2.40) is true, since the variance of the sum in (2.39) depends also on the covariance terms.

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—the estimate is imprecise. See Figure 2.18 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

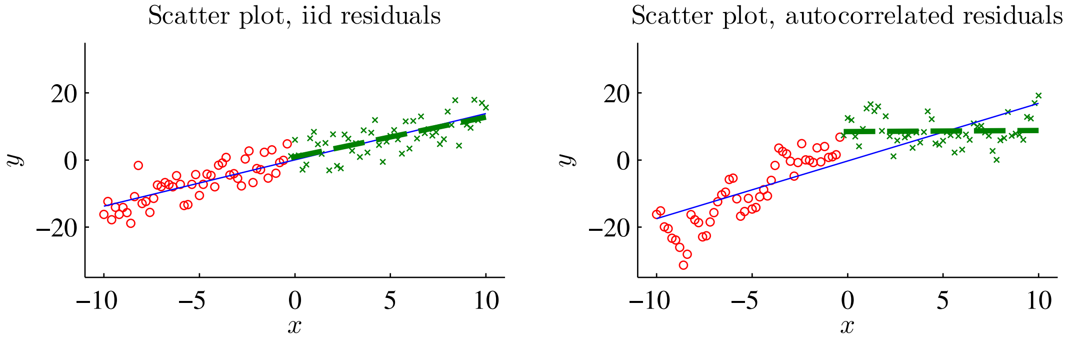
$$\rho_s = \text{Corr}(\hat{u}_t, \hat{u}_{t-s}), s = 1, ..., L. \tag{2.41}$$

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

$$\sqrt{T}\hat{\rho}_s \sim N(0, 1). \tag{2.42}$$

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

$$DW \approx 2 - 2\hat{\rho}_1, \tag{2.43}$$



$y = 0.03 + 1.3x + u$
Solid regression lines are based on all data,
dashed lines are based on late sample (high x values).
The regressor is (strongly) autocorrelated, since
it is an increasing series (-10,-9.9,...,10).

Figure 2.18: Effect of autocorrelation on uncertainty about regression line

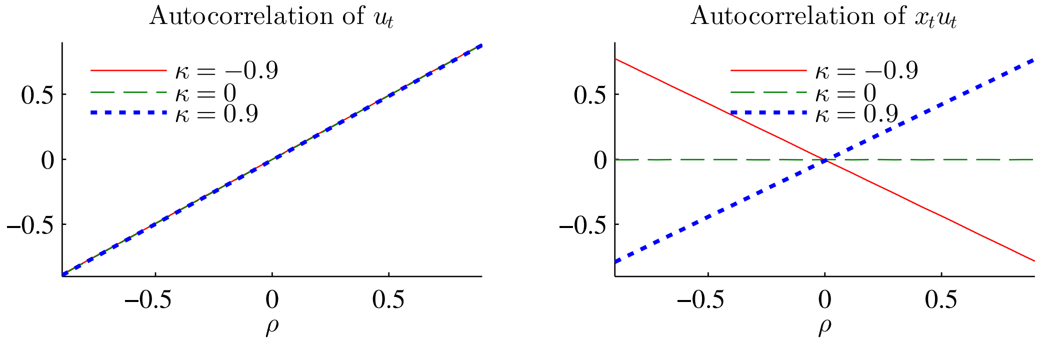
and the null hypothesis is rejected in favour of positive autocorrelation if DW<1.5 or so (depending on sample size and the number of regressors). To extend (2.42) to higher-order autocorrelation, use the Box-Pierce test

$$Q_L = T \sum_{s=1}^{L} \hat{\rho}_s^2 \to^d \chi_L^2. \tag{2.44}$$

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 first line of (2.40) is still correct. However, there might be corrrelation across time periods, so the second line needs to account for terms like $\text{Cov}(x_t u_t, x_{t-s} u_{t-s})$. For instance, for $T = 3$ the middle term of that second line is

$$\begin{aligned}\text{Var}\left(x_1 u_1 + x_2 u_2 + x_3 u_3\right) = {} & \text{Var}(x_1 u_1) + \text{Var}(x_2 u_2) + \text{Var}(x_3 u_3) + \\ & 2\,\text{Cov}(x_2 u_2, x_1 u_1) + 2\,\text{Cov}(x_3 u_3, x_2 u_2) + 2\,\text{Cov}(x_3 u_3, x_1 u_1).\end{aligned} \tag{2.45}$$



Model: $y_t = 0.9x_t + \epsilon_t$,
where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t, \xi_t$ is iid N
$x_t = \kappa x_{t-1} + \eta_t, \eta_t$ is iid N

u_t is the residual from LS estimate of
$y_t = a + bx_t + u_t$

Number of simulations: 25000

Figure 2.19: Autocorrelation of $x_t u_t$ when u_t has autocorrelation ρ

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 middle term becomes

$$\sum_{t=1}^{T} \text{Var}\left(x_t u_t\right) + 2\sum_{s=1}^{m} \sum_{t=s+1}^{T} \text{Cov}\left(x_t u_t, x_{t-s} u_{t-s}\right), \tag{2.46}$$

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 (2.46). For instance, with only one lag ($m = 1$) the calculation is (with several regressors) $\sum_{t=1}^{T} x_t x_t' \hat{u}_t^2 + \sum_{t=2}^{T} \left(x_t x_{t-1}' + x_{t-1} x_t'\right) \hat{u}_t \hat{u}_{t-1}$. Notice also that by exclduing all lags (setting $m = 0$), the Newey-West estimator concides with White's estimator. Hence, Newey-West estimator handles also heteroskedasticity.

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

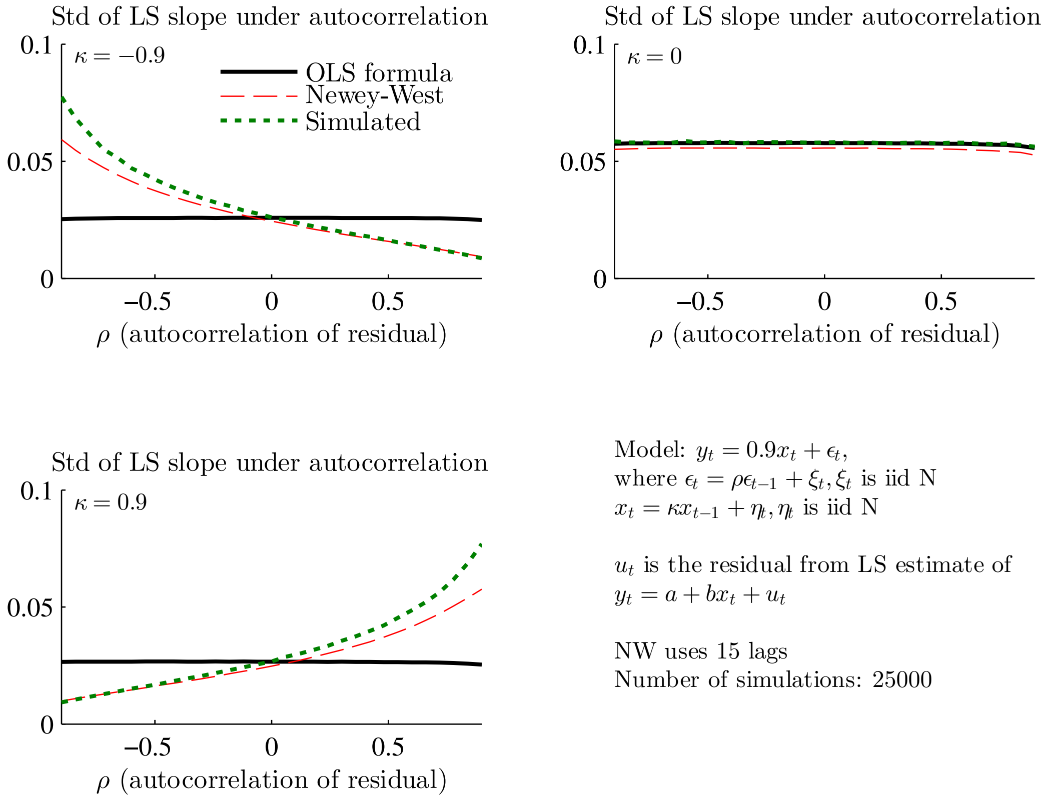


Figure 2.20: Standard error of OLS slope, autocorrelated errors

positively autocorrelated. In contrast, when the regressor has no autocorrelation, then the product does not either. This is illustrated in Figures 2.19–2.21.

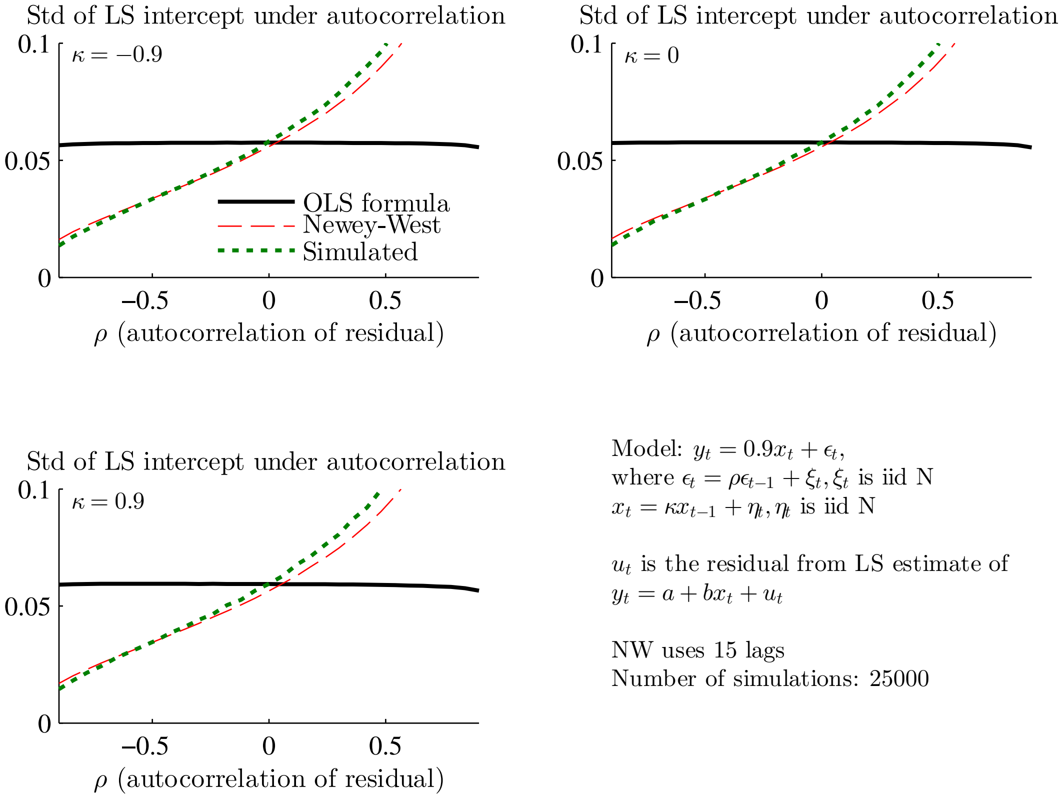
Figures 2.22–2.23 are empirical examples of the importance of using the Newey-West method rather than relying of the iid assumptions. In both cases, the residuals have strong positive autocorrelation.

A 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



Model: $y_t = 0.9x_t + \epsilon_t$,
where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t, \xi_t$ is iid N
$x_t = \kappa x_{t-1} + \eta_t, \eta_t$ is iid N

u_t is the residual from LS estimate of
$y_t = a + bx_t + u_t$

NW uses 15 lags
Number of simulations: 25000

Figure 2.21: Standard error of OLS intercept, autocorrelated errors

element by element

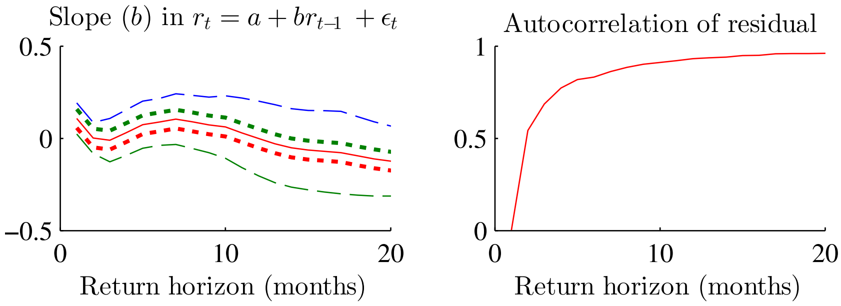
$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + c = \begin{bmatrix} A_{11}+c & A_{12}+c \\ A_{21}+c & A_{22}+c \end{bmatrix}$$

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

Example A.1

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

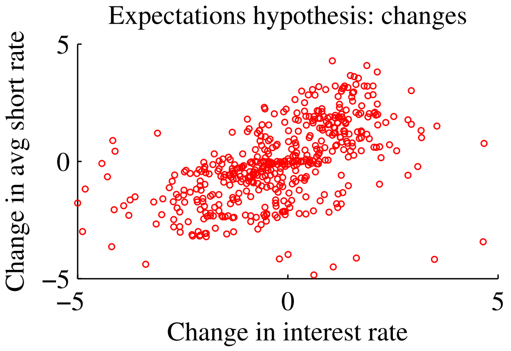
$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} 10 = \begin{bmatrix} 10 & 30 \\ 30 & 40 \end{bmatrix}.$$



Slope with two different 90% conf band, OLS and NW std

Monthly US stock returns 1926:1-2012:12, overlapping data

Figure 2.22: Slope coefficient, LS vs Newey-West standard errors



Overlapping US 12−month interest rates and
next−year average federal funds rate: 1970:1−2012:3

Slope coefficient: 0.54
Std (classical and Newey−West): 0.04 0.13
Autocorrelation of residual: 0.96

Figure 2.23: US 12-month interest and average federal funds rate (next 12 months)

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 A.2 *(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'

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

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 A.3 *(Matrix transpose)*

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

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

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 ith row of the first matrix with the jth 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 A.4 *(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*

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

and xz' the *outer product*

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

(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 A.5 *(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 A.6 *(Matrix inverse) We have*

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

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

A Statistical Tables

n	Critical values		
	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 A.1: Critical values (two-sided test) of t distribution (different degrees of freedom) and normal distribution.

Bibliography

Verbeek, M., 2008, *A guide to modern econometrics*, Wiley, Chichester, 3rd edn.

n	Critical values		
	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 A.2: Critical values of chisquare distribution (different degrees of freedom, n).

$n1$			$n2$			$\chi^2_{n1}/n1$
	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 A.3: 5% Critical values of $F_{n1,n2}$ distribution (different degrees of freedom).

$n1$	$n2$					$\chi^2_{n1}/n1$
	10	30	50	100	300	
1	3.29	2.88	2.81	2.76	2.72	2.71
2	2.92	2.49	2.41	2.36	2.32	2.30
3	2.73	2.28	2.20	2.14	2.10	2.08
4	2.61	2.14	2.06	2.00	1.96	1.94
5	2.52	2.05	1.97	1.91	1.87	1.85
6	2.46	1.98	1.90	1.83	1.79	1.77
7	2.41	1.93	1.84	1.78	1.74	1.72
8	2.38	1.88	1.80	1.73	1.69	1.67
9	2.35	1.85	1.76	1.69	1.65	1.63
10	2.32	1.82	1.73	1.66	1.62	1.60

Table A.4: 10% Critical values of $F_{n1,n2}$ distribution (different degrees of freedom).

3 Regression Diagnostics*

3.1 Misspecifying the Set of Regressors

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 tot really dangerous, but is likely to decrease the precision.

To selecting the regressors, apply 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 and test restrictions,..., keep making it simpler until restrictions are rejected.

Remember that R^2 can never decrease by adding more regressors—not really a good guide. To avoid overfitting, "punish" models with to many parameters. Perhaps consider $\bar{R}^2$ instead

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k}, \tag{3.1}$$

where 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 $0 \leq R^2 \leq 1$). Alternatively, apply Akaike's Information Criterion (AIC) and the Bayesian information criterion (BIC) instead. They are

$$AIC = \ln \hat{\sigma}^2 + 2\frac{k}{T} \tag{3.2}$$

$$BIC = \ln \hat{\sigma}^2 + \frac{k}{T}\ln T. \tag{3.3}$$

These measure also involve a trade-off between fit (low $\hat{\sigma}^2$) 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 $R^2 = 1 - \sigma^2/\operatorname{Var}(y_t)$ so $\sigma^2 = \operatorname{Var}(y_t)(1 - R^2)$) that AIC and

BIC can be rewritten as

$$AIC = \ln \mathrm{Var}(y_t) + \ln(1 - R^2) + 2\frac{k}{T} \tag{3.4}$$

$$BIC = \ln \mathrm{Var}(y_t) + \ln(1 - R^2) + \frac{k}{T} \ln T. \tag{3.5}$$

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 $\bar{R}^2$.

3.2 Comparing Non-Nested Models

When one model is not a special case of another

$$\text{Model A: } y_t = x_t'\beta + \varepsilon_t \tag{3.6}$$

$$\text{Model B: } y_t = z_t'\gamma + v_t \tag{3.7}$$

Non-nested if z is not a subset of x at the same time as x is not a subset of z. For instance, these models could represent alternative economic theories of the same phenomenon. 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). To test this, run the regression

$$y_t = z_t'\gamma + x_{2t}'\delta_A + v_t, \tag{3.8}$$

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.

3.3 Non-Linear Models

Regression analysis typically start with assuming a linear model—which may or may not be a good approximation.

Notice that models that are *non-linear in variables*

$$y_t = a + bx_t^{3.4} + \varepsilon_t, \tag{3.9}$$

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*

$$y_t = \beta_1 + \beta_2 x_t^{\beta_3} + u_t \tag{3.10}$$

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

$$y_t = x_t'\beta + \alpha_2 \hat{y}_t^2 + v_t, \tag{3.11}$$

where $\hat{y}_t = x_t'\hat{b}$ (from linear estimation). Test if $\alpha_2 = 0$.

3.4 Outliers

LS is sensitive to extreme data points. Maybe we need to understand if there are outliers, by plotting data and some regression results.

The starting point (as always in empirical work) is to plot the data: time series plots and histograms—to see if there are extreme data points.

As complement, it is a good idea to try to identify outliers from the regression results. First, estimate on 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 estimates $\hat{\beta}^{(s)}$ 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 $\hat{u}_s^{(s)}/\hat{\sigma}$. 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).

3.5 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

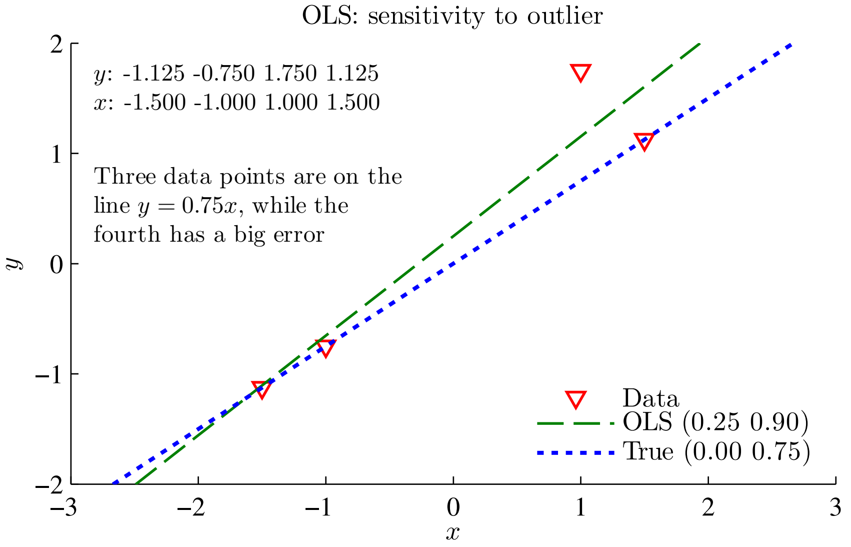


Figure 3.1: Data and regression line from OLS

z_t)

$$y_t = x_t'\beta + g_t z_t'\gamma + \varepsilon_t, \text{ where} \tag{3.12}$$

$$g_t = \begin{cases} 1 & \text{for some subsample} \\ 0 & \text{else} \end{cases} \tag{3.13}$$

and test $\gamma = \mathbf{0}$ (a "Chow test"). Notice that γ measures the change of the coefficients (from one sub sample to another)..

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

$$y_t = x_t'\beta + \varepsilon_t \tag{3.14}$$

on a longer and longer data set ("recursive estimation"). In the standard recursive estimation, the first estimation is done on the sample $t = 1, 2, \ldots, \tau$; while the second estimation is done on $t = 1, 2, \ldots, \tau, \tau + 1$; and so forth until we use the entire sample $t = 1 \ldots, T$. 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 $T - \tau, \ldots, T$; the second $T - \tau - 1, \ldots, T$; and so forth.

Alternatively, a moving data window ("rolling samples") could be used. In this case, the first sample is $t = 1, 2, \ldots, \tau$; but the second is on $t = 2, \ldots, \tau, \tau + 1$, that is, by dropping one observation at the start when the sample is extended at the end. See Figure 3.2 for an illustration.

An alternative is to 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

$$\tilde{x}_t = x_t \lambda^{T-t} \text{ and } \tilde{y}_t = y_t \lambda^{T-t} \tag{3.15}$$

and then estimate

$$\tilde{y}_t = \tilde{x}_t' \beta + \varepsilon_t. \tag{3.16}$$

Notice that also the constant (in x_t) should be scaled in the same way. (Clearly, this method is strongly related to the GLS approach used when residuals are heteroskedastic. Also, the idea of down weighting old data is commonly used to estimate time-varying volatility of returns as in the RISK metrics method.)

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 Figures 3.2–3.3 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 $t = 1, 2, \ldots, \tau$ and use the estimated coefficients (denoted $\beta^{(\tau)}$) to calculate a "forecast" and "forecast error" for $\tau + 1$ as

$$\hat{y}_{\tau+1} = x_{\tau+1}' \beta^{(\tau)} \text{ and } v_{\tau+1} = y_{\tau+1} - \hat{y}_{\tau+1}. \tag{3.17}$$

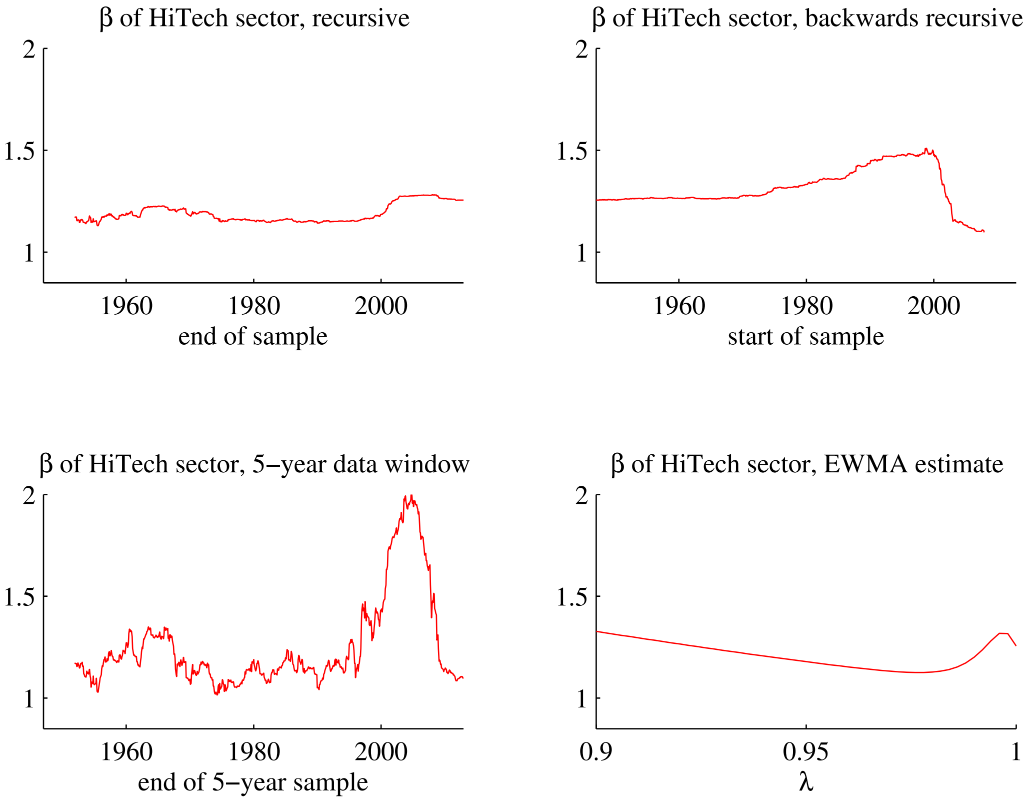


Figure 3.2: Betas of US industry portfolios

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

$$\hat{y}_{\tau+2} = x'_{\tau+2}\beta^{(\tau+1)} \text{ and } v_{\tau+2} = y_{\tau+2} - \hat{y}_{\tau+2}. \tag{3.18}$$

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

$$\begin{aligned} W_\tau &= \frac{v_{\tau+1}}{\sigma} \\ W_{\tau+1} &= \frac{v_{\tau+1} + v_{\tau+2}}{\sigma} \\ W_{\tau+2} &= \frac{v_{\tau+1} + v_{\tau+2} + v_{\tau+3}}{\sigma} \end{aligned} \tag{3.19}$$

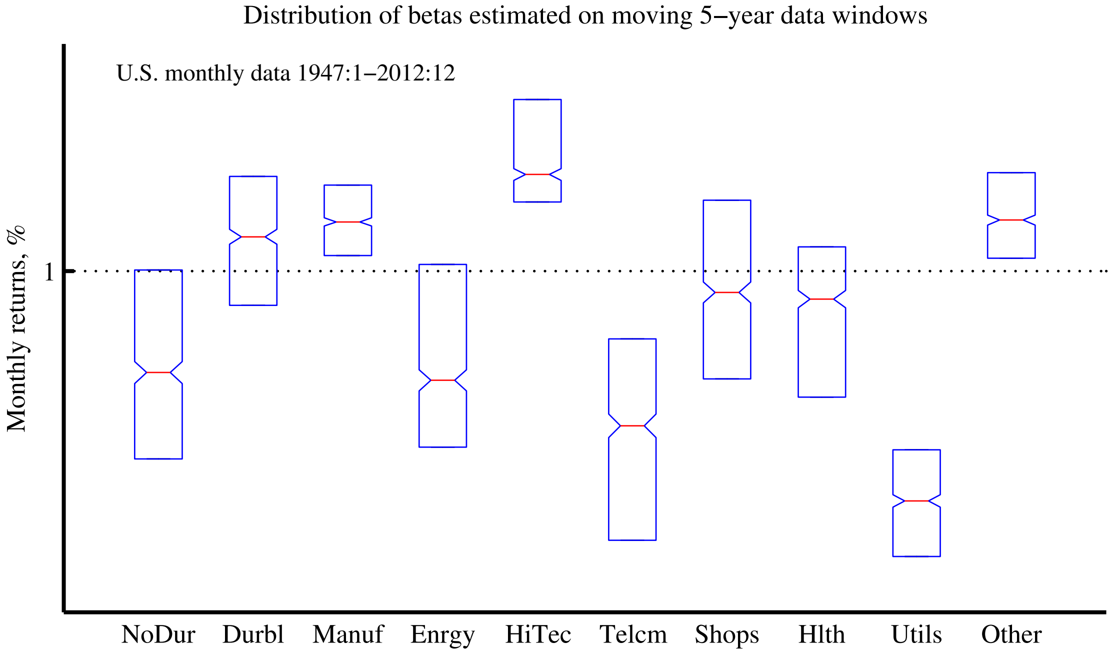


Figure 3.3: Distribution of betas of US industry portfolios (estimated on 5-year data windows)

and so forth. More generally we have the sequence

$$W_t = \sum_{s=\tau}^{t} \frac{v_{s+1}}{\sigma}, \text{ for } t = \tau, ..., T - 1. \tag{3.20}$$

Sixth and finally, plot W_t along with a 95% confidence interval: $\pm 0.948 \left(\sqrt{T-\tau} + 2(t-\tau)/\sqrt{T-\tau}\right)$. Reject stability if any observation is outside.

3.6 Robust Estimation*

3.6.1 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

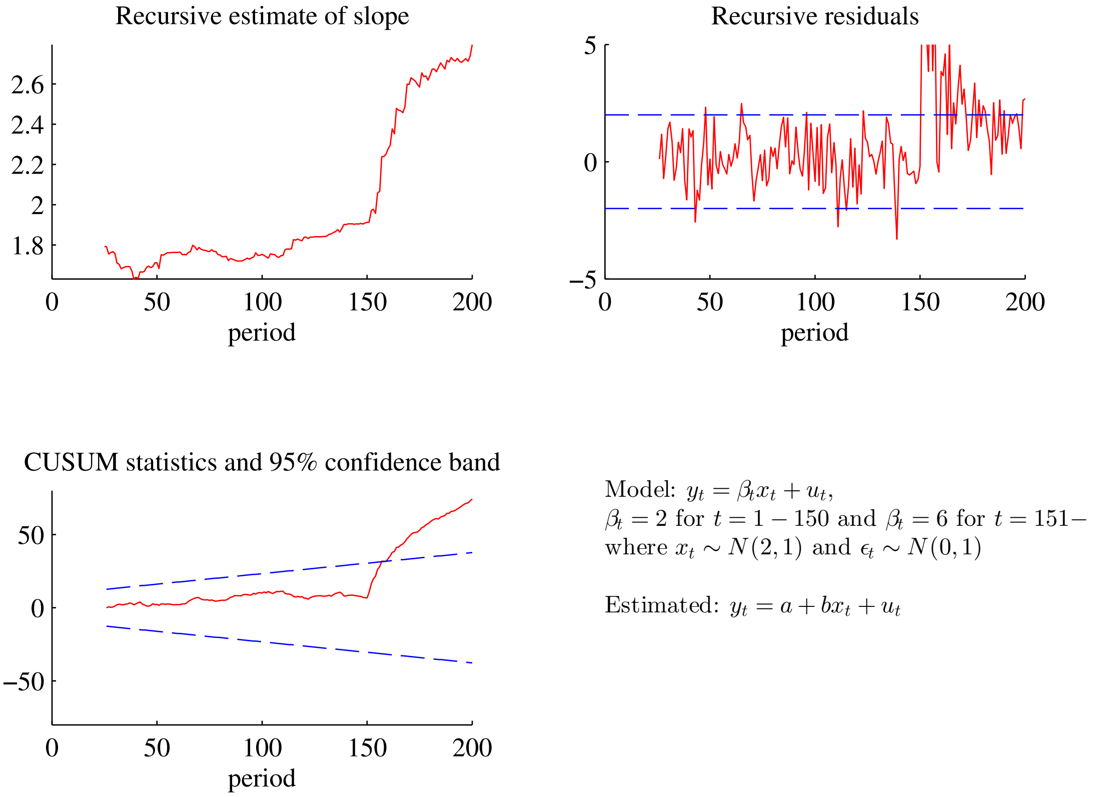


Figure 3.4: Stability test

then decide which data points to exclude. Alternatively, robust estimators can be applied instead of the traditional ones.

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

$$\text{StdRobust} = [\text{quantile}(0.75) - \text{quantile}(0.25)]/1.35, \tag{3.21}$$

or by the *median absolute deviation*

$$\text{StdRobust} = \text{median}(|x_t - \mu|)/0.675. \tag{3.22}$$

Both these would coincide with the standard deviation if data was indeed drawn from a

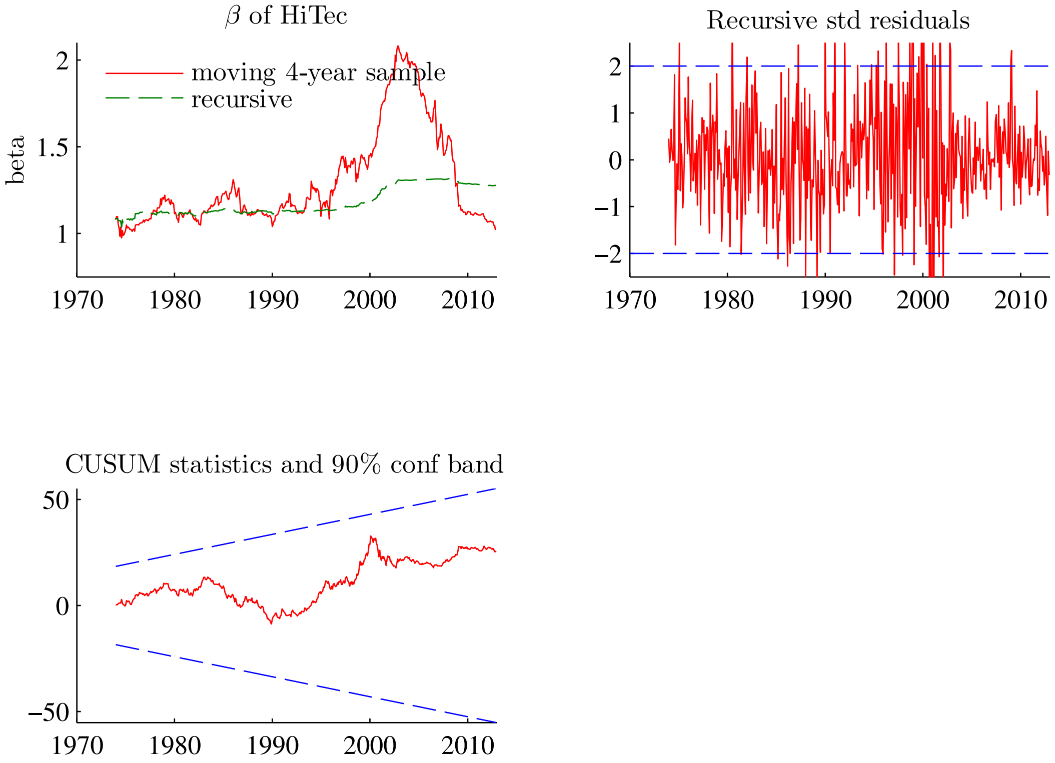


Figure 3.5: CAPM regression on a US industry index

normal distribution without outliers.

A robust covariance can be calculated by using the identity

$$\mathrm{Cov}(x, y) = [\mathrm{Var}(x + y) - \mathrm{Var}(x - y)]/4 \tag{3.23}$$

and using a robust estimator of the variances—like the square of (3.21). A robust correlation is then created by dividing the robust covariance with the two robust standard deviations.

See Figures 3.6–3.7 for empirical examples.

3.6.2 Robust Regression Coefficients

Reference: Amemiya (1985) 4.6

The least absolute deviations (LAD) estimator miminizes the sum of absolute residu-

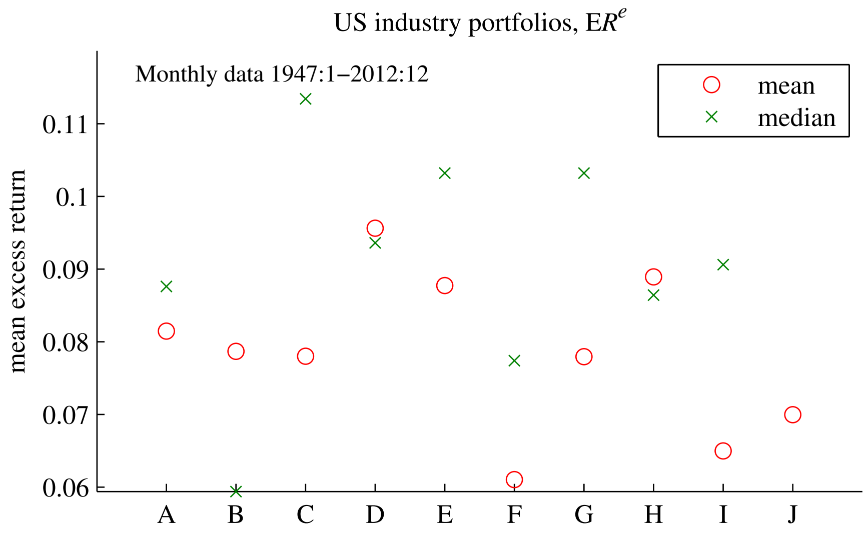


Figure 3.6: Mean excess returns of US industry portfolios

als (rather than the squared residuals)

$$\hat{\beta}_{LAD} = \arg\min_{b} \sum_{t=1}^{T} \left| y_t - x_t' b \right| \tag{3.24}$$

This estimator involve non-linearities, but a simple iteration works nicely. It is typically less sensitive to outliers. (There are also other ways to estimate robust regression coefficients.) This is illustrated in Figure 3.8.

See Figure 3.9 for an empirical example.

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

$$\sqrt{T}(\hat{\beta}_{LAD} - \beta_0) \to^d N\left[0, f(0)^{-2}\Sigma_{xx}^{-1}/4\right], \text{ where } \Sigma_{xx} = \text{plim} \sum\nolimits_{t=1}^{T} x_t x_t'/T, \tag{3.25}$$

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.

Example 3.1 *($N(0,\sigma^2)$) When $u_t \sim N(0,\sigma^2)$, then $f(0) = 1/\sqrt{2\pi\sigma^2}$, so the covariance matrix in (3.25) becomes $\pi\sigma^2\Sigma_{xx}^{-1}/2$. This is $\pi/2$ times larger than when using LS.*

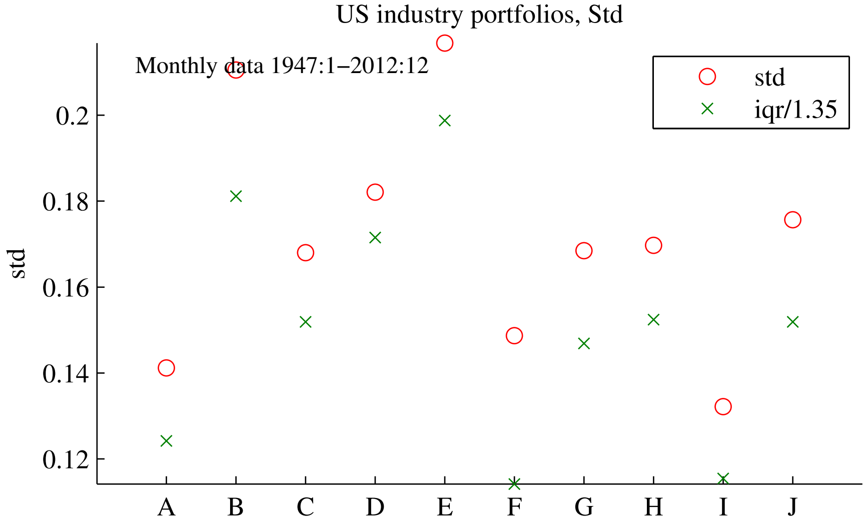


Figure 3.7: Volatility of US industry portfolios

Remark 3.2 *(Algorithm for LAD) The LAD estimator can be written*

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

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

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

$$\hat{\beta}_{LMS} = \arg\min_{\beta} \left[\text{median}\left(\hat{u}_t^2\right)\right], \text{ with } \hat{u}_t = y_t - x_t'\hat{b} \tag{3.26}$$

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} \hat{u}_i^2,\ \hat{u}_1^2 \le \hat{u}_2^2 \le \ldots \text{ and } h \le T. \tag{3.27}$$

Note that the LTS estimator in (3.27) minimizes the sum of the h smallest squared residuals.

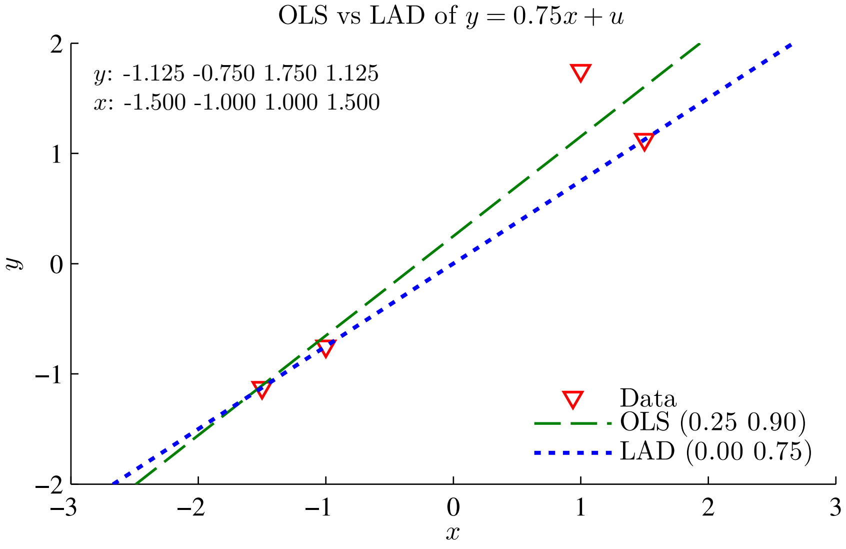


Figure 3.8: Data and regression line from OLS and LAD

Bibliography

Amemiya, T., 1985, *Advanced econometrics*, Harvard University Press, Cambridge, Massachusetts.

Enders, W., 2004, *Applied econometric time series*, John Wiley and Sons, New York, 2nd edn.

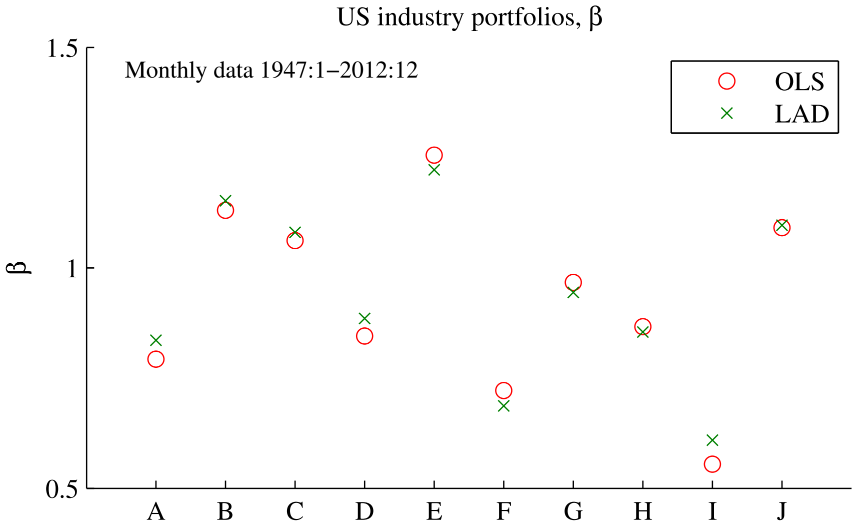


Figure 3.9: Betas of US industry portfolios

4 Asymptotic Results on OLS*

4.1 Properties of the OLS Estimator when “Gauss-Markov” Is False

There are severakl problems when the Gauss-Marlov assumptions are wrong. First, the result that $\mathrm{E}\,\hat{\beta} = \beta$ (unbiased) relied on the assumption that the regressors are fixed or (altermatively) that $\{u_1, ..., u_T\}$ and $\{x_1, ..., x_T\}$ are independent. Otherwise not true (in a finite sample). Second, theresult that $\hat{\beta}$ is normally distributed relied on residuals being normally distributed. Otherwise 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) simulations; *(b)* find results for $T \to \infty$ (“asymptotic properties”) and use as approximation.

4.2 Motivation of Asymptotics

The results from asymptotoc theiry are more general (and prettier) than simulations—and can be used as approximation if sample is large. The basic reasons for this is that most estimators are sample averages and sample averages often have nice properties as $T \to \infty$. In particular, we can make use of the law of large numbers (LLN) and the central limit theorem (CLT). See Figure 4.2

4.3 Asymptotics: Consistency

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

Consistency: if Prob($\hat{\beta}$ deviates much from β) $\to 0$ as $T \to \infty$. (Notation: $\operatorname{plim}\hat{\beta} = \beta$)

LLN: (simple version...) $\operatorname{plim}(\bar{x}) = \mathrm{E}(x)$. OLS (and most other estimators) are sample averages of some sort.

As a simple special case, suppose there is only one regressor and that both the dependent variable and the regressor have zero means. The OLS estimate of the slope coefficient

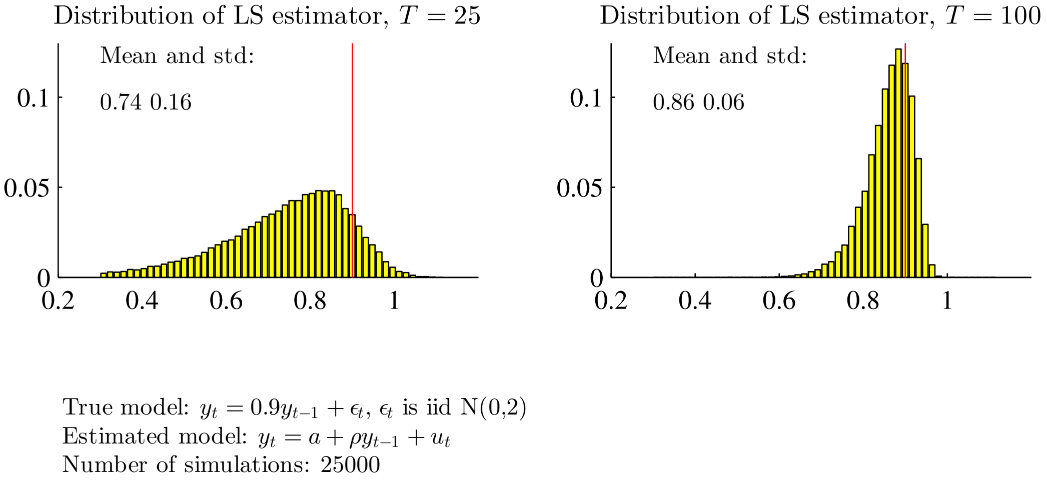


Figure 4.1: Distribution of LS estimator of autoregressive parameter

is then

$$\hat{\beta} = \beta + \left(\sum\nolimits_{t=1}^{T} x_t x_t\right)^{-1} \sum_{t=1}^{T} x_t u_t \tag{4.1}$$

$$= \beta + \left(\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t x_t\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t u_t, \tag{4.2}$$

where u_t are the residuals we could calculate if we knew the true slope coefficient, that is, the true residuals.

This estimate has the probability limit

$$\text{plim}\,\hat{\beta} = \beta + \Sigma_{xx}^{-1}\,\text{E}(x_t u_t), \tag{4.3}$$

where Σ_{xx}^{-1} is a matrix of constants. The key point: is $\text{E}(x_t u_t) = 0$. If not, OLS is not consistent.

Some observations:

1. We can not (easily) test this. OLS *creates* $\hat{\beta}$ and the fitted residuals $\hat{u}_t$ such that $\sum_{t=1}^{T} x_t \hat{u}_t = 0$.

2. The Gauss-Markov assumption that u_t and x_j are independent implies that $\text{E}(x_t u_t) = 0$, so the Gauss-Markov assumptions basically disregards the issue of consistency.

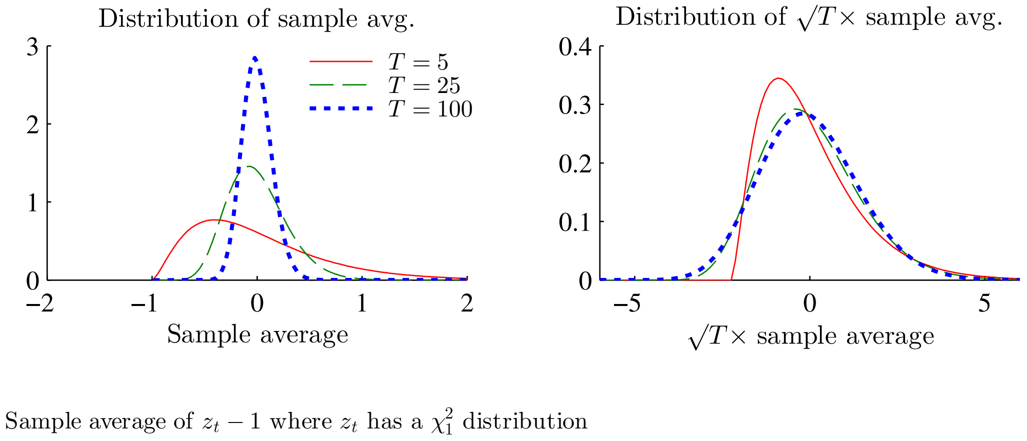


Figure 4.2: Distribution of sample averages

(Assumes that it does not exist.)

3. OLS can be biased, but still be consistent—so it is ok if sample is large. See Figures 4.1 and 4.3. Notice $\text{Cov}(u_{t-1}, x_t) \neq 0$ so biased, but $\text{Cov}(u_t, x_t) = 0$ so consistent)

4. There are cases when $\text{E}(x_t u_t) = 0$ doesn't make sense. More on this later.

5. See Figures 4.1 and 4.3 for examples where OLS is consistent, and Figure 4.4 when it is not.

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

4.4 When LS Cannot be Saved

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

Q. When do we have $\text{Corr}(x, u) \neq 0$?

A. Need to think hard...

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

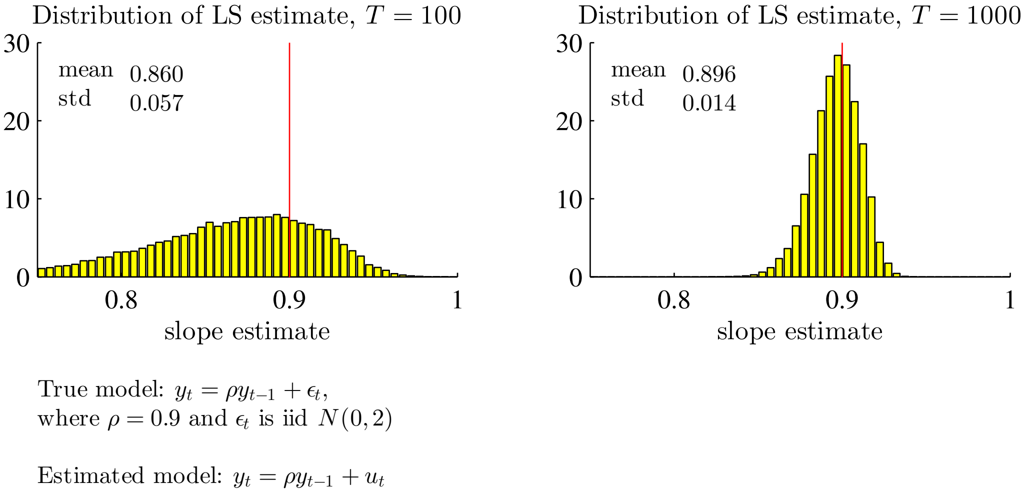


Figure 4.3: Distribution of LS estimator of autoregressive parameter

4.4.1 When LS Cannot be Saved: Excluded Variables

Correct model for log wages

$$y_t = x_{1t}'\beta_1 + x_{2t}\beta_2 + a_t\gamma + v_t \tag{4.4}$$

where x_{1t} measure individual characteristics, x_{2t} years of schooling and u_t ability. Assume that $\gamma > 0$ and $\text{Cov}(x_{2t}, a_t) > 0$ (people with more ability tend to have longer schooling).

Suppose we cann measure ability and therefore estimate

$$\begin{aligned} y_t &= x_{1t}'\beta_1 + x_{2t}\beta_2 + \underbrace{u_t}_{a_t\gamma + v_t} \\ &= x_t'\beta + u_t. \end{aligned} \tag{4.5}$$

From (4.3) $\text{plim}\,\hat{\beta} = \beta + \Sigma_{xx}^{-1}\,\text{E}(x_t u_t)$, so assuming $\text{E}(x_t v_t) = 0$ (that the residual in (4.4) is not correlated with the regressors) gives

$$\begin{aligned} \text{plim}\,\hat{\beta} &= \beta + \Sigma_{xx}^{-1}\,\text{E}(x_t u_t) \\ &= \beta + \Sigma_{xx}^{-1}\,\text{E}(x_t a_t)\gamma. \end{aligned} \tag{4.6}$$

If x_t and a_t are related so $\text{E}(x_t a_t) \neq 0$, then OLS in not consistent. For instance, if x_t

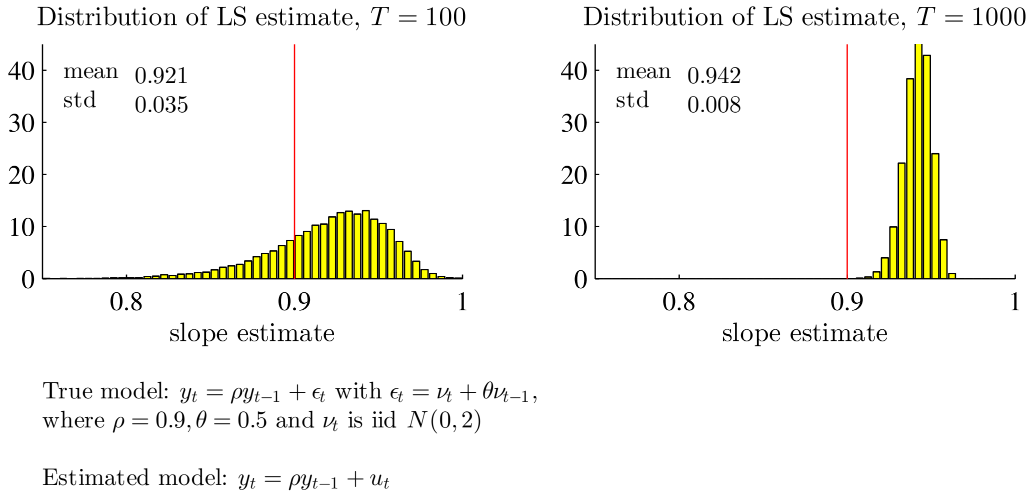


Figure 4.4: Distribution of LS estimator of autoregressive parameter

are zero mean variables, then $\mathrm{E}(x_t a_t) = \mathrm{Cov}(x_t, a_t)$. In our example of how scholling affects wages, $\mathrm{Cov}(x_t, a_t) > 0$ seems reasonable, so $\hat{\beta}_2 > \beta_s$. That is, the OLS estimate is likely to overestimate the returns to schooling β_2, sinnce the estimate $\hat{\beta}_2$ captures also the effect of the excluded ability. In contrast, excluding something that is uncorrelated with other regressors does not create a problem.

Notice the following:

- $\hat{\beta}$ *is* 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 (which clearly also has an effect on y_t).

- $\hat{\beta}_2$ *is not* the right number to use if we want to understand an economic mechanism: "if we increase schooling, x_{2t}, by one unit (but holding all other variables constant), what is the likely effect on y_t?" The reason is that we here need β_2 (or at least a consistent estimate of it).

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

(macroeconomics)

Suppose y_t depends on lags of itself (inertia in the economy?), but the residual is autoregressive

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 \underbrace{y_{t-1}} + u_t \text{ and} \tag{4.7}$$

$$u_t = \upsilon_t + \theta \upsilon_{t-1}, \upsilon_t \text{ iid.} \tag{4.8}$$

This leads to $\text{Cov}(y_{t-1}, u_t) \neq 0$. To see why: $\text{Cov}(y_{t-1}, u_t) = \text{Cov}(y_{t-1}, \upsilon_t + \theta \upsilon_{t-1}) > 0$ (if $\theta > 0$): a positive bias.

As a special case, $\beta_2 = 0$ gives an ARMA(1,1) model, which cannot be estimated by OLS. See Figure 4.4.

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

(microeconomics)

Suppose the correct model

$$y_t = \beta_1 + \beta_2 w_t + \upsilon_t, \tag{4.9}$$

but we estimate with proxy x_t for w_t

$$y_t = \beta_1 + \beta_2 x_t + u_t, \text{ with} \tag{4.10}$$

$$x_t = w_t + \underbrace{e_t}_{\text{measurement error}}. \tag{4.11}$$

In this equation e_t is the (zero mean) measurement error—and we typically assume that it is uncorrelated with the true value (w_t)

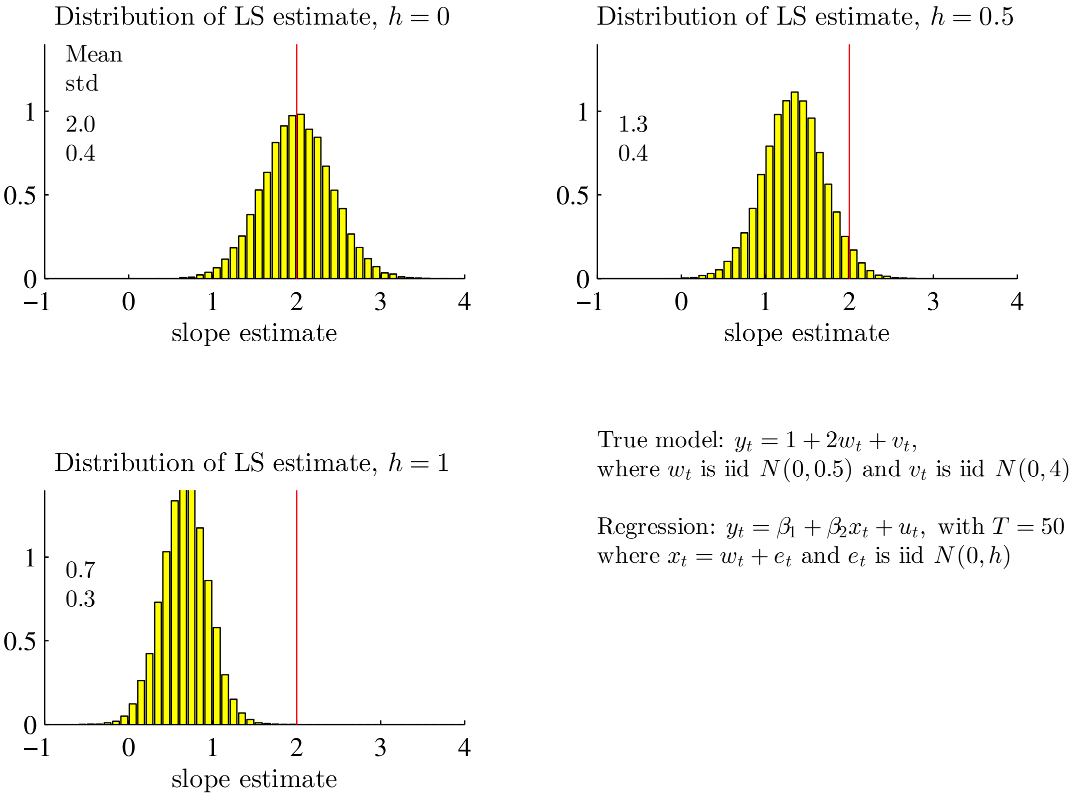
This leads to $\text{Cov}(x_t, u_t) \neq 0$, so OLS is inconsistent. See Figure 4.5.

To see why, solve for $w_t = x_t - e_t$, use in correct model (4.9)

$$\begin{aligned} y_t &= \beta_1 + \beta_2 (x_t - e_t) + \upsilon_t \\ &= \beta_1 + \beta_2 x_t \underbrace{- \beta_2 e_t + \upsilon_t}_{u_t} \end{aligned} \tag{4.12}$$

and from (4.11) we know that x_t is correlated with e_t. In fact, it can be shown that

$$\text{plim}\, \hat{\beta}_2 = \beta_2 \left(1 - \frac{\text{Var}(e_t)}{\text{Var}(w_t) + \text{Var}(e_t)} \right) \tag{4.13}$$



True model: $y_t = 1 + 2w_t + v_t$,
where w_t is iid $N(0, 0.5)$ and v_t is iid $N(0, 4)$

Regression: $y_t = \beta_1 + \beta_2 x_t + u_t$, with $T = 50$
where $x_t = w_t + e_t$ and e_t is iid $N(0, h)$

Figure 4.5: Effect of measurement error in regressor,h is the variance of the errors

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

Proof. (of (4.13)) To simplify, assume that x_t has a zero mean. From (4.3), we then have $\operatorname{plim} \hat{\beta}_2 = \beta_2 + \Sigma_{xx}^{-1} \operatorname{E}(x_t u_t)$. Here, $\Sigma_{xx}^{-1} = 1/\operatorname{Var}(x_t)$, but notice from (4.11) that $\operatorname{Var}(x_t) = \operatorname{Var}(w_t) + \operatorname{Var}(e_t)$ if w_t and e_t are uncorrelated. We also have $\operatorname{E}(x_t u_t) = \operatorname{Cov}(x_t, u_t)$, which from the definition of x_t in (4.11) and of u_t in (4.12) gives

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

Together we get

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

which is (4.13). ■

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

(micro and macro) A simplistic macro model

$$C_t = \beta_1 + \beta_2 Y_t + u_t \tag{4.14}$$

$$Y_t = C_t + I_t \tag{4.15}$$

$$I_t \text{ and } \varepsilon_t \text{ are independent (exogenous)} \tag{4.16}$$

Could be generalized to let Y_t have more action.

Key point: $u_t \to C_t \to Y_t \to C_t \Rightarrow \text{Cov}(Y_t, u_t) > 0$: a positive bias if we try to estimate the consumption equation (4.14).

4.5 Asymptotic Normality

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

CLT: (simple version...) $\sqrt{T}\bar{x} \sim N()$ when T becomes really large. Holds for most random variables. Notice: the distribution of $\bar{x}$ converges to a spike as T increases, but the distribution of $\sqrt{T}\bar{x}$ converges to a nice normal. See Figure 4.2.

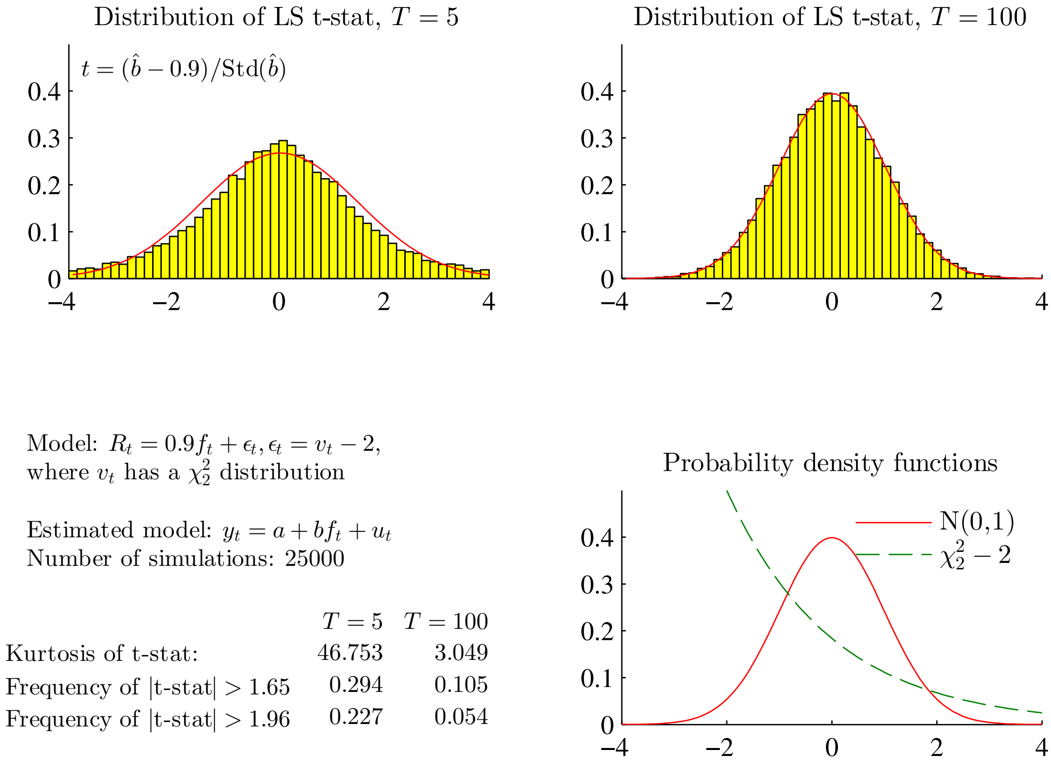
Subtract β from both sides of (4.2), multiply both sides by $\sqrt{T}$

$$\sqrt{T}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t x_t'\right)^{-1}}_{\to \Sigma_{xx}^{-1}} \underbrace{\sqrt{T}\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t u_t}_{\sqrt{T}\times\text{sample average}} \tag{4.17}$$

The first term converges (by a LLN) to a constant, while the second term is $\sqrt{T}\times$sample average (of $x_t u_t$). We should therefore expect that $\hat{\beta}$ is normally distributed in *large* samples—even if the residual doesn't have a normal distribution. See Figure 4.6 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

$$\sqrt{T}(\hat{\beta} - \beta) \to N\left(0, \sigma^2 \Sigma_{xx}^{-1}\right) \text{ or ``}\hat{\beta} \to N\left(\beta, \sigma^2 \Sigma_{xx}^{-1}/T\right)\text{''} \tag{4.18}$$



Model: $R_t = 0.9 f_t + \epsilon_t, \epsilon_t = v_t - 2$,
where v_t has a χ^2_2 distribution

Estimated model: $y_t = a + b f_t + u_t$
Number of simulations: 25000

	$T = 5$	$T = 100$
Kurtosis of t-stat:	46.753	3.049
Frequency of $\lvert$t-stat$\rvert > 1.65$	0.294	0.105
Frequency of $\lvert$t-stat$\rvert > 1.96$	0.227	0.054

Figure 4.6: Results from a Monte Carlo experiment with thick-tailed errors.

Remark 4.1 *Step 1: If* $\mathrm{Var}[\sqrt{T}(\hat{\beta} - \beta)] = \sigma^2 \Sigma_{xx}^{-1}$*, then* $\mathrm{Var}[\sqrt{T}(\hat{\beta} - \beta)/\sqrt{T}] = \sigma^2 \Sigma_{xx}^{-1}/T$*; step 2: if* $\mathrm{E}(\hat{\beta} - \beta) = 0$*, then* $\mathrm{E}(\hat{\beta}) = \beta$*.*

Bibliography

5 Index Models

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

5.1 The Inputs to a MV Analysis

To calculate the mean variance frontier we need to calculate both the expected return and variance of different portfolios (based on n assets). With two assets ($n = 2$) the expected return and the variance of the portfolio are

$$\begin{aligned}
\mathrm{E}(R_p) &= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\
\sigma_P^2 &= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.
\end{aligned} \tag{5.1}$$

In this case we need information on 2 mean returns and 3 elements of the covariance matrix. Clearly, the covariance matrix can alternatively be expressed as

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{5.2}$$

which involves two variances and one correlation (as before, 3 elements).

There are two main problems in estimating these parameters: the number of parameters increase very quickly as the number of assets increases and historical estimates have proved to be somewhat unreliable for future periods.

To illustrate the first problem, notice that with n assets we need the following number of parameters

	Required number of estimates	With 100 assets
μ_i	n	100
σ_{ii}	n	100
σ_{ij}	$n(n-1)/2$	4950

The numerics is not the problem as it is a matter of seconds to estimate a covariance matrix of 100 return series. Instead, the problem is that most portfolio analysis uses lots of judgemental "estimates." These are necessary since there might be new assets (no historical returns series are available) or there might be good reasons to believe that old estimates are not valid anymore. To cut down on the number of parameters, it is often assumed that returns follow some simple model. These notes will discuss so-called single- and multi-index models.

The second problem comes from the empirical observations that estimates from historical data are sometimes poor "forecasts" of future periods (which is what matters for portfolio choice). As an example, the correlation between two asset returns tends to be more "average" than the historical estimate would suggest.

A simple (and often used) way to deal with this is to replace the historical correlation with an average historical correlation. For instance, suppose there are three assets. Then, estimate ρ_{ij} on historical data, but use the average estimate as the "forecast" of all correlations:

$$\text{estimate} \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ & 1 & \rho_{23} \\ & & 1 \end{bmatrix}, \text{calculate } \bar{\rho} = (\hat{\rho}_{12} + \hat{\rho}_{13} + \hat{\rho}_{23})/3, \text{ and use } \begin{bmatrix} 1 & \bar{\rho} & \bar{\rho} \\ & 1 & \bar{\rho} \\ & & 1 \end{bmatrix}.$$

5.2 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_m)

$$\begin{aligned} R_i &= \alpha_i + \beta_i R_m + e_i, \text{ where} \qquad (5.3) \\ \mathrm{E}(e_i) &= 0, \ \mathrm{Cov}\,(e_i, R_m) = 0, \text{ and } \mathrm{Cov}(e_i, e_j) = 0. \end{aligned}$$

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 (5.3) is estimated by Least Squares.

See Figures 5.1 – 5.3 for illustrations.

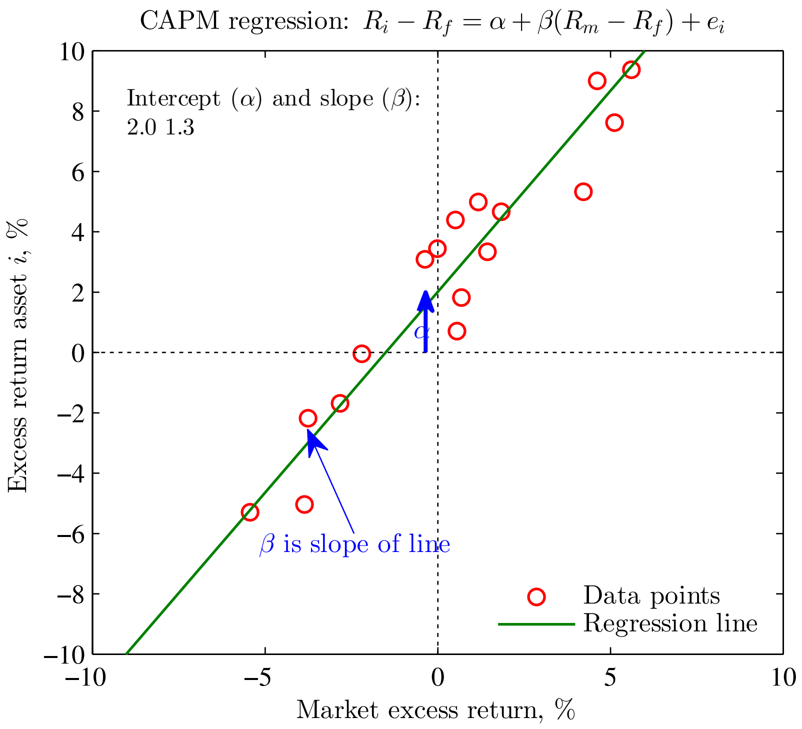


Figure 5.1: CAPM regression

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_i and R_j, say) are due to movements in the common "index" R_m. 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 many cases. In any case, it simplifies the construction of the covariance matrix of the assets enormously—as demonstrated below.

Remark 5.1 *(The market model) The market model is (5.3) without the assumption that* $\text{Cov}(e_i, e_j) = 0$. *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 (5.3) is true, then the variance of asset i and the covariance of assets i and j are

$$\sigma_{ii} = \beta_i^2 \text{Var}(R_m) + \text{Var}(e_i) \tag{5.4}$$

$$\sigma_{ij} = \beta_i \beta_j \text{Var}(R_m) . \tag{5.5}$$

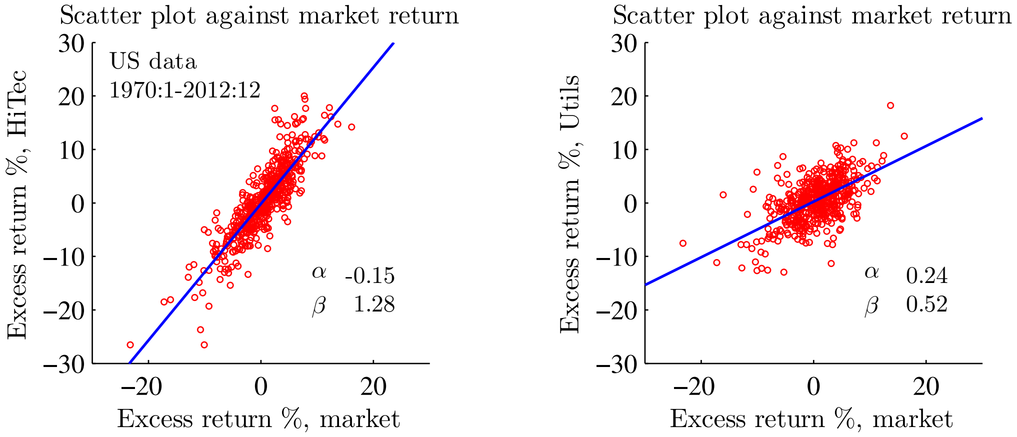


Figure 5.2: Scatter plot against market return

Together, these equations show that we can calculate the whole covariance matrix by having just the variance of the index (to get $\mathrm{Var}\,(R_m)$) and the output from n regressions (to get β_i and $\mathrm{Var}\,(e_i)$ 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)

$$\mathrm{Cov}\left(\begin{bmatrix} R_i \\ R_j \end{bmatrix}\right) = \begin{bmatrix} \beta_i^2 & \beta_i\beta_j \\ \beta_i\beta_j & \beta_j^2 \end{bmatrix} \mathrm{Var}\,(R_m) + \begin{bmatrix} \mathrm{Var}(e_i) & 0 \\ 0 & \mathrm{Var}(e_j) \end{bmatrix}, \text{ or} \tag{5.6}$$

$$= \begin{bmatrix} \beta_i \\ \beta_j \end{bmatrix} \begin{bmatrix} \beta_i & \beta_j \end{bmatrix} \mathrm{Var}\,(R_m) + \begin{bmatrix} \mathrm{Var}(e_i) & 0 \\ 0 & \mathrm{Var}(e_j) \end{bmatrix} \tag{5.7}$$

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

$$\mathrm{Cov}(R) = \beta\beta' \,\mathrm{Var}\,(R_m) + \Sigma. \tag{5.8}$$

See Figure 5.4 for an example based on the Fama-French portfolios detailed in Table 5.2.

	HiTec	Utils
constant	−0.15	0.24
	(−1.00)	(1.58)
market return	1.28	0.52
	(33.58)	(12.77)
R2	0.75	0.34
obs	516.00	516.00
Autocorr (t)	−0.73	0.86
White	6.19	20.42
All slopes	386.67	176.89

Table 5.1: CAPM regressions, monthly returns, %, US data 1970:1-2012:12. Numbers in parentheses are t-stats. Autocorr is a N(0,1) test statistic (autocorrelation); White is a chi-square test statistic (heteroskedasticity), df = K(K+1)/2 - 1; All slopes is a chi-square test statistic (of all slope coeffs), df = K-1

Remark 5.2 *(Fama-French portfolios) The portfolios in Table 5.2 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. For instance, in Table 5.2 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
Size 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 5.2: Numbering of the FF indices in the figures.

Proof. (of (5.4)–(5.5) By using (5.3) and recalling that $\text{Cov}(R_m, e_i) = 0$ direct calcu-

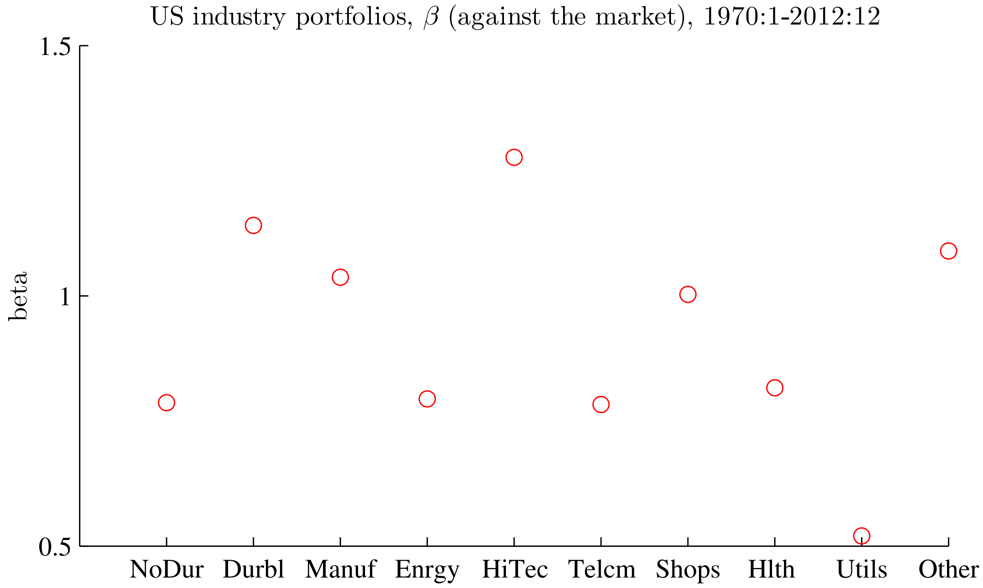


Figure 5.3: βs of US industry portfolios

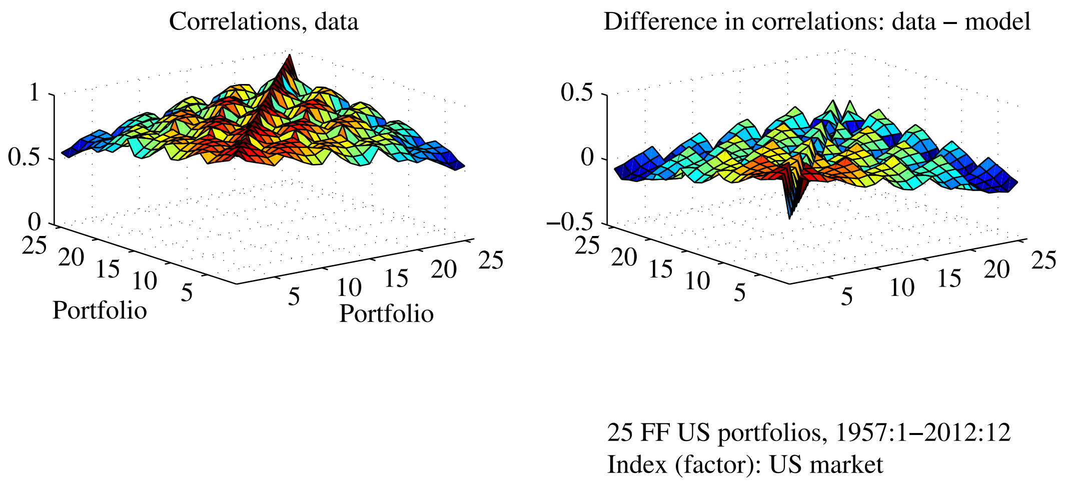
lations give

$$\begin{aligned}
\sigma_{ii} &= \text{Var}\left(R_i\right) \\
&= \text{Var}\left(\alpha_i + \beta_i R_m + e_i\right) \\
&= \text{Var}\left(\beta_i R_m\right) + \text{Var}\left(e_i\right) + 2 \times 0 \\
&= \beta_i^2 \text{Var}\left(R_m\right) + \text{Var}\left(e_i\right).
\end{aligned}$$

Similarly, the covariance of assets i and j is (recalling also that $\text{Cov}\left(e_i, e_j\right) = 0$)

$$\begin{aligned}
\sigma_{ij} &= \text{Cov}\left(R_i, R_j\right) \\
&= \text{Cov}\left(\alpha_i + \beta_i R_m + e_i, \alpha_j + \beta_j R_m + e_j\right) \\
&= \beta_i \beta_j \text{Var}\left(R_m\right) + 0 \\
&= \beta_i \beta_j \text{Var}\left(R_m\right).
\end{aligned}$$

■



Correlations, data

Difference in correlations: data − model

25 FF US portfolios, 1957:1–2012:12
Index (factor): US market

Figure 5.4: Correlations of US portfolios

5.3 Estimating Beta

5.3.1 Estimating Historical Beta: OLS and Other Approaches

Least Squares (LS) is typically used to estimate α_i, β_i and Std(e_i) in (5.3)—and the R^2 is used to assess the quality of the regression.

Remark 5.3 *(R^2 of market model) R^2 of (5.3) measures the fraction of the variance (of R_i) 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)*

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

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

$$\hat{\beta}_{i2} = \gamma_0 + \gamma_1 \hat{\beta}_{i1} + \upsilon_i \tag{5.9}$$

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

Other authors have suggested averaging the OLS estimate ($\hat{\beta}_{i1}$) with some average beta. For instance, $(\hat{\beta}_{i1}+1)/2$ (since the average beta must be unity) or $(\hat{\beta}_{i1}+\Sigma_{i=1}^{n}\hat{\beta}_{i1}/n)/2$ (which will typically be similar since $\Sigma_{i=1}^{n}\hat{\beta}_{i1}/n$ 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, $\hat{\beta}_{i1}$, and some other number, β_0, $(1 - F)\hat{\beta}_{i1} + F\beta_0$ 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_i 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 (5.3). If the residuals are normally distributed, so is the estimator—it is $N(\hat{\beta}_{i1}, \sigma_{\beta 1}^2)$, where we have taken the point estimate to be the mean. If we treat the variance of the LS estimator ($\sigma_{\beta 1}^2$) as known, then the Bayesian estimator of beta is

$$\begin{aligned} b &= (1 - F)\hat{\beta}_{i1} + F\beta_0, \text{ where} \\ F &= \frac{1/\sigma_0^2}{1/\sigma_0^2 + 1/\sigma_{\beta 1}^2} = \frac{\sigma_{\beta 1}^2}{\sigma_0^2 + \sigma_{\beta 1}^2}. \end{aligned} \tag{5.10}$$

When the prior beliefs are very precise ($\sigma_0^2 \to 0$), then $F \to 1$ 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 ($\sigma_0^2 \to \infty$), then $F \to 0$, 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.

5.3.2 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.

5.4 Multi-Index Models

5.4.1 Overview

The multi-index model is just a multivariate extension of the single-index model (5.3)

$$\begin{aligned} R_i &= a_i^* + \textstyle\sum_{k=1}^{K} b_{ik}^* I_k^* + e_i, \text{ where} \\ \mathrm{E}(e_i) &= 0,\ \mathrm{Cov}\left(e_i, I_k^*\right) = 0, \text{ and } \mathrm{Cov}(e_i, e_j) = 0. \end{aligned} \tag{5.11}$$

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 much simpler if the indices are transformed to be uncorrelated so we get the model

$$\begin{aligned} R_i &= a_i + \textstyle\sum_{k=1}^{K} b_{ik} I_k + e_i, \text{ where} \qquad (5.12)\\ \mathrm{E}(e_i) &= 0,\ \mathrm{Cov}\,(e_i, I_k) = 0,\ \mathrm{Cov}(e_i, e_j) = 0 \text{ (unless } i = j\text{), and}\\ \mathrm{Cov}(I_k, I_h) &= 0 \text{ (unless } k = h\text{).} \end{aligned}$$

If this transformation of the indices is linear (and non-singular, so it is can be reversed if we want to), then the fit of the regression is unchanged.

5.4.2 "Rotating" the Indices

There are several ways of transforming the indices to make them uncorrelated, but the following regression approach is perhaps the simplest and may also give the best possibility of interpreting the results:

1. Let the first transformed index equal the original index, $I_1 = I_1^*$ (possibly demeaned). This would often be the market return.

2. Regress the second original index on the first transformed index, $I_2^* = \gamma_0 + \gamma_1 I_1 + \varepsilon_2$. Then, let the second transformed index be the fitted residual, $I_2 = \hat{\varepsilon}_2$.

3. Regress the third original index on the first two transformed indices, $I_3^* = \theta_0 + \theta_1 I_1 + \theta_2 I_2 + \varepsilon_3$. Then, let $I_3 = \hat{\varepsilon}_3$. Follow the same idea for all subsequent indices.

Recall that the fitted residual (from Least Squares) is always uncorrelated with the regressor (by construction). In this case, this means that I_2 is uncorrelated with I_1 (step 2) and that I_3 is uncorrelated with both I_1 and I_2 (step 3). The correlation matrix of the first three rotated indices is therefore

$$\mathrm{Corr}\left(\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \qquad (5.13)$$

This recursive approach also helps in interpreting the transformed indices. Suppose the first index is the market return and that the second original index is an interest rate. The first transformed index (I_1) is then clearly the market return. The second transformed index (I_2) can then be interpreted as the interest rate minus the interest rate expected at the current stock market return—that is, the part of the interest rate that cannot be explained by the stock market return.

More generally, let the kth index ($k = 1, 2, \ldots, K$) be

$$I_k = \hat{\varepsilon}_k, \text{ where } \hat{\varepsilon}_k \text{ is the fitted residual from the regression} \tag{5.14}$$

$$I_k^* = \delta_{k1} + \textstyle\sum_{s=1}^{k-1} \gamma_{ks} I_s + \varepsilon_k. \tag{5.15}$$

Notice that for the first index ($k = 1$), the regression is only $I_1^* = \delta_{11} + \varepsilon_1$, so I_1 equals the demeaned I_1^*.

5.4.3 Multi-Index Model after “Rotating” the Indices

To see why the transformed indices are very convenient for calculating the covariance matrix, consider a two-index model. Then, (5.12) implies that the variance of asset i is

$$\begin{aligned}\sigma_{ii} &= \operatorname{Var}(a_i + b_{i1} I_1 + b_{i2} I_2 + e_i) \\ &= b_{i1}^2 \operatorname{Var}(I_1) + b_{i2}^2 \operatorname{Var}(I_2) + \operatorname{Var}(e_i).\end{aligned} \tag{5.16}$$

Similarly, the covariance of assets i and j is

$$\begin{aligned}\sigma_{ij} &= \operatorname{Cov}\left(a_i + b_{i1} I_1 + b_{i2} I_2 + e_i, a_j + b_{j1} I_1 + b_{j2} I_2 + e_j\right) \\ &= b_{i1} b_{j1} \operatorname{Var}(I_1) + b_{i2} b_{j2} \operatorname{Var}(I_2).\end{aligned} \tag{5.17}$$

More generally, with n assets and K indices we can define b_1 to be an $n \times 1$ vector of the slope coefficients for the first index (b_{i1}, b_{j1}) and b_2 the vector of slope coefficients for the second index and so on. Also, let Σ to be an $n \times n$ matrix with the variances of the residuals along the diagonal. The covariance matrix of the returns is then

$$\operatorname{Cov}(R) = b_1 b_1' \operatorname{Var}(I_1) + b_2 b_2' \operatorname{Var}(I_2) + \ldots + b_K b_K' \operatorname{Var}(I_K) + \Sigma \tag{5.18}$$

$$= \textstyle\sum_{k=1}^{K} b_k b_k' \operatorname{Var}(I_k) + \Sigma. \tag{5.19}$$

See Figure 5.5 for an example.

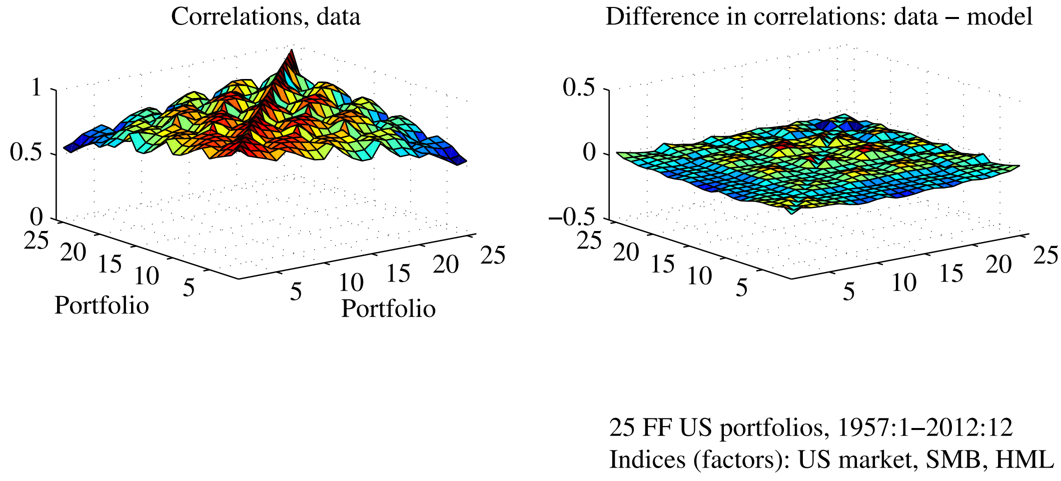


Figure 5.5: Correlations of US portfolios

5.4.4 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.

See Figures 5.6–5.7 for an illustration.

5.5 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

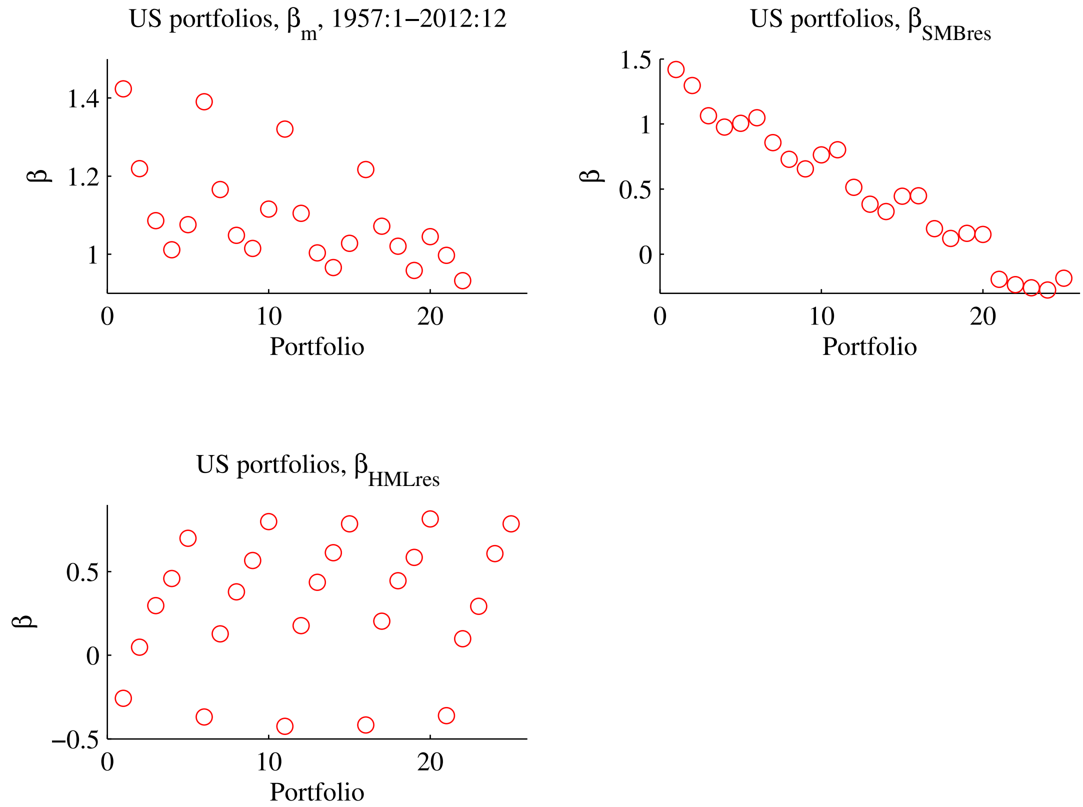


Figure 5.6: Loading (betas) of rotated factors

jth ($j \geq 2$) principal component (pc_{jt}) is similar (and its variance is denoted λ_j), except that is must be uncorrelated with all lower principal components. Remark 5.4 gives a a formal definition.

Remark 5.4 *(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 jth 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 jth largest eigenvalue λ_j (which equals $\text{Var}(pc_{jt}) = w_j' \Sigma w_j$).*

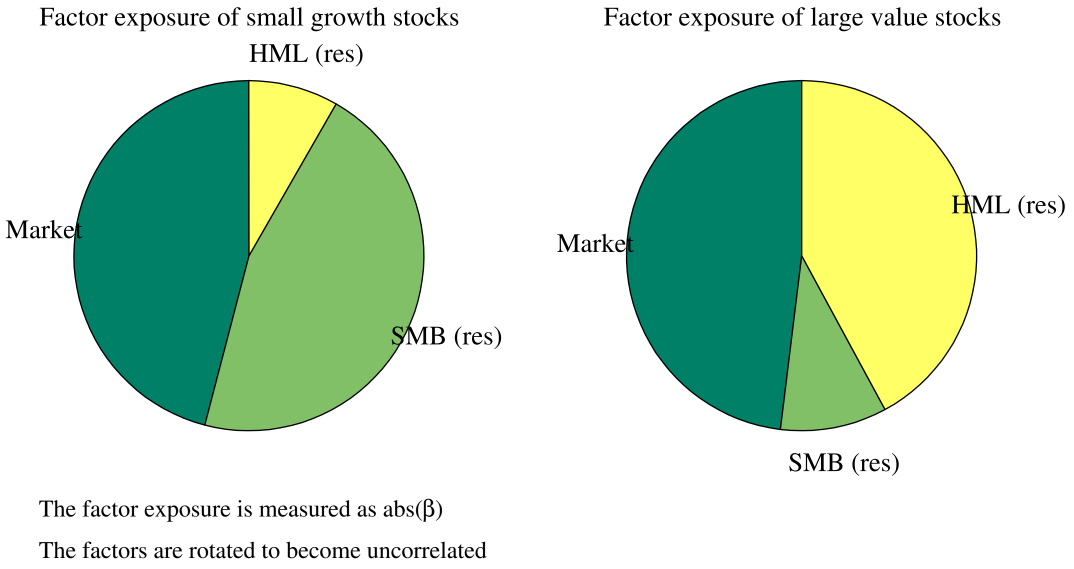


Figure 5.7: Absolute loading (betas) of rotated factors

Let the ith eigenvector be the ith column of the $n \times n$ matrix

$$W = [\ w_1 \quad \cdots \quad w_n\ ]. \tag{5.20}$$

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

$$pc_t = W' z_t. \tag{5.21}$$

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

$$z_t = W pc_t. \tag{5.22}$$

This shows that the ith eigenvector (the ith column of W) can be interpreted as the effect of the ith 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 $-pc_{jt}$).

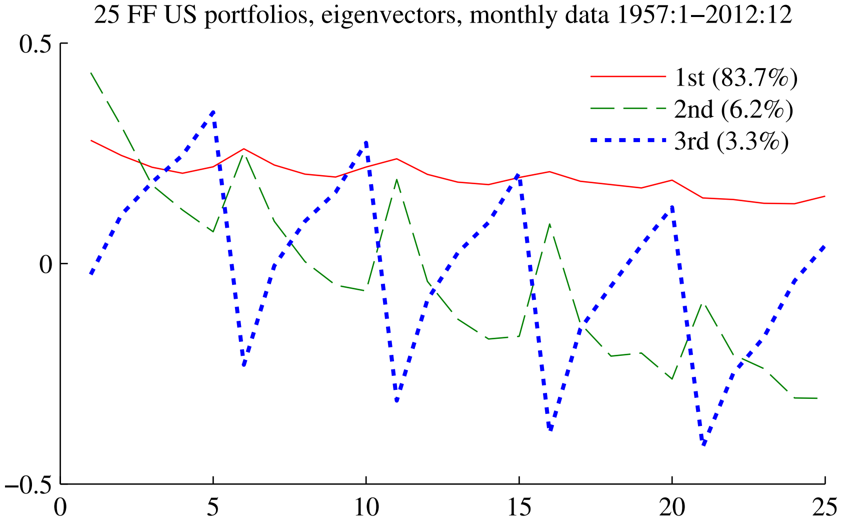


Figure 5.8: Eigenvectors for US portfolio returns

Example 5.5 *(PCA with 2 series) With two series we have*

$$pc_{1t} = \begin{bmatrix} w_{11} \\ w_{21} \end{bmatrix}' \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ and } pc_{2t} = \begin{bmatrix} w_{12} \\ w_{22} \end{bmatrix}' \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ or}$$

$$\begin{bmatrix} pc_{1t} \\ pc_{2t} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}' \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ and}$$

$$\begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} pc_{1t} \\ pc_{2t} \end{bmatrix}.$$

For instance, w_{12} shows how pc_{2t} affects z_{1t}, while w_{22} shows how pc_{2t} affects z_{2t}.

Remark 5.6 *(Data in matrices*) Transpose (5.21) 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 (5.22) 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 ($\text{Cov}(pc_{it}, pc_{jt}) = 0$)), we can think of the sum of their variances ($\Sigma_{i=1}^{n}\lambda_i$) 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

$$\text{relative importance of } pc_j = \lambda_j / \Sigma_{i=1}^{n} \lambda_i . \tag{5.23}$$

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.

5.6 Estimating Expected Returns

The starting point for forming estimates of future mean excess returns is typically historical excess returns. Excess returns are preferred to returns, since this avoids blurring the risk compensation (expected excess return) with long-run movements in inflation (and therefore interest rates). The expected excess return for the future period is typically formed as a judgmental adjustment of the historical excess return. Evidence suggest that the adjustments are hard to make.

It is typically hard to predict movements (around the mean) of asset returns, but a few variables seem to have some predictive power, for instance, the slope of the yield curve, the earnings/price yield, and the book value–market value ratio. Still, the predictive power is typically low.

Makridakis, Wheelwright, and Hyndman (1998) 10.1 show 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.)

Bibliography

Amemiya, T., 1985, *Advanced econometrics*, Harvard University Press, Cambridge, Massachusetts.

Blume, M. E., 1971, “On the Assessment of Risk,” *Journal of Finance*, 26, 1–10.

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.

Greene, W. H., 2003, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 5th edn.

Makridakis, S., S. C. Wheelwright, and R. J. Hyndman, 1998, *Forecasting: methods and applications*, Wiley, New York, 3rd edn.

6 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.

6.1 Market Model

Let $R^e_{it} = R_{it} - R_{ft}$ be the excess return on asset i in excess over the riskfree asset, and let R^e_{mt} be the excess return on the market portfolio. The basic implication of CAPM is that the expected excess return of an asset ($\mathrm{E}\, R^e_{it}$) is linearly related to the expected excess return on the market portfolio ($\mathrm{E}\, R^e_{mt}$) according to

$$\mathrm{E}\, R^e_{it} = \beta_i \,\mathrm{E}\, R^e_{mt}, \text{ where } \beta_i = \frac{\mathrm{Cov}\,(R_i, R_m)}{\mathrm{Var}\,(R_m)}. \tag{6.1}$$

Consider the regression

$$\begin{aligned} R^e_{it} &= \alpha_i + b_i R^e_{mt} + \varepsilon_{it}, \text{ where} \\ \mathrm{E}\, \varepsilon_{it} &= 0 \text{ and } \mathrm{Cov}(R^e_{mt}, \varepsilon_{it}) = 0. \end{aligned} \tag{6.2}$$

The two last conditions are automatically imposed by LS. Take expectations of the regression to get

$$\mathrm{E}\, R^e_{it} = \alpha_i + b_i \,\mathrm{E}\, R^e_{mt}. \tag{6.3}$$

Notice that the LS estimate of b_i is the sample analogue to β_i in (6.1). 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 (6.1) 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}. This is the perspective of CAPM tests.

The t-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

$$\frac{\hat{\alpha}_i}{\operatorname{Std}(\hat{\alpha}_i)} \overset{d}{\to} N(0,1) \text{ under } H_0 : \alpha_i = 0. \tag{6.4}$$

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 (in this respect, at least).

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). 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 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 6.1 shows some results for US industry portfolios.

6.1.1 Interpretation of the CAPM Test

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

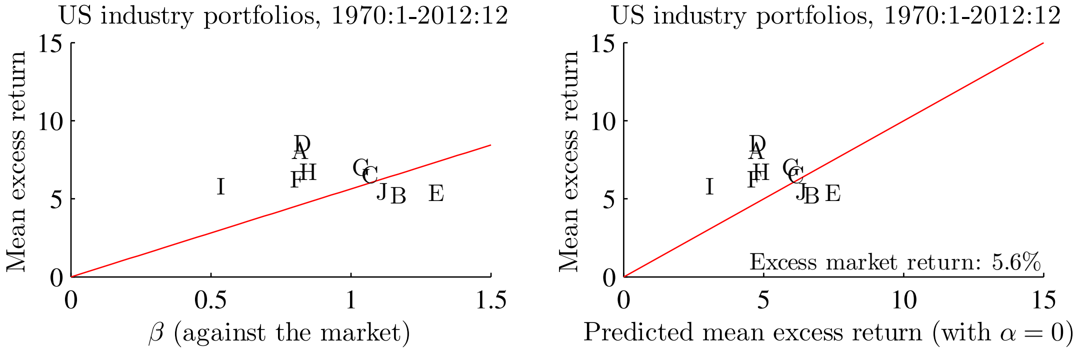
$$\frac{\hat{\alpha}_i^2}{\operatorname{Var}(\hat{\alpha}_i)} \overset{d}{\to} \chi_1^2 \text{ under } H_0\text{: } \alpha_i = 0. \tag{6.5}$$

Tables (A.2)–(A.1) list critical values for t- and chi-square tests

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 (6.5) can be written

$$\frac{\hat{\alpha}_i^2}{\operatorname{Var}(\hat{\alpha}_i)} = \frac{(SR_c)^2 - (SR_m)^2}{[1 + (SR_m)^2]/T}, \tag{6.6}$$

where SR_m is the Sharpe ratio of the market portfolio (as before) 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



	alpha	pval	StdErr
all	NaN	0.04	NaN
A (NoDur)	3.62	0.01	8.70
B (Durbl)	-1.21	0.55	13.66
C (Manuf)	0.70	0.48	6.37
D (Enrgy)	4.06	0.07	14.75
E (HiTec)	-1.82	0.32	11.93
F (Telcm)	1.82	0.29	11.10
G (Shops)	1.37	0.35	9.51
H (Hlth)	2.13	0.21	11.39
I (Utils)	2.87	0.11	11.64
J (Other)	-0.65	0.55	6.99

CAPM
Factor: US market
alpha and StdErr are in annualized %

Figure 6.1: CAPM regressions on US industry indices

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, $\hat{\alpha}_i^2/\operatorname{Var}(\hat{\alpha}_i)$, is zero—and CAPM is not rejected.

Proof. (*Proof of (6.6)) From the CAPM regression (6.2) we have

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

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 $\mu^{e\prime}\Sigma^{-1}\mu^e$, 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, $\mu^{e\prime}\Sigma^{-1}\mu^{e}$, (using both R_{it} and R_{mt}) is

$$\left(\frac{\mu^e_c}{\sigma_c}\right)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + \left(\frac{\mu^e_m}{\sigma_m}\right)^2,$$

which we can write as

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

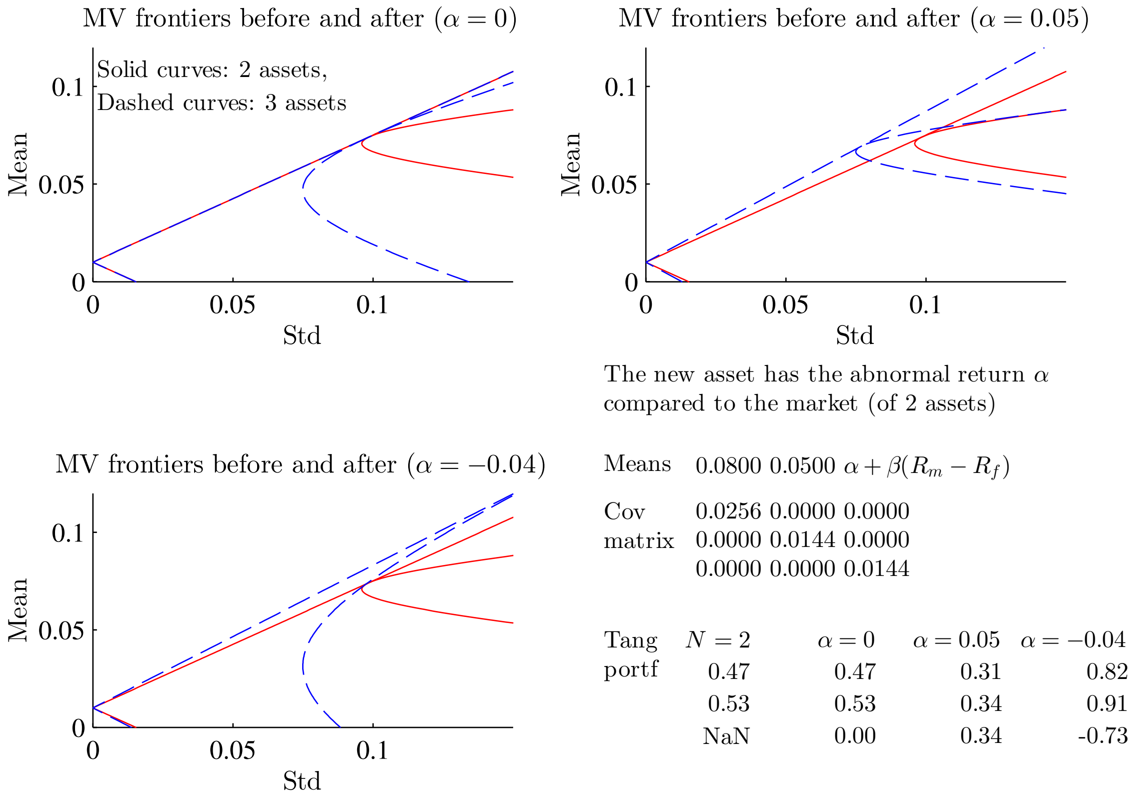
Combine this with (6.8) which shows that $\text{Var}(\hat{\alpha}_i) = [1 + (SR_m)^2]\,\text{Var}(\varepsilon_{it})/T$. ■

This is illustrated in Figure 6.2 which shows the effect of adding an asset to the investment opportunity set. In this case, the new asset has a zero beta (since it is uncorrelated with all original assets), but the same type of result holds for any new asset. The basic point is that the market model tests if the new assets moves 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 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.

6.1.2 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 actual small sample *size of the test* is thus larger than the asymptotic (or "nominal") size (see Campbell, Lo, and MacKinlay (1997) Table 5.1). The practical consequence is that we should either used adjusted critical values (from Monte Carlo or bootstrap simulations)—or more pragmatically, that we should only believe in strong rejections of the null hypothesis.

To study the power of the test (the frequency of rejections of a false null hypothesis) we have to specify an alternative data generating process (for instance, how much extra return in excess of that motivated by CAPM) and the size of the test (the critical value to use). Once that is done, 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



The new asset has the abnormal return α compared to the market (of 2 assets)

Means	0.0800	0.0500	$\alpha + \beta(R_m - R_f)$
Cov matrix	0.0256	0.0000	0.0000
	0.0000	0.0144	0.0000
	0.0000	0.0000	0.0144

Tang portf	$N = 2$	$\alpha = 0$	$\alpha = 0.05$	$\alpha = -0.04$
	0.47	0.47	0.31	0.82
	0.53	0.53	0.34	0.91
	NaN	0.00	0.34	-0.73

Figure 6.2: Effect on MV frontier of adding assets

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

$$\text{Var}(\hat{\alpha}_i) = \left[1 + \frac{(\mu_m^e)^2}{\text{Var}\left(R_m^e\right)}\right] \text{Var}(\varepsilon_{it})/T \tag{6.7}$$

$$= \left[1 + (SR_m)^2\right] \text{Var}(\varepsilon_{it})/T, \tag{6.8}$$

where 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 bit uncertainty about how risky asset i is then translates in a large uncertainty about what the risk-adjusted return should be.

Example 6.1 *Suppose we have monthly data with* $\widehat{\alpha}_i = 0.2\%$ *(that is,* $0.2\% \times 12 = 2.4\%$ *per year),* $\text{Std}(\varepsilon_{it}) = 3\%$ *(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 (6.4) of at least 1.65, so*

$$\frac{0.2}{\sqrt{1+0.15^2}3/\sqrt{T}} \geq 1.65 \text{ or } T \geq 626.$$

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 (6.8)) 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

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

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

$$\Sigma_{xx} = \text{E}\sum\nolimits_{t=1}^T x_t x_t'/T = \text{E}\frac{1}{T}\sum\nolimits_{t=1}^T \begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} = \begin{bmatrix} 1 & \text{E}\,f_t \\ \text{E}\,f_t & \text{E}\,f_t^2 \end{bmatrix}, \text{ so}$$

$$\sigma^2\Sigma_{xx}^{-1} = \frac{\sigma^2}{\text{E}\,f_t^2 - (\text{E}\,f_t)^2}\begin{bmatrix} \text{E}\,f_t^2 & -\text{E}\,f_t \\ -\text{E}\,f_t & 1 \end{bmatrix} = \frac{\sigma^2}{\text{Var}(f_t)}\begin{bmatrix} \text{Var}(f_t) + (\text{E}\,f_t)^2 & -\text{E}\,f_t \\ -\text{E}\,f_t & 1 \end{bmatrix}.$$

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

6.1.3 Several Assets

In most cases there are several (n) test assets, and we actually want to test if all the α_i (for $i = 1, 2, ..., n$) are zero. 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. 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 judgement on. For a more formal approach, see Section 6.1.4.

A quite different approach to study a cross-section of assets is to first perform a CAPM regression (6.2) and then the following cross-sectional regression

$$\sum_{t=1}^{T} R_{it}^{e}/T = \gamma + \lambda\hat{\beta}_i + u_i, \tag{6.9}$$

where $\sum_{t=1}^{T} R_{it}^{e}/T$ 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 contains measurement errors (it is only an uncertain estimate), which typically tend to bias the slope coefficient 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 (6.9) to have a high R^2—since it should be unity in a very large sample (if CAPM holds).

6.1.4 Several Assets: SURE Approach

This section outlines how we can set up a formal test of CAPM when there are several test assets.

For simplicity, suppose we have two test assets. Stack (6.2) for the two equations are

$$R_{1t}^{e} = \alpha_1 + b_1 R_{mt}^{e} + \varepsilon_{1t}, \tag{6.10}$$

$$R_{2t}^{e} = \alpha_2 + b_2 R_{mt}^{e} + \varepsilon_{2t} \tag{6.11}$$

where $\mathrm{E}\,\varepsilon_{it} = 0$ and $\mathrm{Cov}(R_{mt}^{e}, \varepsilon_{it}) = 0$. 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 (6.10) on a traditional form

$$R^e_{1t} = x_t'\beta_1 + \varepsilon_{1t}, \text{ where } x_t = \begin{bmatrix} 1 \\ R^e_{mt} \end{bmatrix}, \beta_1 = \begin{bmatrix} \alpha_1 \\ b_1 \end{bmatrix}, \tag{6.12}$$

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

Define

$$\hat{\Sigma}_{xx} = \sum\nolimits_{t=1}^{T} x_t x_t'/T, \text{ and } \hat{\sigma}_{ij} = \sum\nolimits_{t=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{jt}/T, \tag{6.13}$$

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

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \hat{\sigma}_{ij}\hat{\Sigma}_{xx}^{-1}/T. \tag{6.14}$$

(In many text books, this is written $\hat{\sigma}_{ij}(X'X)^{-1}$.)

The null hypothesis in our two-asset case is

$$H_0 : \alpha_1 = 0 \text{ and } \alpha_2 = 0. \tag{6.15}$$

In a large sample, the estimator is normally distributed (this follows from the fact that the LS estimator is a form of sample average, so we can apply a central limit theorem). Therefore, under the null hypothesis we have the following result. From (6.8) we know that the upper left element of Σ_{xx}^{-1}/T equals $[1 + (SR_m)^2]/T$. Then

$$\begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}[1 + (SR_m)^2]/T\right) \text{ (asymptotically).} \tag{6.16}$$

In practice we use the sample moments for the covariance matrix. Notice that the zero means in (6.16) 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. Notice that the zero means in (6.16) come from the null hypothesis: the distribution is (as usual) constructed by pretending that the null hypothesis is true.

We can now construct a chi-square test by using the following fact.

Remark 6.2 *If the* $n \times 1$ *vector* $v \sim N(0, \Omega)$*, then* $v'\Omega^{-1}v \sim \chi^2_n$*.*

To apply this, form the test static

$$T \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix}' [1 + (SR_m)^2]^{-1} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \sim \chi_2^2. \qquad (6.17)$$

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

6.1.5 Representative Results of the CAPM Test

One of the more interesting 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 ith size quintile *and* the jth BE/ME quintile:

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

Tables 6.1–6.2 summarize some basic properties of these portfolios.

	Book value/Market value				
	1	2	3	4	5
Size 1	3.3	9.2	9.6	11.7	13.2
2	5.4	8.4	10.5	10.8	12.0
3	5.7	8.9	8.8	10.3	12.0
4	6.8	6.7	8.6	9.7	9.6
5	5.2	5.8	6.1	5.9	7.3

Table 6.1: Mean excess returns (annualised %), US data 1957:1–2012:12. Size 1: smallest 20% of the stocks, Size 5: largest 20% of the stocks. B/M 1: the 20% of the stocks with the smallest ratio of book to market value (growth stocks). B/M 5: the 20% of the stocks with the highest ratio of book to market value (value stocks).

	Book value/Market value				
	1	2	3	4	5
Size 1	1.4	1.2	1.1	1.0	1.1
2	1.4	1.2	1.0	1.0	1.1
3	1.3	1.1	1.0	1.0	1.0
4	1.2	1.1	1.0	1.0	1.0
5	1.0	0.9	0.9	0.8	0.9

Table 6.2: Beta against the market portfolio, US data 1957:1–2012:12. Size 1: smallest 20% of the stocks, Size 5: largest 20% of the stocks. B/M 1: the 20% of the stocks with the smallest ratio of book to market value (growth stocks). B/M 5: the 20% of the stocks with the highest ratio of book to market value (value stocks).

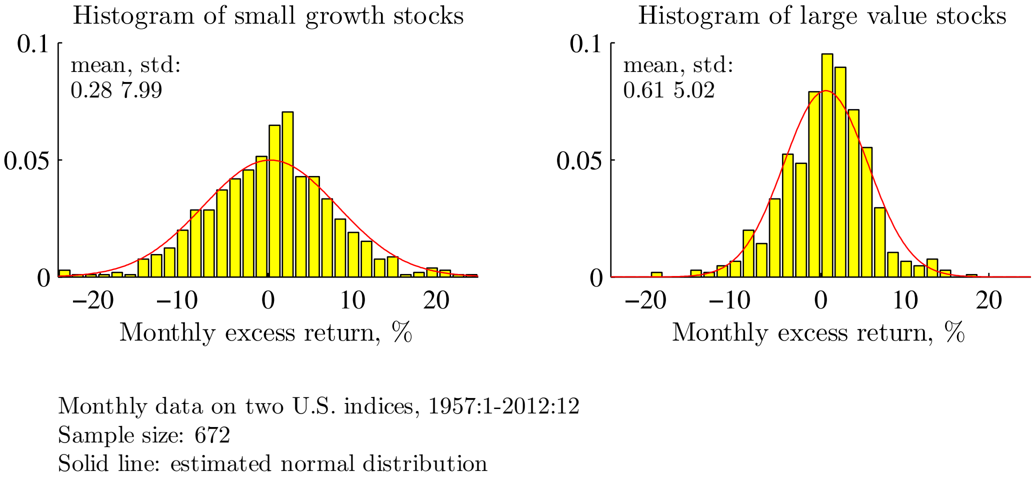


Figure 6.3: Comparison of small growth stock and large value stocks

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 $\mathrm{E}\, R^e_{it} = \beta_i \lambda$ where λ is the risk premium (excess return) on the market portfolio).

However, it is found that there is almost no relation between $\mathrm{E}\, R^e_{it}$ and β_i (there is a cloud in the $\beta_i \times \mathrm{E}\, R^e_{it}$ space, see Cochrane (2001) 20.2, Figure 20.9). 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 quantiles) between $\mathrm{E}\, R^e_{it}$ and β_i—as predicted by CAPM

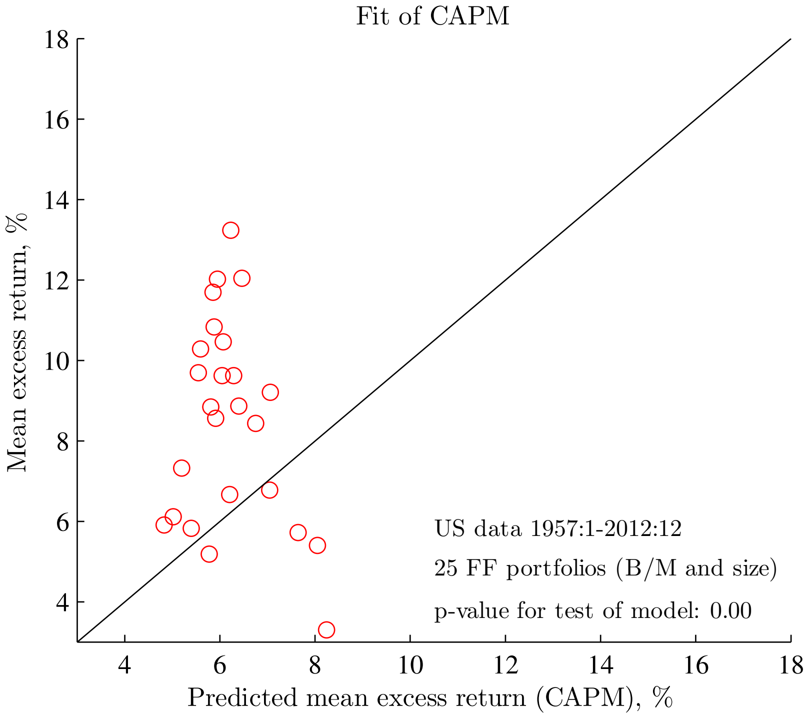


Figure 6.4: CAPM, FF portfolios

(see Cochrane (2001) 20.2, Figure 20.10). Second, *within a size quintile* there is a negative relation (across BE/ME quantiles) between $\mathrm{E}\,R^e_{it}$ and β_i—in stark contrast to CAPM (see Cochrane (2001) 20.2, Figure 20.11).

Figure 6.1 shows some results for US industry portfolios and Figures 6.4–6.6 for US size/book-to-market portfolios.

6.1.6 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).

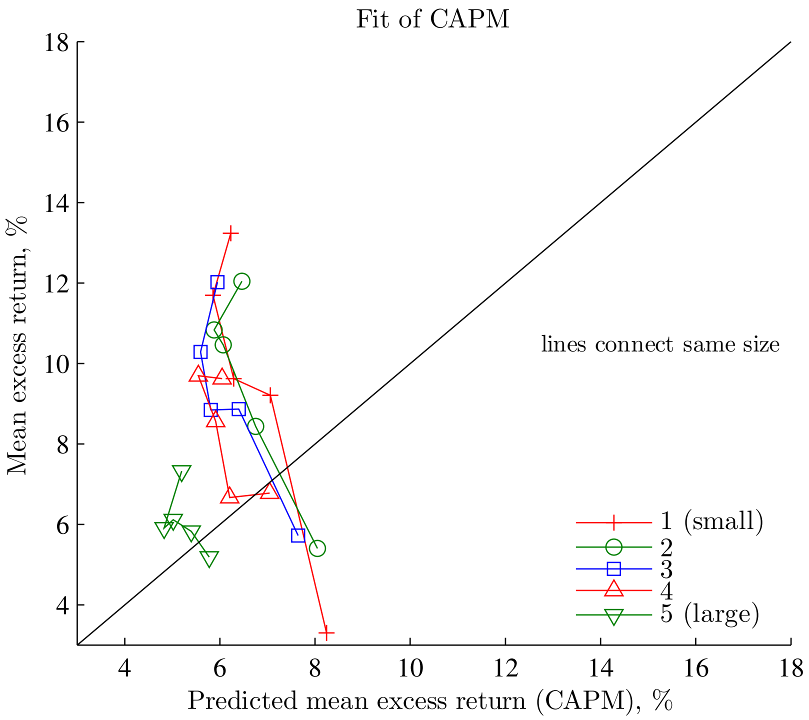


Figure 6.5: CAPM, FF portfolios

6.2 Calendar Time and Cross Sectional Regression*

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 ($\bar{R}^e_{jt}$ for group j)

$$\bar{R}^e_{jt} = \frac{1}{N_j}\textstyle\sum_{i \in \text{Group } j} R^e_{it}, \tag{6.18}$$

where N_j is the number of individuals in group j.

Then, we run a factor model

$$\bar{R}^e_{jt} = x_t'\beta_j + v_{jt}, \text{ for } j = 1, 2, \ldots, M \tag{6.19}$$

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

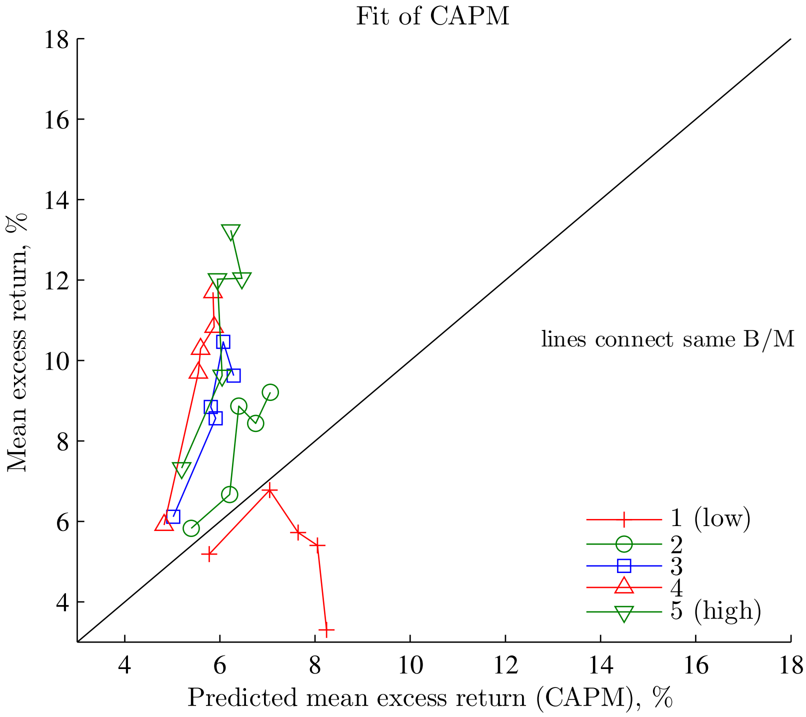


Figure 6.6: CAPM, FF portfolios

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 Mth group than for the for first group.

Example 6.3 *(CalTime with two investor groups) With two investor groups, estimate the following SURE system*

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

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.

The *cross sectional regression* (CrossReg) approach is to first estimate the factor model for each investor

$$R^e_{it} = x_t'\beta_i + \varepsilon_{it}, \text{ for } i = 1, 2, \ldots, N \tag{6.20}$$

and to then regress the (estimated) betas for the pth factor (for instance, the intercept) on the investor characteristics

$$\hat{\beta}_{pi} = z_i'c_p + w_{pi}. \tag{6.21}$$

In this second-stage regression, the investor characteristics z_i could be a dummy variable (for age roup, say) or a continuous variable (age, say). Notice that using a continuos investor characteristics assumes that the relation between the characteristics and the beta is linear—something that is not assumed in the CalTime approach. (This saves degrees of freedom, but may sometimes be a very strong assumption.) However, a potential problem with the CrossReg approach is that it is often important to account for the cross-sectional correlation of the residuals.

6.3 Several Factors

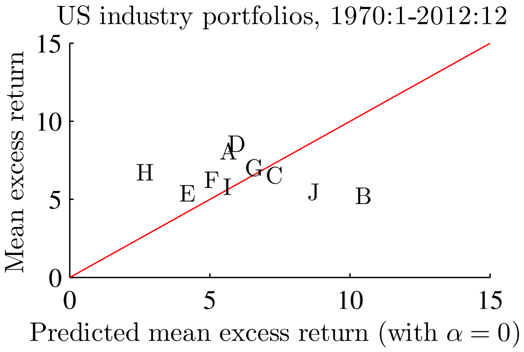
In multifactor models, (6.2) is still valid—provided we reinterpret b_i and R^e_{mt} as vectors, so $b_i R^e_{mt}$ stands for $b_{io}R^e_{ot} + b_{ip}R^e_{pt} + ...$

$$R^e_{it} = \alpha + b_{io}R^e_{ot} + b_{ip}R^e_{pt} + ... + \varepsilon_{it}. \tag{6.22}$$

In this case, (6.2) is a multiple regression, but the test (6.4) 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 6.4 *(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*



	alpha	pval	StdErr
all	NaN	0.00	NaN
A (NoDur)	2.70	0.04	8.47
B (Durbl)	-4.96	0.01	12.36
C (Manuf)	-0.46	0.62	6.08
D (Enrgy)	2.90	0.19	14.30
E (HiTec)	1.46	0.34	9.96
F (Telcm)	1.43	0.41	10.86
G (Shops)	0.76	0.60	9.40
H (Hlth)	4.35	0.01	10.66
I (Utils)	0.33	0.85	10.45
J (Other)	-3.00	0.00	5.86

Fama-French model
Factors: US market, SMB (size), and HML (book-to-market)
alpha and StdErr are in annualized %

Figure 6.7: Fama-French regressions on US industry indices

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 returns, you add the riskfree rate: $R_i - R_j + R_f$, implicitly assuming that your portfolio consists 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 6.7 shows some results for the Fama-French model on US industry portfolios and Figures 6.8–6.10 on the 25 Fama-French portfolios.

6.4 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.

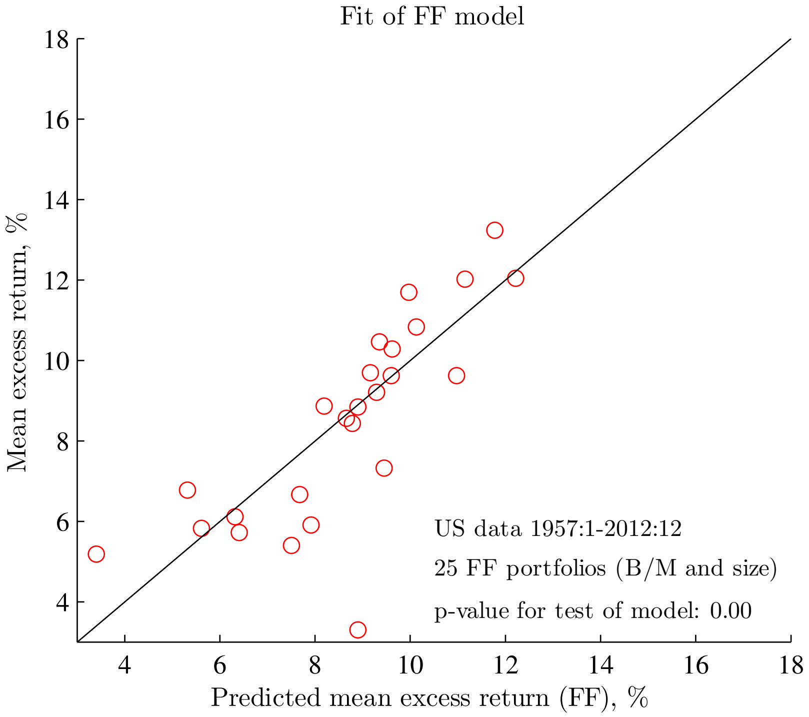


Figure 6.8: FF, FF portfolios

- First, estimate the betas β_i ($i = 1, \ldots, n$) from (6.2) (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 $i = 1, \ldots, n$) regression

$$R^e_{it} = \lambda_t' \hat{\beta}_i + \varepsilon_{it}, \tag{6.23}$$

where $\hat{\beta}_i$ are the regressors. (Note the difference to the traditional cross-sectional approach discussed in (6.9), where the second stage regression regressed $\mathrm{E}\, R^e_{it}$ on $\hat{\beta}_i$, while the Fama-French approach runs one regression for every time period.)

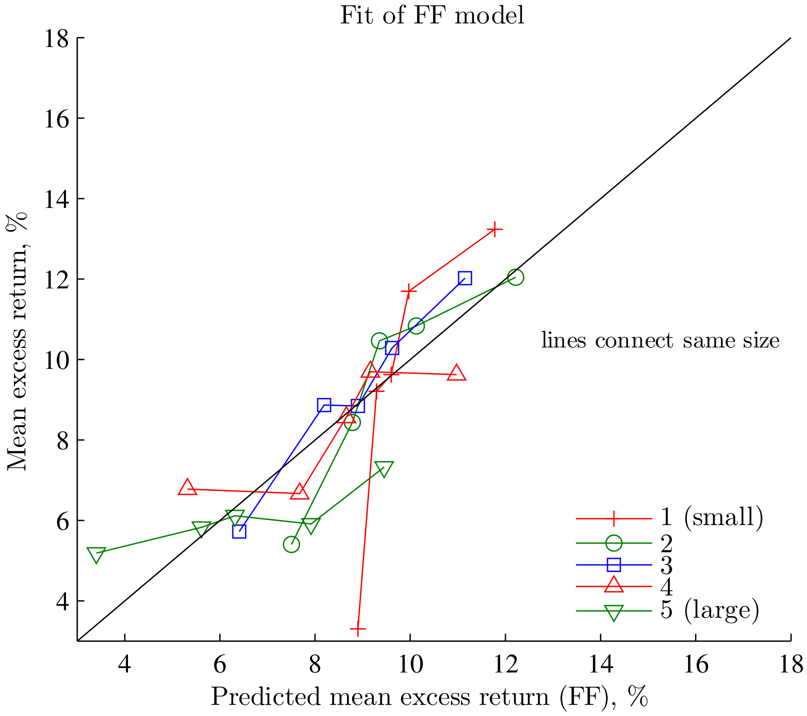


Figure 6.9: FF, FF portfolios

- Third, estimate the time averages

$$\hat{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it} \text{ for } i = 1, \dots, n, \text{ (for every asset)} \tag{6.24}$$

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\hat{\lambda}_t. \tag{6.25}$$

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-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

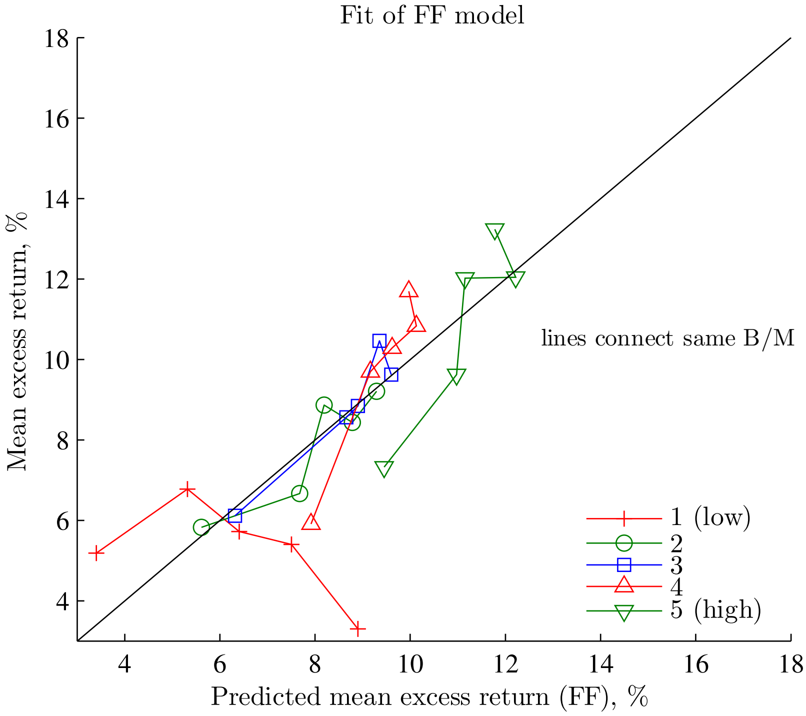


Figure 6.10: FF, FF portfolios

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 (6.23) 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 (6.24) that ε_i is the time average of the residual for asset i, ε_{it}), by forming a t-test $\hat{\varepsilon}_i /\operatorname{Std}(\hat{\varepsilon}_i)$. 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

$$\operatorname{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2 . \tag{6.26}$$

Since $\hat{\varepsilon}_i$ is the sample average of $\hat{\varepsilon}_{it}$, the variance of the former is the variance of the latter divided by T (the sample size)—provided $\hat{\varepsilon}_{it}$ is iid. That is,

$$\text{Var}(\hat{\varepsilon}_i) = \frac{1}{T}\,\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T^2}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{it} - \hat{\varepsilon}_i\right)^2. \tag{6.27}$$

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

$$\text{Var}(\hat{\lambda}) = \frac{1}{T^2}\sum_{t=1}^{T}(\hat{\lambda}_t - \hat{\lambda})^2. \tag{6.28}$$

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.

A Statistical Tables

n	Critical values		
	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 A.1: Critical values (two-sided test) of t distribution (different degrees of freedom) and normal distribution.

n	Critical values		
	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 A.2: Critical values of chisquare distribution (different degrees of freedom, n).

Bibliography

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The econometrics of financial markets*, Princeton University Press, Princeton, New Jersey.

Chen, N.-F., R. Roll, and S. A. Ross, 1986, "Economic forces and the stock market," *Journal of Business*, 59, 383–403.

Cochrane, J. H., 2001, *Asset pricing*, Princeton University Press, Princeton, New Jersey.

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.

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.

Gibbons, M., S. Ross, and J. Shanken, 1989, "A test of the efficiency of a given portfolio," *Econometrica*, 57, 1121–1152.

MacKinlay, C., 1995, “Multifactor models do not explain deviations from the CAPM,” *Journal of Financial Economics*, 38, 3–28.

Wooldridge, J. M., 2002, *Econometric analysis of cross section and panel data*, MIT Press.

7 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.

7.1 Descriptive Statistics

The sth *autocovariance* of y_t is estimated by

$$\widehat{\text{Cov}}\left(y_t, y_{t-s}\right) = \textstyle\sum_{t=1}^{T}\left(y_t - \bar{y}\right)\left(y_{t-s} - \bar{y}\right)/T, \text{ where } \bar{y} = \textstyle\sum_{t=1}^{T} y_t/T. \tag{7.1}$$

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

The sth *autocorrelation* is estimated as

$$\hat{\rho}_s = \frac{\widehat{\text{Cov}}\left(y_t, y_{t-s}\right)}{\widehat{\text{Std}}\left(y_t\right)^2}. \tag{7.2}$$

Compared with a traditional estimate of a correlation we here impose that the standard deviation of y_t and y_{t-p} 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 typically carry over to processes which are similar to the Gaussian—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

$$\sqrt{T}\begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \to^d N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right). \tag{7.3}$$

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,

$$\sqrt{T}\hat{\rho}_s \overset{d}{\to} N(0,1), \tag{7.4}$$

so $\sqrt{T}\hat{\rho}_s$ can be used as a t-stat.

Example 7.1 *(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 -1.65 and another 5% above 1.65, we can reject the null hypothesis at the 10% level if* $\sqrt{T}|\hat{\rho}_1| > 1.65$*. With* $T = 100$*, we therefore need* $|\hat{\rho}_1| > 1.65/\sqrt{100} = 0.165$ *for rejection, and with* $T = 1000$ *we need* $|\hat{\rho}_1| > 1.65/\sqrt{1000} \approx 0.052$*.*

The *Box-Pierce test* follows directly from the result in (7.3), since it shows that $\sqrt{T}\hat{\rho}_i$ and $\sqrt{T}\hat{\rho}_j$ are iid N(0,1) variables. Therefore, the sum of the square of them is distributed as a χ^2 variable. The test statistics typically used is

$$Q_L = T\sum_{s=1}^{L}\hat{\rho}_s^2 \to^d \chi_L^2. \tag{7.5}$$

Example 7.2 *(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 (7.5), 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).

The *pth partial autocorrelation* is discussed in Section 7.4.6.

7.2 Stationarity

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

$$\mathrm{E}\,y_t = \mu < \infty \tag{7.6}$$

$$\mathrm{Var}(y_t) = \gamma_0 < \infty \tag{7.7}$$

$$\mathrm{Cov}(y_t, y_{t-s}) = \gamma_s < \infty \tag{7.8}$$

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

$$\lim_{|s|\to\infty} \rho_s = 0 \text{ for any stationary series.} \tag{7.9}$$

7.3 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

$$\begin{aligned} \mathrm{E}\,\varepsilon_t &= 0 \\ \mathrm{Var}\,(\varepsilon_t) &= \sigma^2, \text{ and} \\ \mathrm{Cov}\,(\varepsilon_{t-s}, \varepsilon_t) &= 0 \text{ if } s \neq 0. \end{aligned} \tag{7.10}$$

If, in addition, ε_t is normally distributed, then it is said to be Gaussian white noise. This process can clearly not be forecasted.

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

$$y_t = \mu + \varepsilon_t, \tag{7.11}$$

where μ is a constant. This process is most easily estimated by estimating the sample mean and variance in the usual way (as in (7.1) with $p = 0$) or my OLS with a constant as the only regressor.

7.4 Autoregression (AR)

7.4.1 AR(1)

In this section we study the *first-order autoregressive* process, AR(1), in some detail in order to understand the basic concepts of autoregressive processes. The process is assumed to have a zero mean (or is demeaned, that an original variable minus its mean, for instance $y_t = x_t - \bar{x}_t$)—but it is straightforward to put in any mean or trend.

An AR(1) is

$$y_t = a y_{t-1} + \varepsilon_t, \text{ with } \mathrm{Var}(\varepsilon_t) = \sigma^2, \tag{7.12}$$

where ε_t is the white noise process in (7.10) which is uncorrelated with y_{t-1}. If $-1 <$

$a < 1$, then the effect of a shock eventually dies out: y_t is stationary.

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 $|a| < 1$)

$$\text{Var}\,(y_t) = \sigma^2/(1 - a^2) \tag{7.13}$$

$$\text{Corr}\,(y_t, y_{t-s}) = a^s, \tag{7.14}$$

so the variance and autocorrelation are increasing in a (assuming $a > 0$).

See Figure 7.1 for an illustration.

Remark 7.3 *(Autocorrelation and autoregression). Notice that the OLS estimate of a in (7.12) is essentially the same as the sample autocorrelation coefficient in (7.2). This follows from the fact that the slope coefficient is* $\widehat{\text{Cov}}\,(y_t, y_{t-1})\,/\widehat{\text{Var}}(y_{t-1})$. *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 7.4 *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. (Why?)*

If $a = 1$ in (7.12), 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 etc. 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.

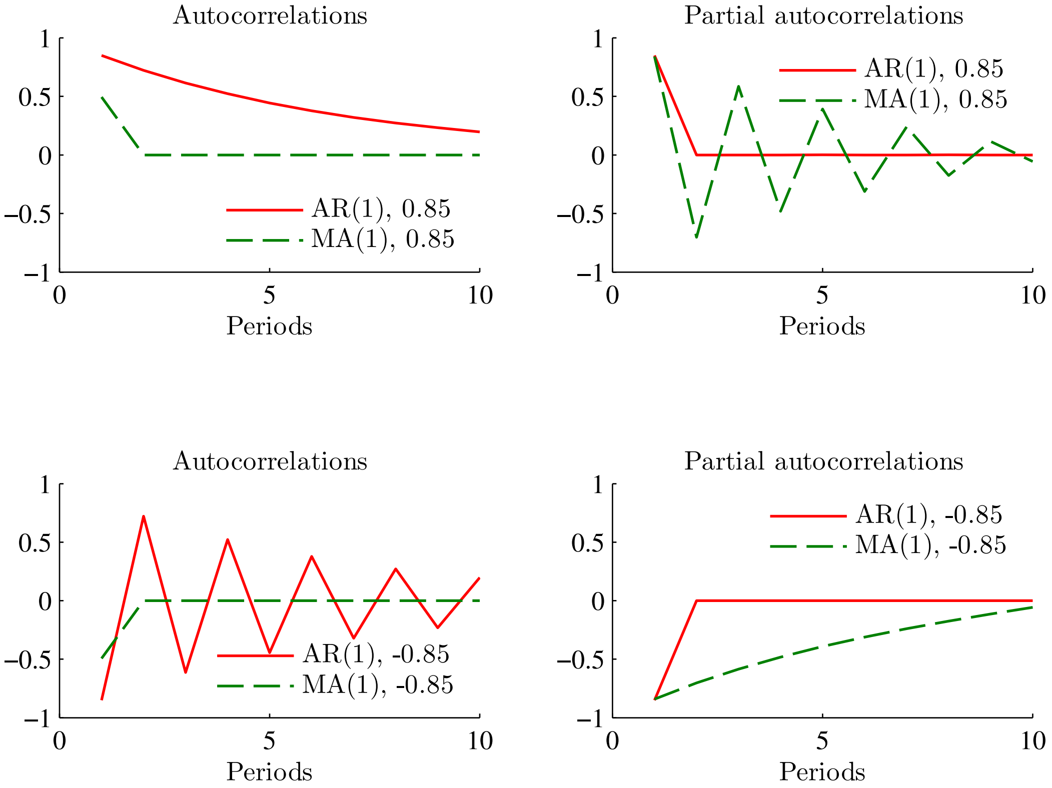


Figure 7.1: Autocorrelations and partial autocorrelations

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

Solve (7.12) backwards by repeated substitution

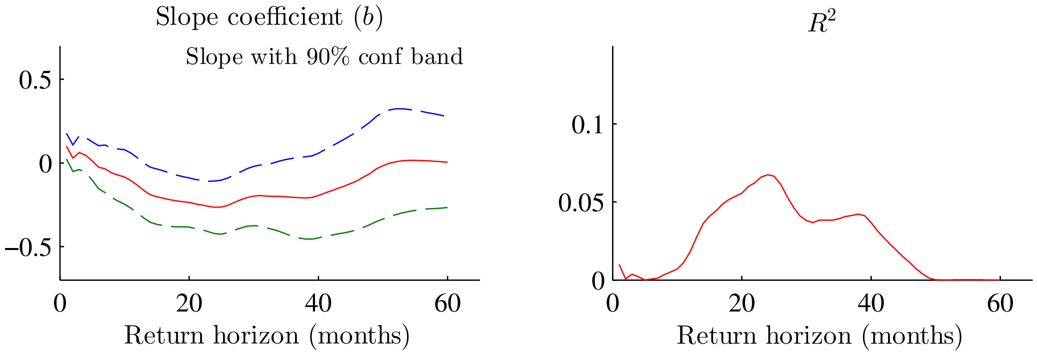
$$y_t = a\underbrace{(ay_{t-2} + \varepsilon_{t-1})}_{y_{t-1}} + \varepsilon_t \tag{7.15}$$

$$= a^2 y_{t-2} + a\varepsilon_{t-1} + \varepsilon_t \tag{7.16}$$

$$\vdots \tag{7.17}$$

$$= a^{K+1} y_{t-K-1} + \sum_{s=0}^{K} a^s \varepsilon_{t-s}. \tag{7.18}$$

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



Slope coefficient (b)

R^2

Monthly US stock returns 1957:1-2012:12

Regression: $r_t = a + br_{t-1} + \epsilon_t$

Figure 7.2: Predicting US stock returns (various investment horizons) with lagged returns.

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

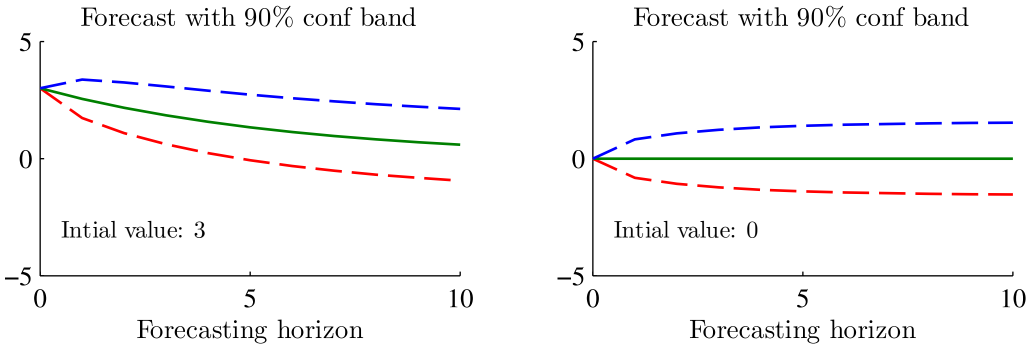
$$\begin{aligned} y_t &= \varepsilon_t + a\varepsilon_{t-1} + a^2\varepsilon_{t-2} + ... \\ &= \sum_{s=0}^{\infty} a^s \varepsilon_{t-s}. \end{aligned} \tag{7.19}$$

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 (7.12) as the sum of the variances of the two components on the right hand side

$$\begin{aligned} \text{Var}\,(y_t) &= \text{Var}\,(ay_{t-1}) + \text{Var}\,(\varepsilon_t) \\ &= a^2\,\text{Var}\,(y_{t-1}) + \text{Var}\,(\varepsilon_t) \\ &= \text{Var}\,(\varepsilon_t)\,/(1-a^2),\ \text{since}\ \text{Var}\,(y_{t-1}) = \text{Var}\,(y_t)\,. \end{aligned} \tag{7.20}$$

In this calculation, we use the fact that Var$\,(y_{t-1})$ and Var$\,(y_t)$ are equal. Formally, this follows from that they are both linear functions of current and past ε_s terms (see (7.19)), which have the same variance over time (ε_t is assumed to be white noise).

Note from (7.20) that the variance of y_t is increasing in the absolute value of a, which is illustrated in Figure 7.3. The intuition is that a large $|a|$ implies that a shock have effect over many time periods and thereby create movements (volatility) in y.



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

Figure 7.3: Properties of AR(1) process

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

$$\begin{aligned}\text{Cov}\left(y_t, y_{t-1}\right) &= \text{Cov}\left(a y_{t-1} + \varepsilon_t, y_{t-1}\right)\\ &= a\,\text{Cov}\left(y_{t-1}, y_{t-1}\right)\\ &= a\,\text{Var}\left(y_t\right). \qquad (7.21)\end{aligned}$$

We can then calculate the first-order autocorrelation as

$$\begin{aligned}\text{Corr}\left(y_t, y_{t-1}\right) &= \frac{\text{Cov}\left(y_t, y_{t-1}\right)}{\text{Std}(y_t)\,\text{Std}(y_{t-1})}\\ &= a. \qquad (7.22)\end{aligned}$$

It is straightforward to show that

$$\text{Corr}\left(y_t, y_{t-s}\right) = \text{Corr}\left(y_{t+s}, y_t\right) = a^s. \qquad (7.23)$$

7.4.3 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 long sample.

We want to *forecast y_{t+1} using information available in t*. From (7.12) we get

$$y_{t+1} = ay_t + \varepsilon_{t+1}. \tag{7.24}$$

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

$$\mathrm{E}_t\, y_{t+1} = ay_t, \text{ and} \tag{7.25}$$

$$y_{t+1} - \mathrm{E}_t\, y_{t+1} = \varepsilon_{t+1} \text{ with variance } \sigma^2. \tag{7.26}$$

We may also want to forecast y_{t+2} using the information in t. To do that note that (7.12) gives

$$\begin{aligned} y_{t+2} &= ay_{t+1} + \varepsilon_{t+2} \\ &= a\underbrace{(ay_t + \varepsilon_{t+1})}_{y_{t+1}} + \varepsilon_{t+2} \\ &= a^2 y_t + a\varepsilon_{t+1} + \varepsilon_{t+2}. \end{aligned} \tag{7.27}$$

Since the $\mathrm{E}_t\, \varepsilon_{t+1}$ and $\mathrm{E}_t\, \varepsilon_{t+2}$ are both zero, we get that

$$\mathrm{E}_t\, y_{t+2} = a^2 y_t, \text{ and} \tag{7.28}$$

$$y_{t+2} - \mathrm{E}_t\, y_{t+2} = a\varepsilon_{t+1} + \varepsilon_{t+2} \text{ with variance } a^2\sigma^2 + \sigma^2. \tag{7.29}$$

More generally, we have

$$\mathrm{E}_t\, y_{t+s} = a^s y_t, \tag{7.30}$$

$$\mathrm{Var}\,(y_{t+s} - \mathrm{E}_t\, y_{t+s}) = \left(1 + a^2 + a^4 + \ldots + a^{2(s-1)}\right)\sigma^2 \tag{7.31}$$

$$= \frac{a^{2s} - 1}{a^2 - 1}\sigma^2. \tag{7.32}$$

Example 7.5 *If $y_t = 3, a = 0.85$ and $\sigma = 0.5$, then the forecasts and the forecast error variances become*

Horizon s	$\mathrm{E}_t\, y_{t+s}$	$\mathrm{Var}\,(y_{t+s} - \mathrm{E}_t\, y_{t+s})$
1	$0.85 \times 3 = 2.55$	0.25
2	$0.85^2 \times 3 = 2.17$	$\left(0.85^2 + 1\right) \times 0.5^2 = 0.43$
25	$0.85^{25} \times 3 = 0.05$	$\frac{0.85^{50}-1}{0.85^2-1} \times 0.5^2 = 0.90$

Notice that the point forecast converge towards zero and the variance of the forecast error variance to the unconditional variance (see Example 7.4).

If the shocks ε_t, are normally distributed, then we can calculate 90% confidence intervals around the point forecasts in (7.25) and (7.28) as

$$\text{90\% confidence band of } \mathrm{E}_t\, y_{t+1} : a y_t \pm 1.65 \times \sigma \tag{7.33}$$

$$\text{90\% confidence band of } \mathrm{E}_t\, y_{t+2} : a^2 y_t \pm 1.65 \times \sqrt{a^2\sigma^2 + \sigma^2}. \tag{7.34}$$

(Recall that 90% of the probability mass is within the interval -1.65 to 1.65 in the N(0,1) distribution). To get 95% confidence bands, replace 1.65 by 1.96. Figure 7.3 gives an example.

Example 7.6 *Continuing Example 7.5, we get the following 90% confidence bands*

Horizon s	
1	$2.55 \pm 1.65 \times \sqrt{0.25} \approx [1.7, 3.4]$
2	$2.17 \pm 1.65 \times \sqrt{0.43} \approx [1.1, 3.2]$
25	$0.05 \pm 1.65 \times \sqrt{0.90} \approx [-1.5, 1.6]$

.

Remark 7.7 *(White noise as special case of AR(1).) When $a = 0$ in (7.12), then the AR(1) collapses to a white noise process. The forecast is then a constant (zero) for all forecasting horizons, see (7.30), and the forecast error variance is also the same for all horizons, see (7.32).*

7.4.4 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 $\bar{x}_t$ and then let the y_t in the AR(1) model (7.12) be a demeaned variable $y_t = x_t - \bar{x}_t$.

To include a constant μ in the theoretical expressions, we just need to substitute $x_t - \mu$ for y_t everywhere. For instance, in (7.12) we would get

$$\begin{aligned} x_t - \mu &= a\,(x_{t-1} - \mu) + \varepsilon_t \text{ or} \\ x_t &= (1-a)\,\mu + a x_{t-1} + \varepsilon_t. \end{aligned} \tag{7.35}$$

Estimation by LS will therefore give an intercept that equals $(1-a)\,\mu$ and a slope coefficient that equals a.

7.4.5 AR(p)

The *pth-order autoregressive* process, AR(p), is a straightforward extension of the AR(1)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ...a_p y_{t-p} + \varepsilon_t. \tag{7.36}$$

All the previous calculations can be made on this process as well—it is just a bit messier. This process can also be estimated with OLS since ε_t is uncorrelated with lags of y_t. Adding a constant is straightforward by substituting $x_t - \mu$ for y_t everywhere

7.4.6 Partial Autocorrelations

The pth partial autocorrelation tries to measure the direct relation between y_t and y_{t-p}, where the indirect effects of $y_{t-1}, ..., y_{t-p+1}$ are eliminated. 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. 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)

$$y_t = a y_{t-1} + \varepsilon_t. \tag{7.37}$$

The second partial autocorrelation is estimated by the second slope coefficient (a_2) in an AR(2)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t, \tag{7.38}$$

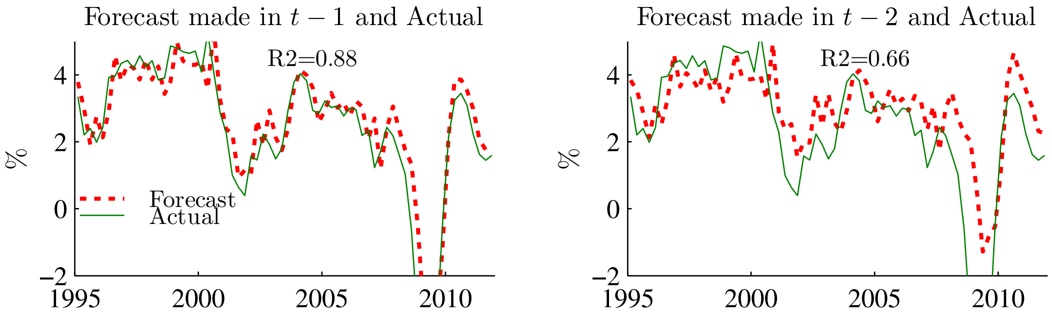
and so forth. The general pattern is that the pth partial autocorrelation is estimated by the slope coefficient of the pth lag in an AR(p), where we let p go from 1,2,3...

See Figure 7.1 for an illustration.

7.4.7 Forecasting with an AR(2)*

As an example, consider making a forecast of y_{t+1} based on the information in t by using an AR(2)

$$y_{t+1} = a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1}. \tag{7.39}$$



Model: AR(2) of US 4-quarter GDP growth
Estimated on data for 1947-1994

R^2 is corr(forecast,actual)^2 for 1995–
y_t and $\mathrm{E}_{t-s} y_t$ are plotted in t

Comparison of forecast errors from
autoregression (AR) and random walk (RW):

	1-quarter	2-quarter
MSE(AR)/MSE(RW)	0.70	0.72
MAE(AR)/MAE(RW)	0.95	0.98
R^2(AR)/R^2(RW)	1.07	1.27

Figure 7.4: Forecasting with an AR(2)

This immediately gives the one-period point forecast

$$\mathrm{E}_t\, y_{t+1} = a_1 y_t + a_2 y_{t-1}. \tag{7.40}$$

We can use (7.39) to write y_{t+2} as

$$\begin{aligned} y_{t+2} &= a_1 y_{t+1} + a_2 y_t + \varepsilon_{t+2} \\ &= a_1 \underbrace{(a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1})}_{y_{t+1}} + a_2 y_t + \varepsilon_{t+2} \\ &= (a_1^2 + a_2) y_t + a_1 a_2 y_{t-1} + a_1 \varepsilon_{t+1} + \varepsilon_{t+2}. \end{aligned} \tag{7.41}$$

Figure 7.4 gives an empirical example.

The expressions for the forecasts and forecast error variances quickly get somewhat messy—and even more so with an AR of higher order than two. There is a simple, and approximately correct, shortcut that can be taken. Note that 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

$$y_{t+1} = a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1} \tag{7.42}$$

$$y_{t+2} = b_1 y_t + b_2 y_{t-1} + v_{t+2}. \tag{7.43}$$

Clearly, (7.42) is the same as (7.39) and the estimated coefficients can therefore be used to make one-period forecasts, and the variance of ε_{t+1} is a good estimator of the variance of the one-period forecast error. The coefficients in (7.43) will be very similar to what we get by combining the a_1 and a_2 coefficients as in (7.41): b_1 will be similar to $a_1^2 + a_2$ and b_2 to $a_1 a_2$ (in an infinite sample they should be identical). Equation (7.43) can therefore be used to make two-period forecasts, and the variance of v_{t+2} can be taken to be the forecast error variance for this forecast.

7.5 Moving Average (MA)

A q^{th}*-order moving average process* is

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}, \tag{7.44}$$

where the innovation ε_t is white noise (usually Gaussian). It is straightforward to add a constant to capture a non-zero mean.

Estimation of MA processes is typically done by setting up the likelihood function and then using some numerical method to maximize it; 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.

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 7.1 for an illustration.

7.6 ARMA(p,q)

When the autocorrelations and partial autocorrelations look a bit like both an AR and an MA model, then a combination (ARMA) might be appropriate.

Autoregressive-moving average models add a moving average structure to an AR model. For instance, an ARMA(2,1) could be

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{7.45}$$

where ε_t is white noise. This type of model is much harder to estimate than the autoregressive model (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.

It is straightforward to add a constant to capture a non-zero mean.

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.

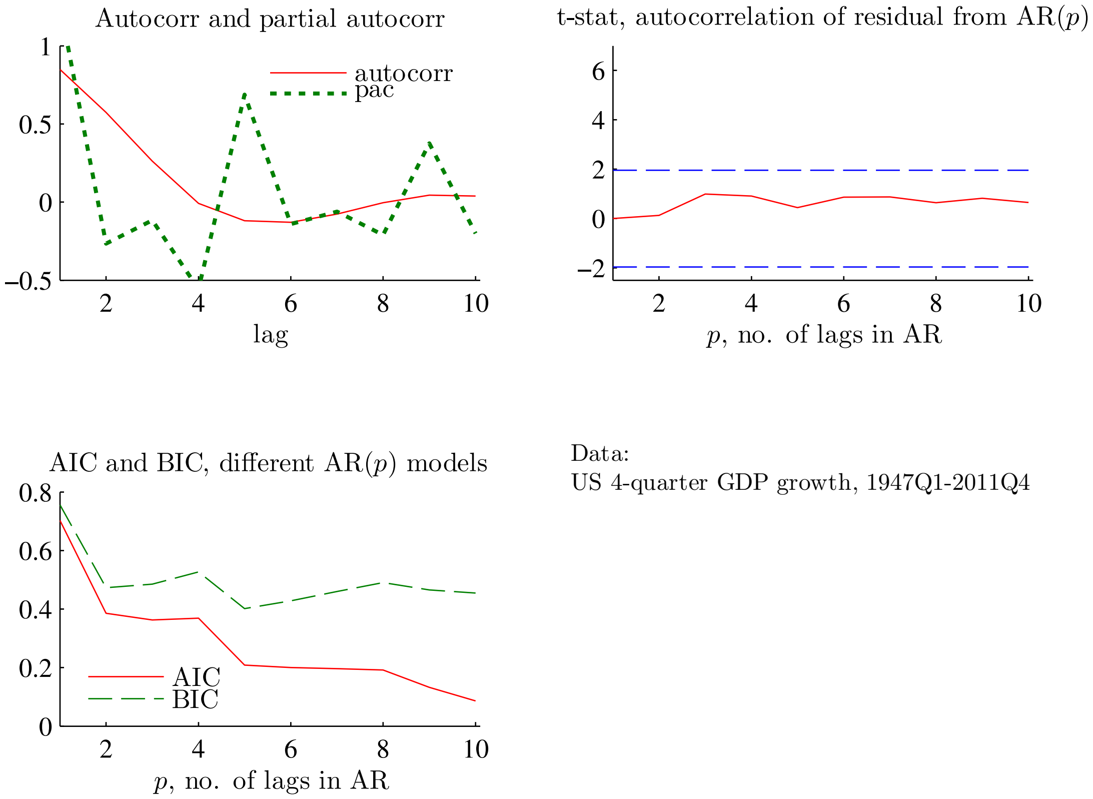
Remark 7.8 *In an ARMA model, both the autocorrelations and partial autocorrelations decay to zero for long lags.*

To choose a model, study the ACF and PACF—and check that residual are close to white noise (or at least not autocorrelated). To avoid overfitting, "punish" models with to many parameters. Akaike's Information Criterion (AIC) and the Bayesian information criterion (BIC) are

$$AIC = \ln \hat{\sigma}^2 + 2\frac{p+q+1}{T} \tag{7.46}$$

$$BIC = \ln \hat{\sigma}^2 + \frac{p+q+1}{T} \ln T. \tag{7.47}$$

trade-off between fit (low $\hat{\sigma}^2$) and number of parameters ($p+q$). Choose the model with the lowest AIC or BIC. (AIC often exaggerates the length)



Data:
US 4-quarter GDP growth, 1947Q1-2011Q4

Figure 7.5: Example of choosing lag length in an AR model

7.7 VAR(p)

The vector autoregression is a multivariate version of an AR(1) process: we can think of y_t and ε_t in (7.36) as vectors and the a_i as matrices.

For instance the VAR(1) of two variables (x_t and z_t) is (in matrix form)

$$\begin{bmatrix} x_{t+1} \\ z_{t+1} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_t \\ z_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{xt+1} \\ \varepsilon_{zt+1} \end{bmatrix}, \tag{7.48}$$

or equivalently

$$x_{t+1} = a_{11}x_t + a_{12}z_t + \varepsilon_{xt+1}, \text{ and} \tag{7.49}$$

$$z_{t+1} = a_{21}x_t + a_{22}z_t + \varepsilon_{zt+1}. \tag{7.50}$$

Both (7.49) and (7.50) are regression equations, which can be estimated with OLS (since ε_{xt+1} and ε_{zt+1} are uncorrelated with x_t and z_t).

With the information available in t, that is, information about x_t and z_t, (7.49) and (7.50) can be used to forecast one step ahead as

$$\mathrm{E}_t\, x_{t+1} = a_{11}x_t + a_{12}z_t \tag{7.51}$$

$$\mathrm{E}_t\, z_{t+1} = a_{21}x_t + a_{22}z_t. \tag{7.52}$$

We also want to make a forecast of x_{t+2} based on the information in t. Clearly, it must be the case that

$$\mathrm{E}_t\, x_{t+2} = a_{11}\,\mathrm{E}_t\, x_{t+1} + a_{12}\,\mathrm{E}_t\, z_{t+1} \tag{7.53}$$

$$\mathrm{E}_t\, z_{t+2} = a_{21}\,\mathrm{E}_t\, x_{t+1} + a_{22}\,\mathrm{E}_t\, z_{t+1}. \tag{7.54}$$

We already have values for $\mathrm{E}_t\, x_{t+1}$ and $\mathrm{E}_t\, z_{t+1}$ from (7.51) and (7.52) which we can use. For instance, for $\mathrm{E}_t\, x_{t+2}$ we get

$$\begin{aligned}\mathrm{E}_t\, x_{t+2} &= a_{11}\underbrace{(a_{11}x_t + a_{12}z_t)}_{\mathrm{E}_t\, x_{t+1}} + a_{12}\underbrace{(a_{21}x_t + a_{22}z_t)}_{\mathrm{E}_t\, z_{t+1}}\\ &= \left(a_{11}^2 + a_{12}a_{21}\right)x_t + \left(a_{12}a_{22} + a_{11}a_{12}\right)z_t. \end{aligned}\tag{7.55}$$

This has the same form as the one-period forecast in (7.51), 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 (7.48) 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.

The forecast uncertainty about the one-period forecast is simple: the forecast error $x_{t+1} - \mathrm{E}_t\, x_{t+1} = \varepsilon_{xt+1}$. The two-period forecast error, $x_{t+2} - \mathrm{E}_t\, x_{t+2}$, 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 messy 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 (7.48) and that we want to forecast x one and two periods ahead. From (7.51) and (7.55) we see that the regression equations should be of the form

$$x_{t+1} = \delta_1 x_t + \delta_2 z_t + u_{t+1}, \text{ and} \tag{7.56}$$

$$x_{t+2} = \gamma_1 x_t + \gamma_2 z_t + w_{t+s}. \tag{7.57}$$

With estimated coefficients (OLS can be used), it is straightforward to calculate forecasts and forecast error variances.

In a more general VAR(p) model we need to include p lags of both x and z in the regression equations ($p = 1$ in (7.56) and (7.57)).

7.7.1 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 economic causality (Christmas cards may Granger cause Christmas). In (7.56) z does Granger cause x if $\delta_2 \neq 0$, which can be tested with an F-test. 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.

7.8 Impulse Response Function

Any stationary process can be rewritten on ("inverted to") MA form

Example: AR(1)→MA(∞)

$$\begin{aligned} y_t &= \theta y_{t-1} + \varepsilon_t \\ &= \theta \underbrace{(\theta y_{t-2} + \varepsilon_{t-1})}_{y_{t-1}} + \varepsilon_t = \theta^2 y_{t-2} + \theta \varepsilon_{t-1} + \varepsilon_t \\ &\vdots \\ &= \varepsilon_t + \theta \varepsilon_{t-1} + \theta^2 \varepsilon_{t-2} + \ldots \end{aligned} \tag{7.58}$$

The MA form can be interpreted as giving the *impulse response* (the dynamic response to a shock in period t). Set all $\varepsilon_s = 0$, except $\varepsilon_t = 1$. From (7.44) we have

$$\begin{aligned} y_t &= 1 \\ y_{t+1} &= \theta_1 \\ y_{t+2} &= \theta_2, \text{ etc} \end{aligned} \tag{7.59}$$

See Figure 7.6 for an illustration.

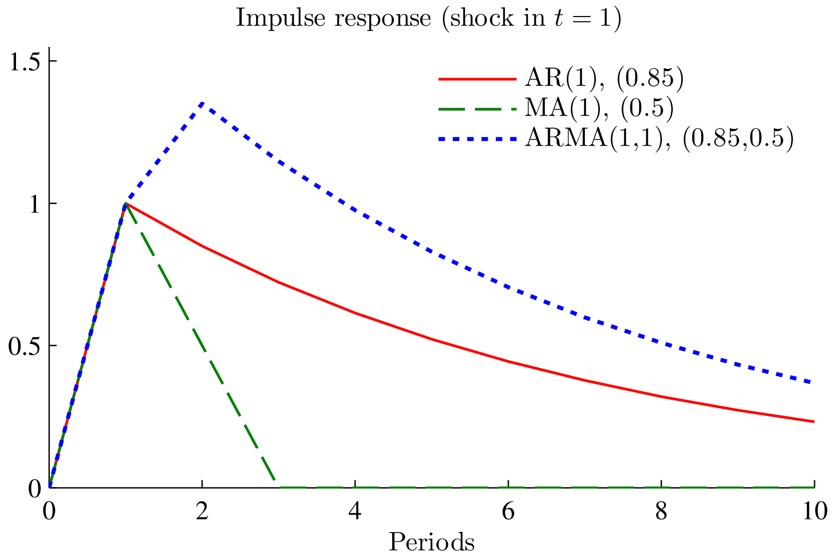


Figure 7.6: Impulse responses

7.9 Non-stationary Processes

7.9.1 Introduction

A *trend-stationary process* can be made stationary by subtracting a linear trend. The simplest example is

$$y_t = \mu + \beta t + \varepsilon_t \tag{7.60}$$

where ε_t is white noise.

A *unit root* process can be made stationary only by taking a difference. The simplest example is the *random walk* with drift

$$y_t = \mu + y_{t-1} + \varepsilon_t, \tag{7.61}$$

where ε_t is white noise. The name "unit root process" comes from the fact that the largest eigenvalues of the canonical form (the VAR(1) form of the AR(p)) is one. Such a process is said to be integrated of order one (often denoted I(1)) and can be made stationary by taking first differences. (So the first difference is an I(0) series.)

Example 7.9 *(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*. This is most easily seen for the random walk. Substitute repeatedly in (7.61) to get

$$\begin{aligned} y_t &= \mu + (\mu + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &\vdots \\ &= t\mu + y_0 + \sum_{s=1}^{t} \varepsilon_s. \end{aligned} \tag{7.62}$$

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 7.7 for an illustration.

A consequence of the permanent effect of a shock is that the variance of the conditional distribution grows without bound as the forecasting horizon is extended. For instance, for the random walk with drift, (7.62), 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.

In contrast, the conditional distribution from the trend stationary model, (7.60), is $N(st, \sigma^2)$.

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.

Example 7.10 *(Two unit roots.) Suppose y_t in Example (7.9) is actually the first differ-*

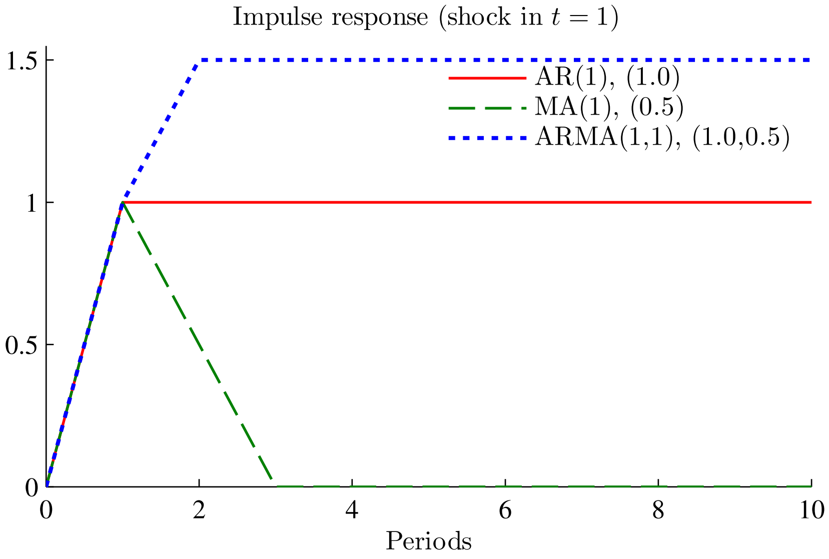


Figure 7.7: Impulse responses

ence 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 7.11 *(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 7.12 *(Lag operator*) A common and convenient way of dealing with leads and* *lags is the* lag operator, L. *It is such that*

$$L^s y_t = y_{t-s}$$

For instance, the AR(1) model

$$\begin{aligned} y_t &= \theta \underbrace{y_{t-1}}_{Ly_t} + \varepsilon_t, \textit{ or} \\ (1-\theta L)\, y_t &= \varepsilon_t, \textit{ or} \\ \theta(L) y_t &= \varepsilon_t, \end{aligned}$$

where $\theta(L) = (1-\theta L)$ *is a lag polynomial. Similarly, an ARMA(2,1) can be written*

$$\begin{aligned} y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} &= \varepsilon_t + \alpha_1 \varepsilon_{t-1} \\ \left(1 - \theta_1 L - \theta_2 L^2\right) y_t &= (1 + \alpha_1 L)\, \varepsilon_t. \end{aligned}$$

7.9.2 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 a trend. The same holds for unit root processes, even if they have no deterministic trends. However, 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 7.8 for an empirical example and Figures 7.9–7.11 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 first difference. For instance, a first difference of the random walk with drift is

$$\begin{aligned} \Delta y_t &= y_t - y_{t-1} \\ &= \mu + \varepsilon_t, \end{aligned} \tag{7.63}$$

which is white noise (any finite difference, like $y_t - y_{t-s}$, will give a stationary series),

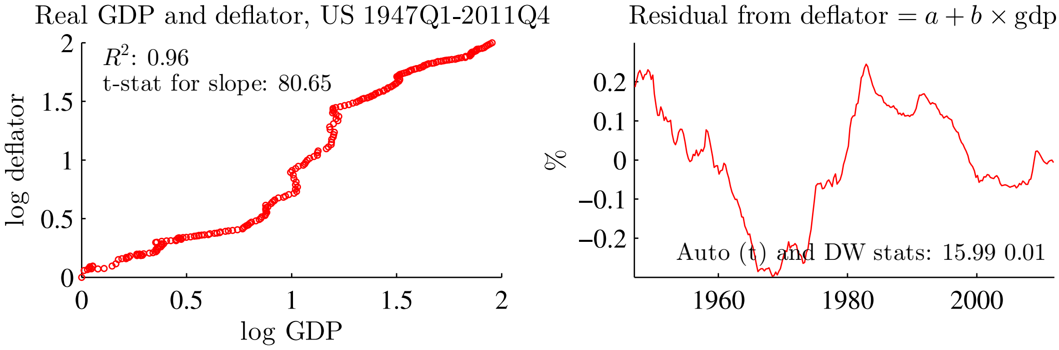


Figure 7.8: Example of a spurious regression

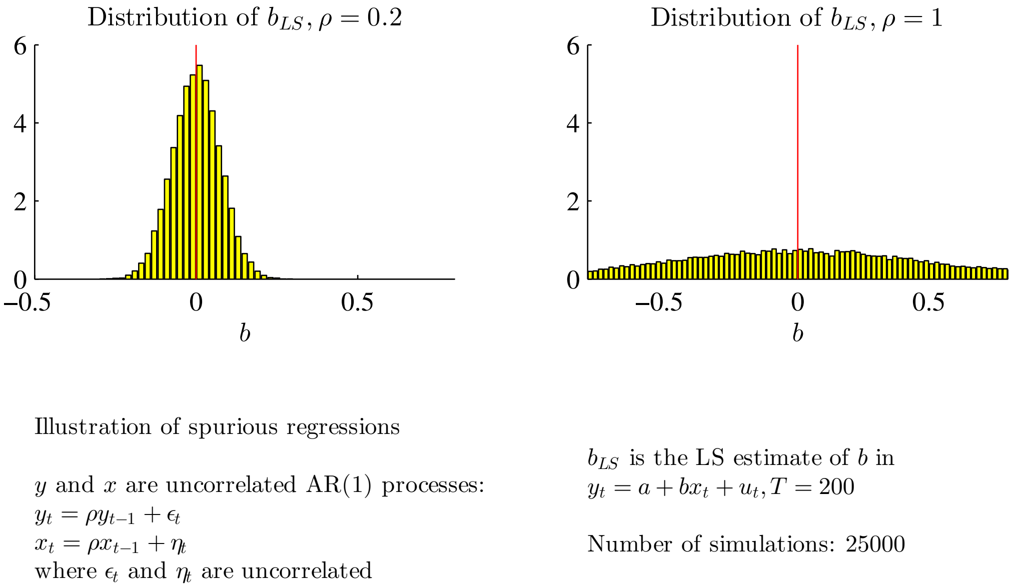


Illustration of spurious regressions

y and x are uncorrelated AR(1) processes:
$y_t = \rho y_{t-1} + \epsilon_t$
$x_t = \rho x_{t-1} + \eta_t$
where ϵ_t and η_t are uncorrelated

b_{LS} is the LS estimate of b in
$y_t = a + bx_t + u_t, T = 200$

Number of simulations: 25000

Figure 7.9: Distribution of LS estimator when y_t and x_t are independent AR(1) processes

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, (7.60), gives

$$y_t - y_{t-1} = \beta + \varepsilon_t - \varepsilon_{t-1}. \tag{7.64}$$

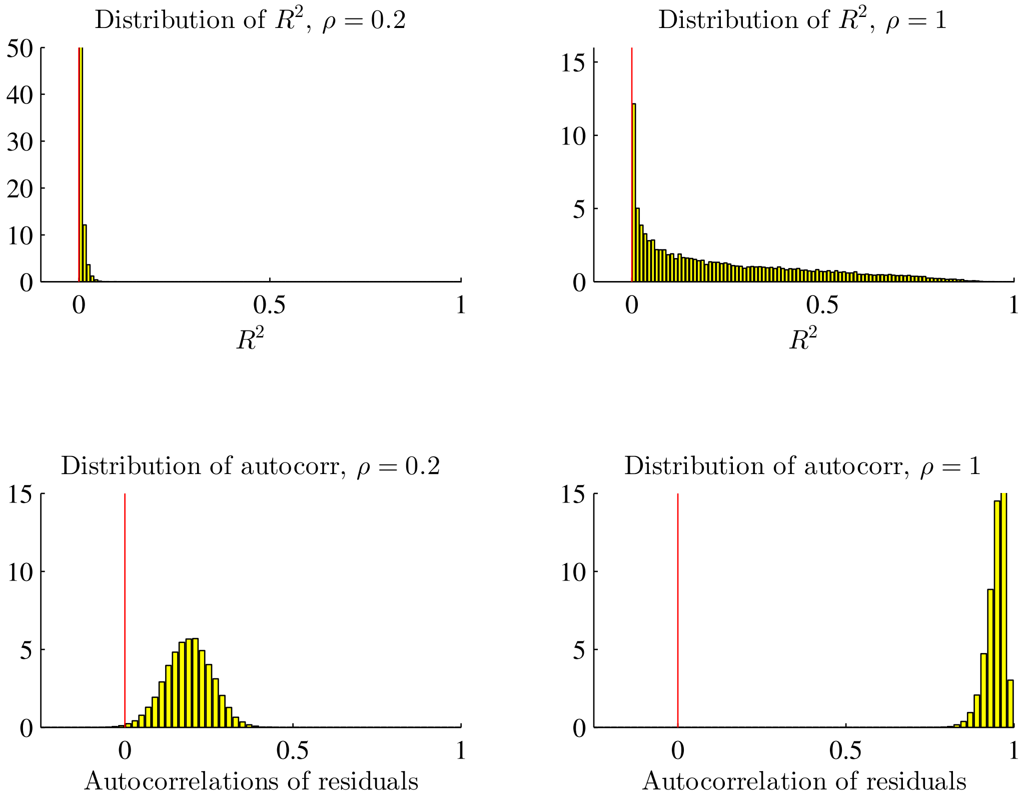


Figure 7.10: Distribution of R^2 and autorrelation of residuals. See Figure 7.9

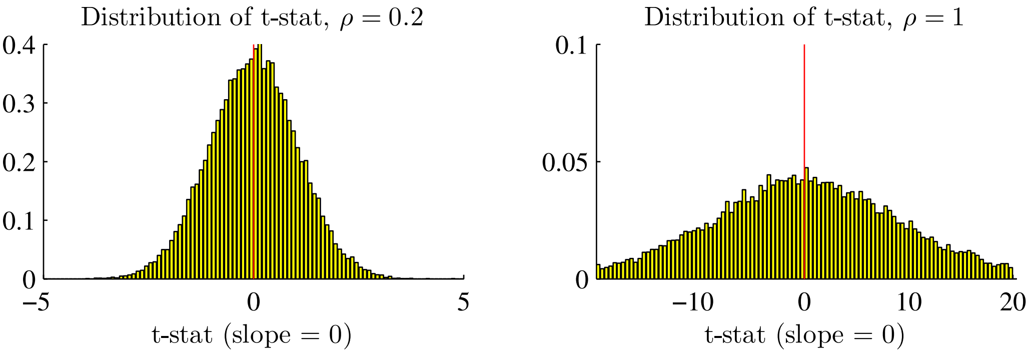


Figure 7.11: Distribution of t-statistics. See Figure 7.9

Its unclear if this is an improvement: the trend is gone, but the errors are now of MA(1) type (in fact, non-invertible, and therefore tricky, in particular for estimation).

7.9.3 Testing for a Unit Root*

Suppose we run an OLS regression of

$$y_t = a y_{t-1} + \varepsilon_t, \tag{7.65}$$

where the true value of $|a| < 1$. The asymptotic distribution of the LS estimator is

$$\sqrt{T}\left(\hat{a} - a\right) \sim N\left(0, 1 - a^2\right). \tag{7.66}$$

(The variance follows from the standard OLS formula where the variance of the estimator is $\sigma^2\left(X'X/T\right)^{-1}$. Here plim $X'X/T = \text{Var}\left(y_t\right)$ which we know is $\sigma^2/\left(1 - a^2\right)$).

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 (7.66). 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 (7.65).

In practice, the approach is to run the regression (7.65) with a constant (and perhaps even a time trend), calculate the test statistic

$$DF = \frac{\hat{a} - 1}{\text{Std}(\hat{a})}, \tag{7.67}$$

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 (7.65)), do an *augmented DickeyFuller test* (ADF)

$$\begin{aligned} y_t &= \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_{2t}, \text{ or} \\ \Delta y_t &= \delta + \left(\theta_1 + \theta_2 - 1\right) y_{t-1} - \theta_2 \Delta y_{t-1} + \varepsilon_{2t}, \end{aligned} \tag{7.68}$$

and test if $\theta_1 + \theta_2 - 1 = 0$ (against the alternative, < 0) 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. It has three steps. First, regress

$$y_t = a + \varepsilon_t. \tag{7.69}$$

Second, define

$$S_t = \textstyle\sum_{s=1}^{t} \hat{\varepsilon}_s \text{ for } t = 1, ..., T \text{ and let} \tag{7.70}$$

$$\hat{\sigma}^2 = \text{Var}(\hat{\varepsilon}_t). \tag{7.71}$$

Third, the test statistic is

$$KPSS = \frac{1}{T^2} \textstyle\sum_{t=1}^{T} S_t^2 / \hat{\sigma}^2 \tag{7.72}$$

Reject stationarity if $KPSS > 0.463$ (a 5% critical value). We could also include a linear trend in (KPSSReg). 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 of 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.

7.9.4 Cointegration*

An exception to the "spurious regression" result: Y_t and X_t are I(1) but share a common stochastic trend such that

$$y_t - \alpha - \beta x_t \text{ is I(0)}. \tag{7.73}$$

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 (7.73), we call $(1, -\beta)$ the *cointegrating vector*, since

$$\begin{bmatrix} 1 & -\beta \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} \text{ is I(0)} \tag{7.74}$$

Example 7.13 *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)*

$$\begin{aligned} y_t &= \gamma A_t + z_t \\ x_t &= \delta A_t + w_t \end{aligned}$$

From the second equation $A_t = (x_t - w_t)/\delta$, use in first equation

$$\underbrace{y_t}_{I(1)} = \frac{\gamma}{\delta}\underbrace{x_t}_{I(1)} + \underbrace{z_t - \frac{\gamma}{\delta}w_t}_{I(0)}$$

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

$$y_t = a + bx_t + \varepsilon_t \tag{7.75}$$

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).

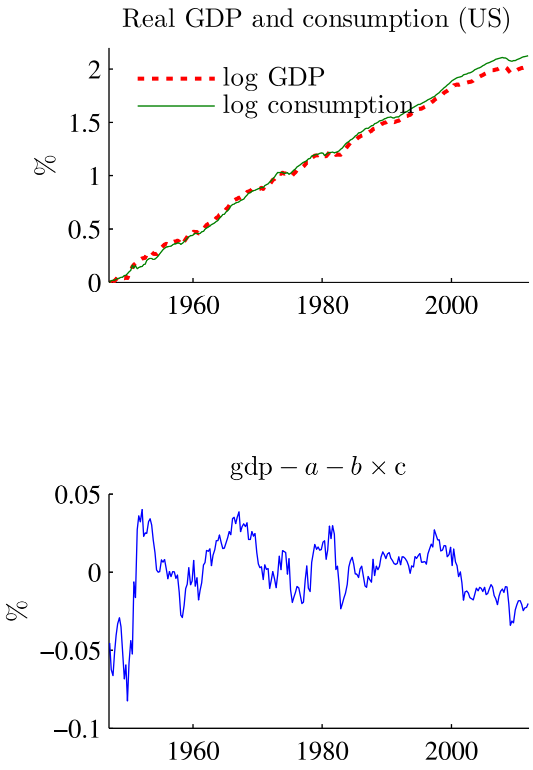
One way to incorporate the cointegration in a model of the short-run dynamics is to use a *Error-Correction Model*, for instance,

$$\begin{aligned} \Delta y_t &= \delta + \phi_1 \Delta x_{t-1} - \gamma\left(y_{t-1} - \beta x_{t-1}\right) + \varepsilon_t \text{ or perhaps} \qquad (7.76) \\ &= \delta + \phi_1 \Delta x_{t-1} + \theta_1 \Delta y_{t-1} - \gamma\left(y_{t-1} - \beta x_{t-1}\right) + \varepsilon_t \end{aligned}$$

Recall: (y_t, x_t) are I(1), but $y_{t-1} - \beta x_{t-1}$ is I(0), so all terms in (7.76) 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 ack 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 (7.76) and estimate the rest of the parameters. (Standard inference applies to them.)



t-stat for autocorr of residual
in ADF test of gdp or c

	gdp	c
1 lag in ADF test	5.82	1.28
2 lags in ADF test	-0.51	-0.36
3 lags in ADF test	0.21	-0.08

ADF test of series

	gdp	c
1 lag in ADF test	-1.79	-0.23
2 lags in ADF test	-1.97	-1.32
3 lags in ADF test	-1.56	-1.37

t-stat for autocorr of residual
in ADF test of gdp − a − b × c

1 lag in ADF test	-1.19
2 lags in ADF test	-0.04
3 lags in ADF test	-0.26

ADF test

1 lag in ADF test	-4.03
2 lags in ADF test	-4.16
3 lags in ADF test	-4.29

Figure 7.12: Unit root tests, US quarterly macro data

Bibliography

Brockwell, P. J., and R. A. Davis, 1991, *Time series: theory and methods*, Springer Verlag, New York, second edn.

Newbold, P., 1995, *Statistics for business and economics*, Prentice-Hall, 4th edn.

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.

	Δgdp
Coint res$_{t-1}$	-0.10
	(-2.56)
Δgdp$_{t-1}$	0.15
	(1.61)
Δc$_{t-1}$	0.33
	(3.00)
Δgdp$_{t-2}$	0.02
	(0.24)
Δc$_{t-2}$	0.22
	(2.34)
const	0.00
	(2.01)
R2	0.25
obs	257.00

Table 7.1: Error-correction model for log real US GDP growth, 1947Q1-2011Q4. Numbers in parentheses are t-stats. The 'Coint res' is the residual from regressing the log GDP level on the log consumption level.

8 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.

8.1 Autocorrelations

8.1.1 Autocorrelation Coefficients and the Box-Pierce Test

The autocovariances of the y_t process can be estimated as

$$\hat{\gamma}_s = \frac{1}{T} \sum_{t=1+s}^{T} (y_t - \bar{y})(y_{t-s} - \bar{y}), \text{ with} \tag{8.1}$$

$$\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t. \tag{8.2}$$

(We typically divide by T in (8.1) even if we have only $T-s$ full observations to estimate γ_s from.) Autocorrelations are then estimated as

$$\hat{\rho}_s = \hat{\gamma}_s / \hat{\gamma}_0. \tag{8.3}$$

The sampling properties of $\hat{\rho}_s$ are complicated, but there are several useful large sample results for Gaussian processes (these results typically carry over to processes which are similar to the Gaussian—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

$$\sqrt{T} \begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \to^d N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \tag{8.4}$$

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).

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 -1.65 and another 5% above 1.65, we can reject the null hypothesis at the 10% level if* $\sqrt{T}|\hat{\rho}_1| > 1.65$. *With* $T = 100$, *we therefore need* $|\hat{\rho}_1| > 1.65/\sqrt{100} = 0.165$ *for rejection, and with* $T = 1000$ *we need* $|\hat{\rho}_1| > 1.65/\sqrt{1000} \approx 0.052$.

The *Box-Pierce test* follows directly from the result in (8.4), since it shows that $\sqrt{T}\hat{\rho}_i$ and $\sqrt{T}\hat{\rho}_j$ are iid N(0,1) variables. Therefore, the sum of the square of them is distributed as a χ^2 variable. The test statistics typically used is

$$Q_L = T\sum_{s=1}^{L}\hat{\rho}_s^2 \to^d \chi_L^2. \tag{8.5}$$

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).

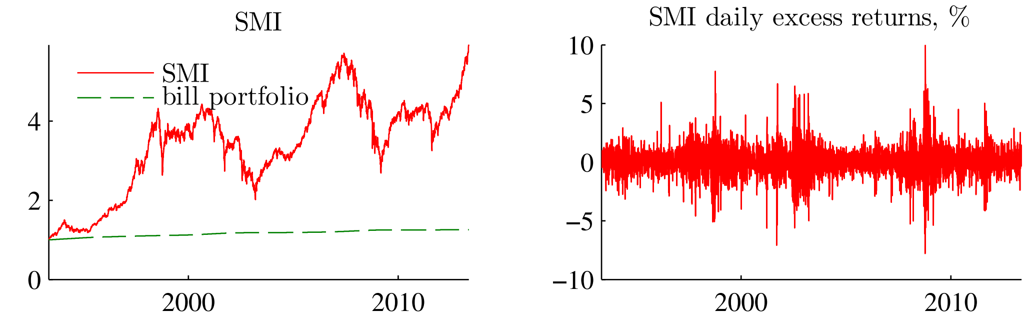
8.1.2 Autoregressions

An alternative way of testing autocorrelations is to estimate an AR model

$$y_t = c + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \varepsilon_t, \tag{8.6}$$

and then test if all slope coefficients $(a_1, a_2, ..., a_p)$ are zero with a χ^2 or F test. This approach is somewhat less general than the Box-Pierce test, but most stationary time series processes can be well approximated by an AR of relatively low order.

See Figure 8.4 for an illustration.



Daily SMI data, 1993:5-2013:5
1st order autocorrelation of returns (daily, weekly, monthly): 0.03 -0.11 0.04
1st order autocorrelation of absolute returns (daily, weekly, monthly): 0.28 0.31 0.19

Figure 8.1: Time series properties of SMI

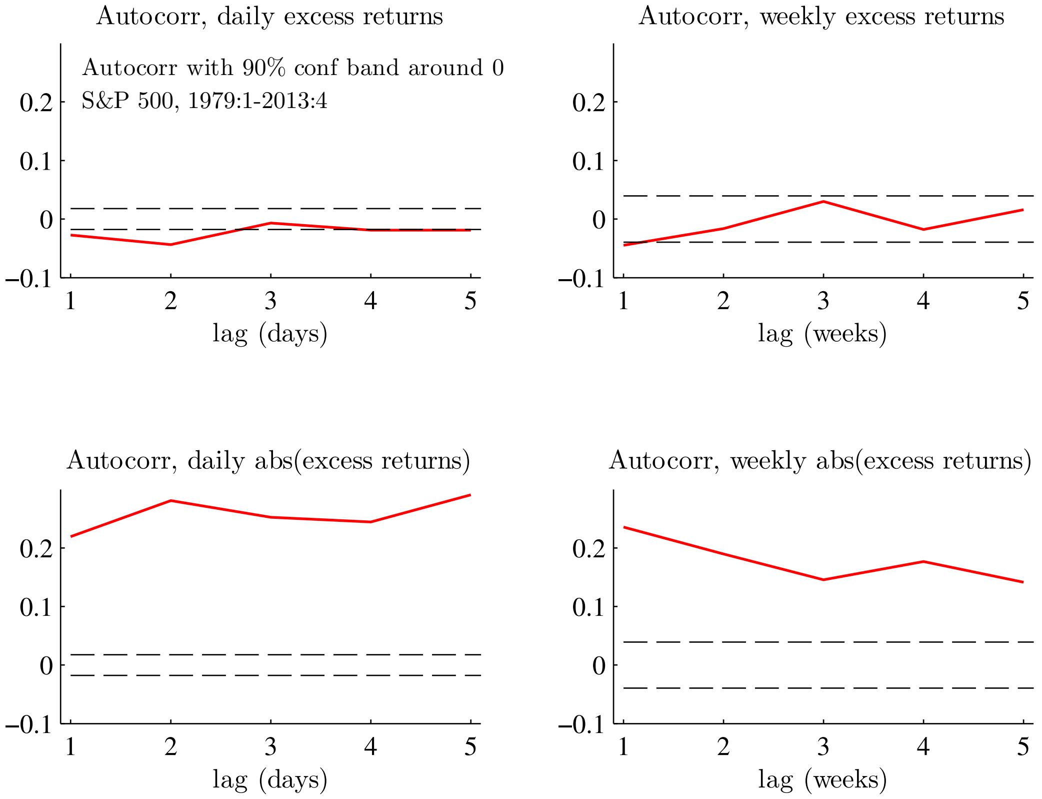


Figure 8.2: Predictability of US stock returns

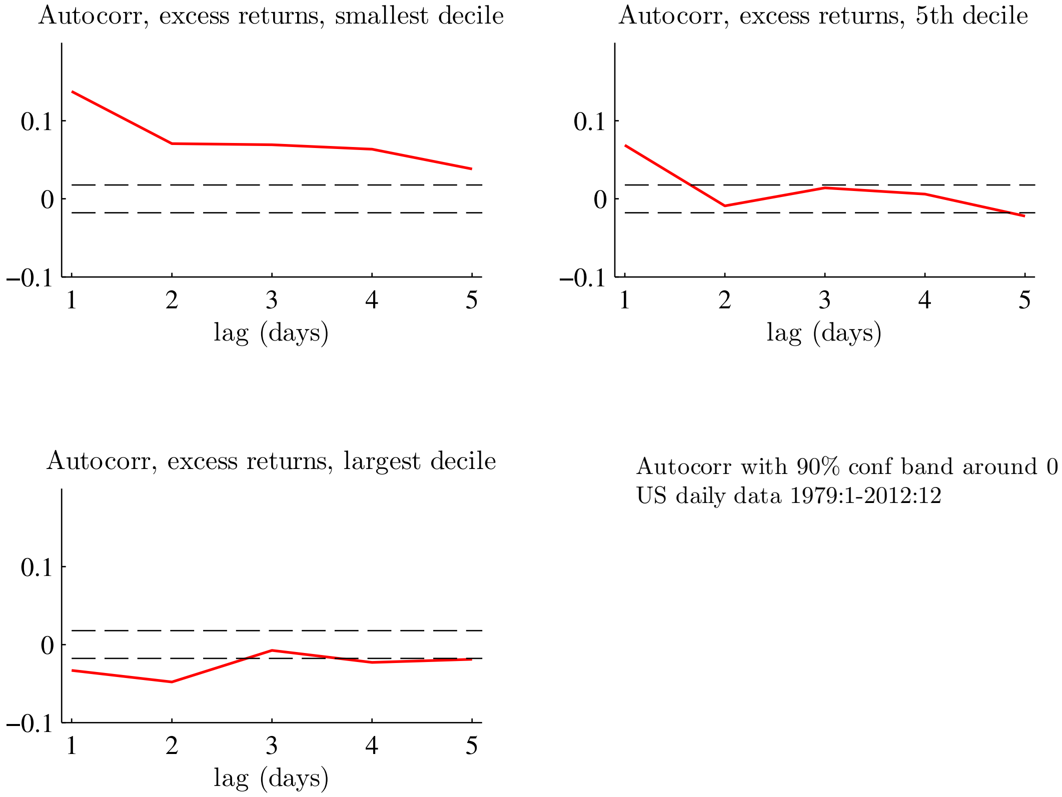


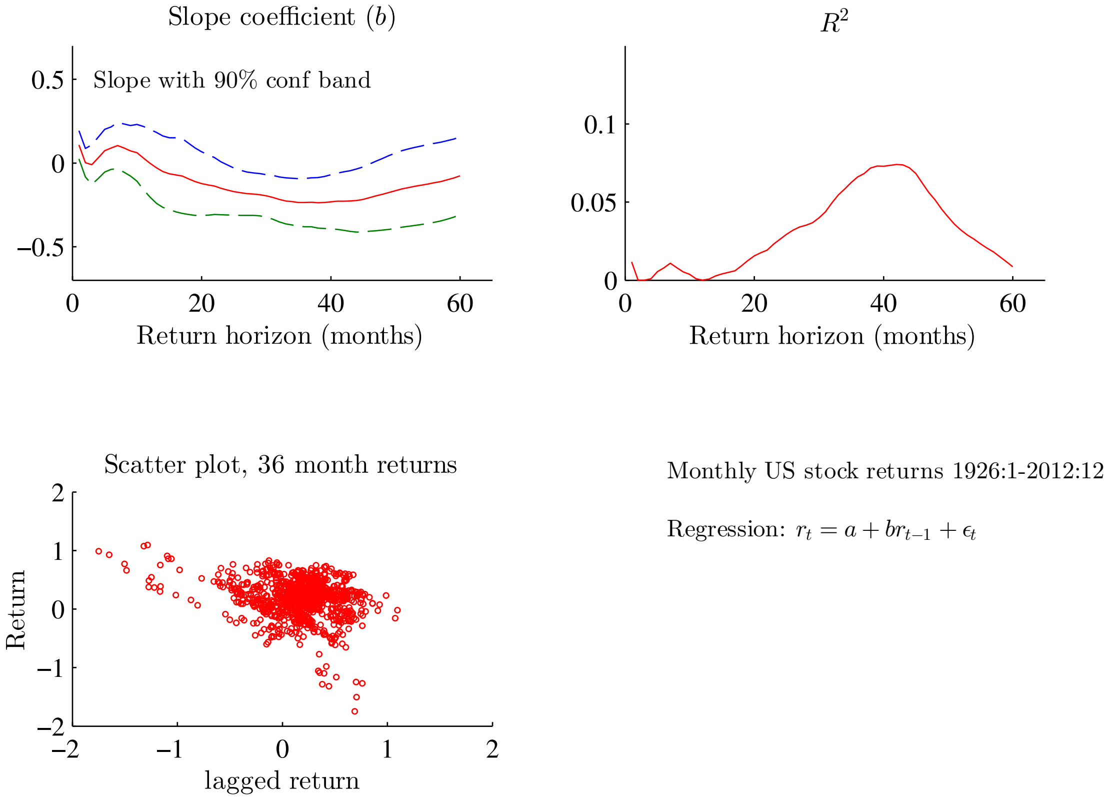
Figure 8.3: Predictability of US stock returns, size deciles

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

$$y_t = c + a\delta(y_{t-1} \leq 0)y_{t-1} + b\delta(y_{t-1} > 0)y_{t-1} + \varepsilon_t \text{, where} \tag{8.7}$$
$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

See Figure 8.5 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 (use the weekly return for a particular weekday only, say Wednesdays), the standard LS



Monthly US stock returns 1926:1-2012:12

Regression: $r_t = a + br_{t-1} + \epsilon_t$

Figure 8.4: Predictability of US stock returns

expression 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.

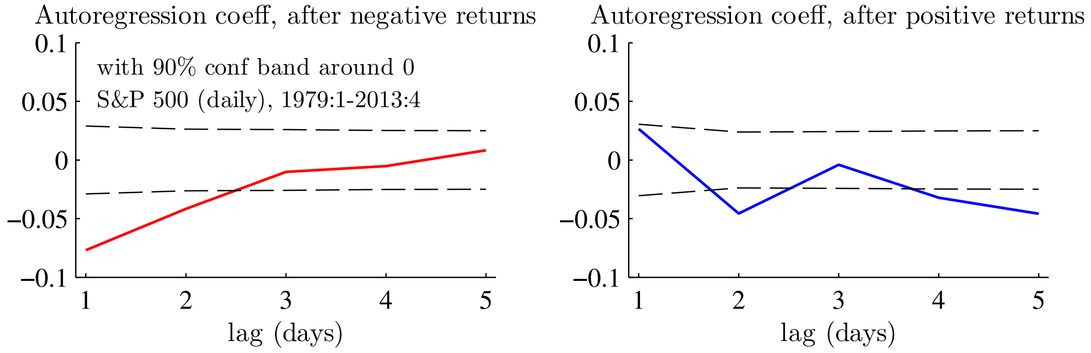
Remark 8.3 *(Overlapping returns*) Consider an AR(1) for the two-period return, $y_{t-1}+$ y_t*

$$y_{t+1} + y_{t+2} = a + b_2 (y_{t-1} + y_t) + \varepsilon_{t+2}.$$

Two successive observations with non-overlapping returns are then

$$\begin{aligned}
y_{t+1} + y_{t+2} &= a + b_2 (y_{t-1} + y_t) + \varepsilon_{t+2} \\
y_{t+3} + y_{t+4} &= a + b_2 (y_{t+1} + y_{t+2}) + \varepsilon_{t+4}.
\end{aligned}$$

Suppose that y_t is not autocorrelated, so the slope coefficient $b_2 = 0$. We can then write



Based on the following regression:
$r_t = \alpha + \beta(1 - Q_{t-1})r_{t-1} + \gamma Q_{t-1} r_{t-1} + \epsilon_t$
$Q_{t-1} = 1$ if $r_{t-1} > 0$, and zero otherwise

Figure 8.5: Predictability of US stock returns, results from a regression with interactive dummies

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

$$\varepsilon_{t+2} = -a + y_{t+1} + y_{t+2}$$
$$\varepsilon_{t+3} = -a + y_{t+2} + y_{t+3},$$

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.

8.1.3 Autoregressions versus Autocorrelations*

It is straightforward to see the relation between autocorrelations and the AR model when the AR model is the true process. This relation is given by the *Yule-Walker equations*.

For an AR(1), the autoregression coefficient is simply the first autocorrelation coefficient. For an AR(2), $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$, we have

$$\begin{bmatrix} \operatorname{Cov}(y_t, y_t) \\ \operatorname{Cov}(y_{t-1}, y_t) \\ \operatorname{Cov}(y_{t-2}, y_t) \end{bmatrix} = \begin{bmatrix} \operatorname{Cov}(y_t, a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t) \\ \operatorname{Cov}(y_{t-1}, a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t) \\ \operatorname{Cov}(y_{t-2}, a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t) \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \operatorname{Cov}(y_t, y_{t-1}) + a_2 \operatorname{Cov}(y_t, y_{t-2}) + \operatorname{Cov}(y_t, \varepsilon_t) \\ a_1 \operatorname{Cov}(y_{t-1}, y_{t-1}) + a_2 \operatorname{Cov}(y_{t-1}, y_{t-2}) \\ a_1 \operatorname{Cov}(y_{t-2}, y_{t-1}) + a_2 \operatorname{Cov}(y_{t-2}, y_{t-2}) \end{bmatrix}, \text{ or}$$

$$\begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} a_1\gamma_1 + a_2\gamma_2 + \operatorname{Var}(\varepsilon_t) \\ a_1\gamma_0 + a_2\gamma_1 \\ a_1\gamma_1 + a_2\gamma_0 \end{bmatrix}. \tag{8.8}$$

To transform to autocorrelation, divide by γ_0. The last two equations are then

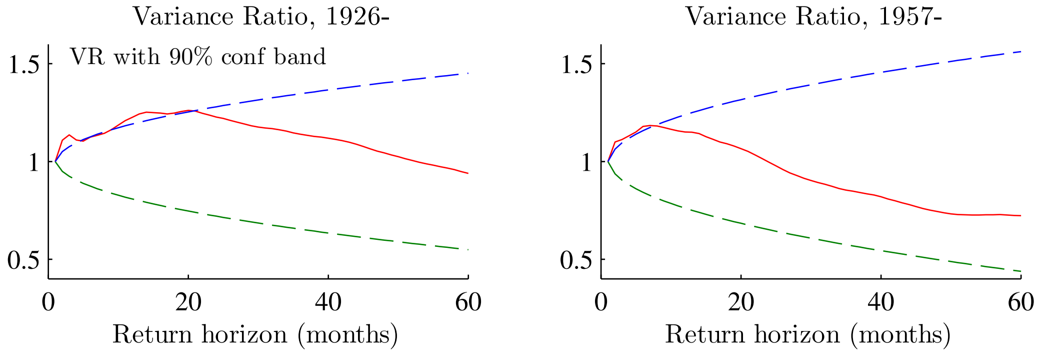
$$\begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2\rho_1 \\ a_1\rho_1 + a_2 \end{bmatrix} \text{ or } \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} a_1/(1-a_2) \\ a_1^2/(1-a_2) + a_2 \end{bmatrix}. \tag{8.9}$$

If we know the parameters of the AR(2) model (a_1, a_2, and $\operatorname{Var}(\varepsilon_t)$), then we can solve for the autocorrelations. Alternatively, if we know the autocorrelations, then we can solve for the autoregression coefficients. This demonstrates that testing if all the autocorrelations are zero is essentially the same as testing if all the autoregressive coefficients are zero. Note, however, that the transformation is non-linear, which may make a difference in small samples.

8.1.4 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

$$VR_q = \frac{\operatorname{Var}\left(\sum_{s=0}^{q-1} y_{t-s}\right)}{q \operatorname{Var}(y_t)}. \tag{8.10}$$



Monthly US stock returns 1926:1-2012:12

The confidence bands use the asymptotic sampling distribution of the variance ratios

Figure 8.6: Variance ratios, US excess stock returns

To see that this is related to predictability, consider the 2-period variance ratio.

$$
\begin{aligned}
VR_2 &= \frac{\operatorname{Var}(y_t + y_{t-1})}{2\operatorname{Var}(y_t)} && (8.11)\\
&= \frac{\operatorname{Var}(y_t) + \operatorname{Var}(y_{t-1}) + 2\operatorname{Cov}(y_t, y_{t-1})}{2\operatorname{Var}(y_t)}\\
&= 1 + \frac{\operatorname{Cov}(y_t, y_{t-1})}{\operatorname{Var}(y_t)}\\
&= 1 + \rho_1. && (8.12)
\end{aligned}
$$

It is clear from (8.12) 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 (8.10) 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 (8.10)

$$
\begin{aligned}
VR_q &= \frac{\operatorname{Var}\left(\sum_{s=0}^{q-1} y_{t-s}\right)}{q\operatorname{Var}(y_t)} \\
&= \sum_{s=-(q-1)}^{q-1}\left(1-\frac{|s|}{q}\right)\rho_s \text{ or} \\
&= 1+2\sum_{s=1}^{q-1}\left(1-\frac{s}{q}\right)\rho_s. \qquad (8.13)
\end{aligned}
$$

To estimate VR_q, we first estimate the autocorrelation coefficients (using all available data points for each estimation) and then calculate (8.13).

Remark 8.4 *(*Sampling distribution of $\widehat{VR}_q$) Under the null hypothesis that there is no autocorrelation, (8.4) and (8.13) give*

$$
\sqrt{T}\left(\widehat{VR}_q-1\right)\to^d N\left[0,\sum_{s=1}^{q-1}4\left(1-\frac{s}{q}\right)^2\right].
$$

Example 8.5 *(Sampling distributions of $\widehat{VR}_2$ and $\widehat{VR}_3$)*

$$
\begin{aligned}
&\sqrt{T}\left(\widehat{VR}_2-1\right)\to^d N(0,1) \textit{ or } \widehat{VR}_2\to^d N(1,1/T) \\
\textit{and } &\sqrt{T}\left(\widehat{VR}_3-1\right)\to^d N(0,20/9) \textit{ or } \widehat{VR}_3\to^d N\left[1,(20/9)/T\right].
\end{aligned}
$$

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).

See Figure 8.6 for an illustration.

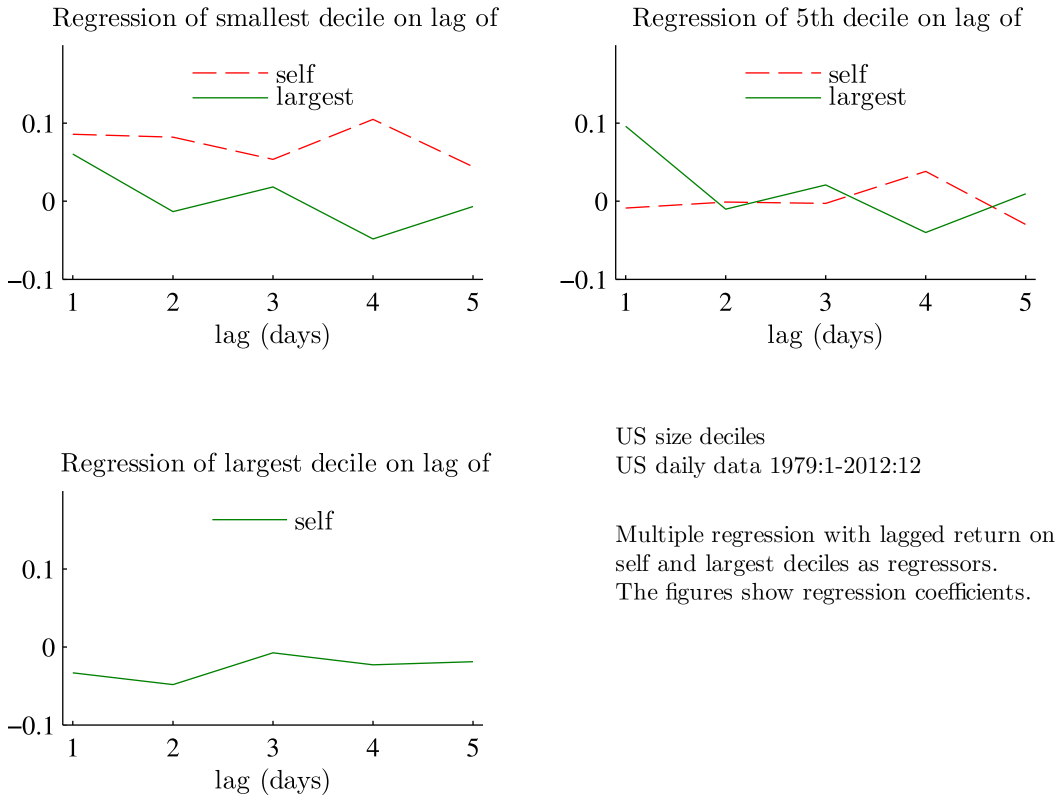
8.2 Other Predictors and Methods

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.

8.2.1 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.

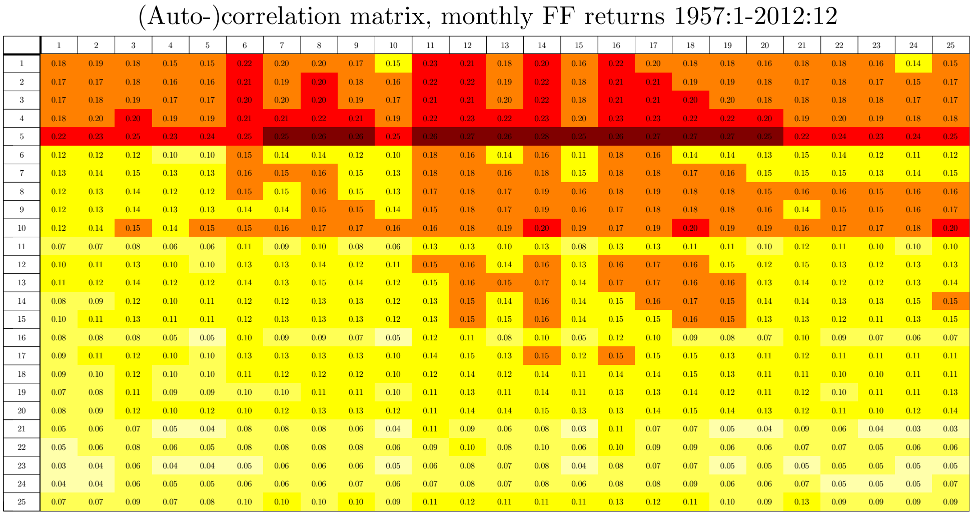
See Figure 8.7 for an illustration.



US size deciles
US daily data 1979:1-2012:12

Multiple regression with lagged return on
self and largest deciles as regressors.
The figures show regression coefficients.

Figure 8.7: Coefficients from multiple prediction regressions



(Auto-)correlation matrix, monthly FF returns 1957:1-2012:12

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0.18	0.19	0.18	0.15	0.15	0.22	0.20	0.20	0.17	0.15	0.23	0.21	0.18	0.20	0.16	0.22	0.20	0.18	0.18	0.16	0.18	0.18	0.16	0.14	0.15
2	0.17	0.17	0.18	0.16	0.16	0.21	0.19	0.20	0.18	0.16	0.22	0.22	0.19	0.22	0.18	0.21	0.21	0.19	0.19	0.18	0.17	0.18	0.17	0.15	0.17
3	0.17	0.18	0.19	0.17	0.17	0.20	0.20	0.20	0.19	0.17	0.21	0.21	0.20	0.22	0.18	0.21	0.21	0.20	0.20	0.18	0.18	0.18	0.18	0.17	0.17
4	0.18	0.20	0.20	0.19	0.19	0.21	0.21	0.22	0.21	0.19	0.22	0.23	0.22	0.23	0.20	0.23	0.23	0.22	0.22	0.20	0.19	0.20	0.19	0.18	0.18
5	0.22	0.23	0.25	0.23	0.24	0.25	0.25	0.26	0.26	0.25	0.26	0.27	0.26	0.28	0.25	0.26	0.27	0.27	0.27	0.25	0.22	0.24	0.23	0.24	0.25
6	0.12	0.12	0.12	0.10	0.10	0.15	0.14	0.14	0.12	0.10	0.18	0.16	0.14	0.16	0.11	0.18	0.16	0.14	0.14	0.13	0.15	0.14	0.12	0.11	0.12
7	0.13	0.14	0.15	0.13	0.13	0.16	0.15	0.16	0.15	0.13	0.18	0.18	0.16	0.18	0.15	0.18	0.18	0.17	0.16	0.15	0.15	0.15	0.13	0.14	0.15
8	0.12	0.13	0.14	0.12	0.12	0.15	0.15	0.16	0.15	0.13	0.17	0.18	0.17	0.19	0.16	0.18	0.19	0.18	0.18	0.15	0.16	0.16	0.15	0.16	0.16
9	0.12	0.13	0.14	0.13	0.13	0.14	0.14	0.15	0.15	0.14	0.15	0.18	0.17	0.19	0.16	0.17	0.18	0.18	0.18	0.16	0.14	0.15	0.15	0.16	0.17
10	0.12	0.14	0.15	0.14	0.15	0.15	0.16	0.17	0.17	0.16	0.16	0.18	0.19	0.20	0.19	0.17	0.19	0.20	0.19	0.19	0.16	0.17	0.17	0.18	0.20
11	0.07	0.07	0.08	0.06	0.06	0.11	0.09	0.10	0.08	0.06	0.13	0.13	0.10	0.13	0.08	0.13	0.13	0.11	0.11	0.10	0.12	0.11	0.10	0.10	0.10
12	0.10	0.11	0.13	0.10	0.10	0.13	0.13	0.14	0.12	0.11	0.15	0.16	0.14	0.16	0.13	0.16	0.17	0.16	0.15	0.12	0.15	0.13	0.12	0.13	0.13
13	0.11	0.12	0.14	0.12	0.12	0.14	0.13	0.15	0.14	0.12	0.15	0.16	0.15	0.17	0.14	0.17	0.17	0.16	0.16	0.13	0.14	0.12	0.12	0.13	0.14
14	0.08	0.09	0.12	0.10	0.11	0.12	0.12	0.13	0.13	0.12	0.13	0.15	0.14	0.16	0.14	0.15	0.16	0.17	0.15	0.14	0.14	0.13	0.13	0.15	0.15
15	0.10	0.11	0.13	0.11	0.11	0.12	0.13	0.13	0.13	0.12	0.13	0.15	0.15	0.16	0.14	0.15	0.15	0.16	0.15	0.13	0.13	0.12	0.11	0.13	0.15
16	0.08	0.08	0.08	0.05	0.05	0.10	0.09	0.09	0.07	0.05	0.12	0.11	0.08	0.10	0.05	0.12	0.10	0.09	0.08	0.07	0.10	0.09	0.07	0.06	0.07
17	0.09	0.11	0.12	0.10	0.10	0.13	0.13	0.13	0.13	0.10	0.14	0.15	0.13	0.15	0.12	0.15	0.15	0.15	0.13	0.11	0.12	0.11	0.11	0.11	0.11
18	0.09	0.10	0.12	0.10	0.10	0.11	0.12	0.12	0.12	0.10	0.12	0.14	0.12	0.14	0.11	0.14	0.14	0.15	0.13	0.11	0.11	0.10	0.10	0.11	0.11
19	0.07	0.08	0.11	0.09	0.09	0.10	0.10	0.11	0.11	0.10	0.11	0.13	0.11	0.14	0.11	0.13	0.13	0.14	0.12	0.11	0.12	0.10	0.11	0.11	0.13
20	0.08	0.09	0.12	0.10	0.12	0.10	0.12	0.13	0.13	0.12	0.11	0.14	0.14	0.15	0.13	0.13	0.14	0.15	0.14	0.13	0.12	0.11	0.10	0.12	0.14
21	0.05	0.06	0.07	0.05	0.04	0.08	0.08	0.08	0.06	0.04	0.11	0.09	0.06	0.08	0.03	0.11	0.07	0.07	0.05	0.04	0.09	0.06	0.04	0.03	0.03
22	0.05	0.06	0.08	0.06	0.05	0.08	0.08	0.08	0.08	0.06	0.09	0.10	0.08	0.10	0.06	0.10	0.09	0.09	0.06	0.06	0.07	0.07	0.05	0.06	0.06
23	0.03	0.04	0.06	0.04	0.04	0.05	0.06	0.06	0.06	0.05	0.06	0.08	0.07	0.08	0.04	0.08	0.07	0.07	0.05	0.05	0.05	0.05	0.05	0.05	0.05
24	0.04	0.04	0.06	0.05	0.05	0.06	0.06	0.06	0.07	0.06	0.07	0.08	0.07	0.08	0.06	0.08	0.08	0.09	0.06	0.06	0.07	0.05	0.05	0.05	0.07
25	0.07	0.07	0.09	0.07	0.08	0.10	0.10	0.10	0.10	0.09	0.11	0.12	0.11	0.11	0.11	0.13	0.12	0.11	0.10	0.09	0.13	0.09	0.09	0.09	0.09

Figure 8.8: Illustration of the cross-autocorrelations, $\mathrm{Corr}(R_t, R_{t-k})$, monthly FF data. Dark colors indicate high correlations, light colors indicate low correlations.

8.2.2 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 dividend-price ratio (here in logs)

$$\sum_{s=1}^{q} r_{t+s} = \alpha + \beta_q(d_t - p_t) + \varepsilon_{t+q}. \tag{8.14}$$

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 8.9 for an illustration.

8.2.3 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-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

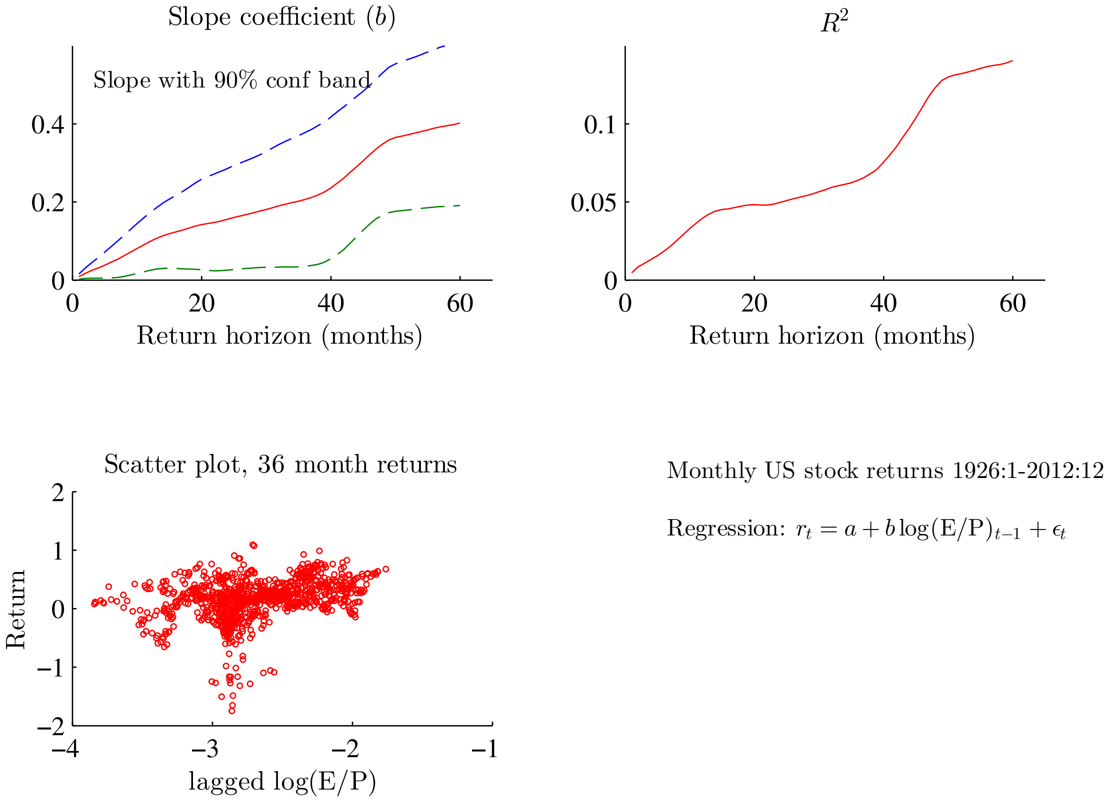


Figure 8.9: Predictability of US stock returns

without autocorrelation.

8.3 Out-of-Sample Forecasting Performance

8.3.1 In-Sample versus Out-of-Sample Forecasting

References: Goyal and Welch (2008), and Campbell and Thompson (2008)

Goyal and Welch (2008) find that the evidence of predictability of equity returns disappears when out-of-sample forecasts are considered. Campbell and Thompson (2008) claim that there is still some out-of-sample predictability, provided we put restrictions on the estimated models.

Campbell and Thompson (2008) first report that only few variables (earnings price ratio, T-bill rate and the inflation rate) have significant predictive power for one-month stock returns in the full sample (1871–2003 or early 1920s–2003, depending on predictor).

To gauge the out-of-sample predictability, they estimate the prediction equation using data up to and including $t-1$, and then make a forecast for period t. The forecasting performance of the equation is then compared with using the historical average as the predictor. Notice that this historical average is also estimated on data up to an 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): one model has just an intercept, the other has also a predictor. The comparison is done in terms of the RMSE and an "out-of-sample R^2"

$$R^2_{OS} = 1 - \frac{1}{T}\sum\nolimits_{t=s}^{T}(r_t - \hat{r}_t)^2 \Big/ \frac{1}{T}\sum\nolimits_{t=s}^{T}(r_t - \tilde{r}_t)^2, \tag{8.15}$$

where s is the first period with an out-of-sample forecast, $\hat{r}_t$ is the forecast based on the prediction model (estimated on data up to and including $t-1$) and $\tilde{r}_t$ 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 (also estimated on data up to and including $t-1$) as the benchmark prediction. That is, the benchmark prediction is that the return in t will equal the historical average.

The evidence shows that the out-of-sample forecasting performance is very weak—as claimed by Goyal and Welch (2008).

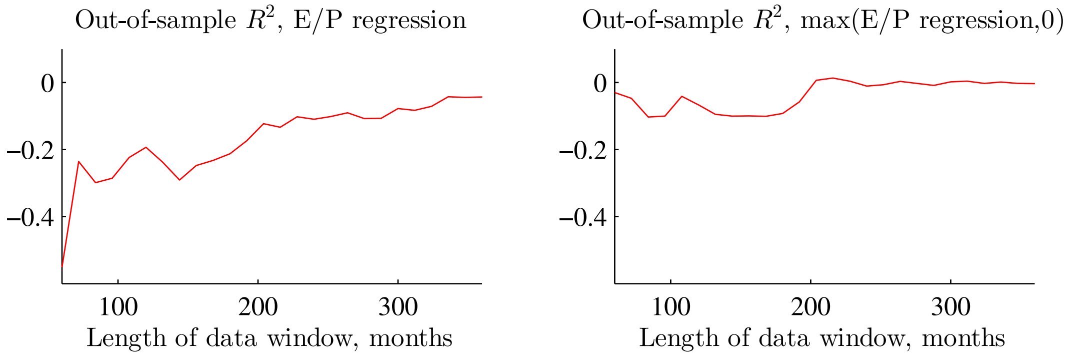
It is argued that forecasting equations can easily give strange results when they are estimated on a small data set (as they are early in the sample). They therefore try different restrictions: setting the slope coefficient to zero whenever the sign is "wrong," setting the prediction (or the historical average) to zero whenever the value is negative. This improves the results a bit—although the predictive performance is still weak.

See Figure 8.10 for an illustration.

8.3.2 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



US stock returns (1-year, in excess of riskfree) 1926:1-2012:12

Estimation is done on moving data window,
forecasts are made out of sample for: 1957:1-2012:12

Figure 8.10: Predictability of US stock returns, in-sample and out-of-sample

the performance of a portfolio by running the following regression

$$\begin{aligned} R_{1t} - R_{ft} &= \alpha + \beta(R_{mt} - R_{ft}) + \varepsilon_t, \text{ with} \\ \mathrm{E}\,\varepsilon_t &= 0 \text{ and } \operatorname{Cov}(R_{mt} - R_{ft}, \varepsilon_t) = 0, \end{aligned} \tag{8.16}$$

where $R_{1t} - R_{ft}$ is the excess return on the portfolio being studied and $R_{mt} - R_{ft}$ the excess returns of a vector of benchmark portfolios (for instance, only the market portfolio if we want to rely on CAPM; returns times conditional information if we want to allow for time-variation in expected benchmark returns). Neutral performance (mean-variance intersection, that is, that the tangency portfolio is unchanged and the two MV frontiers intersect there) requires $\alpha = 0$, which can be tested with a t test.

See Figure 8.11 for an illustration.

8.3.3 More Evidence on Out-of-Sample Forecasting Performance

Figures 8.12–8.16 illustrate the *out-of-sample performance on daily returns*. Figure 8.12 shows that extreme S&P 500 returns are followed by mean-reverting movements the following day—which suggests that a trading strategy should sell after a high return and buy after a low return. However, extreme returns are rare, so Figure 8.13 tries a simpler strate-

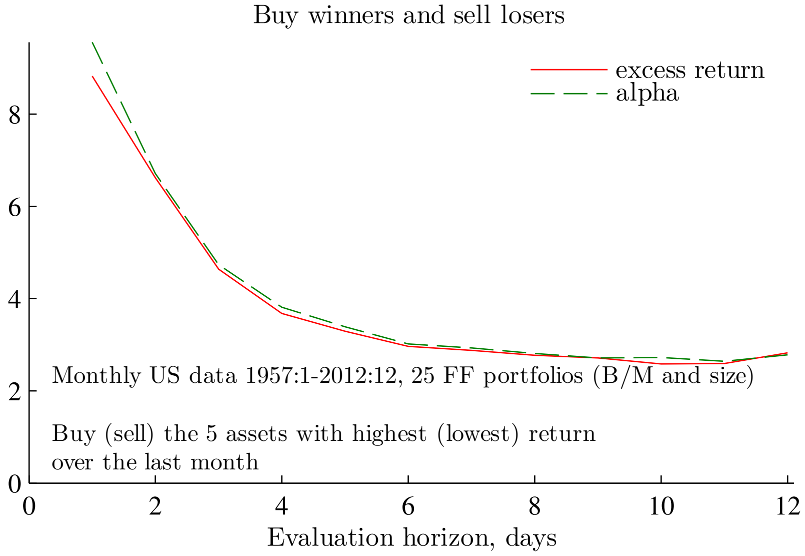


Figure 8.11: Predictability of US stock returns, momentum strategy

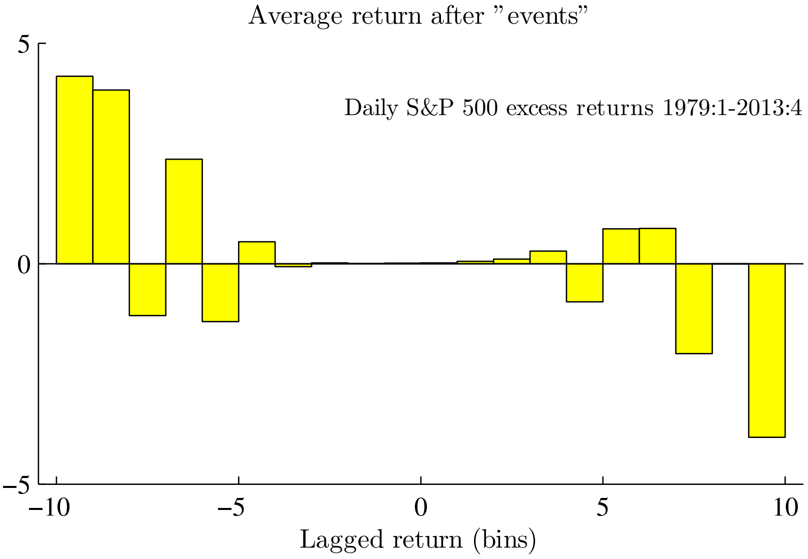
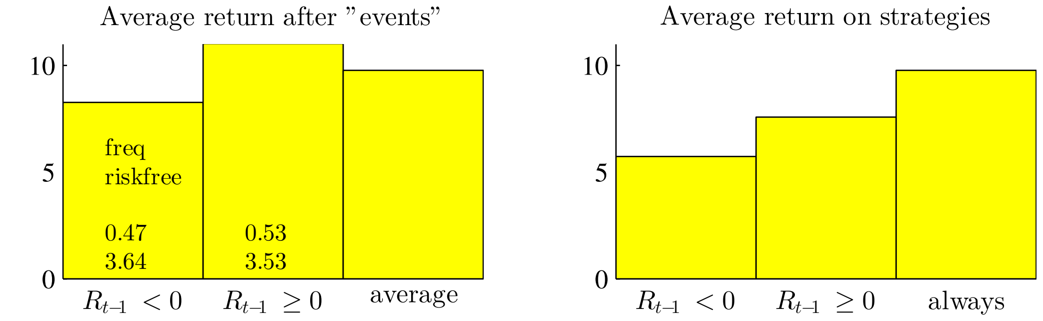


Figure 8.12: Short-run predictability of US stock returns, out-of-sample

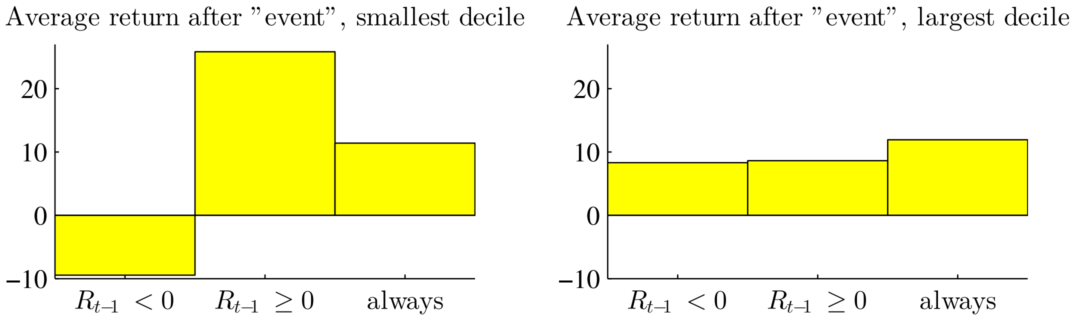
gies: buy after a negative return (or hold T-bills), or instead buy after a positive return (or hold T-bills). It turns out that the latter has a higher average return, which suggests that the extreme mean-reverting movements in Figure 8.12 are actually dominated by smaller mo-



Daily S&P 500 returns 1979:1-2013:4
The returns are annualized

Strategies (rebalanced daily):
hold stocks if condition is met;
otherwise, hold T-bills

Figure 8.13: Short-run predictability of US stock returns, out-of-sample

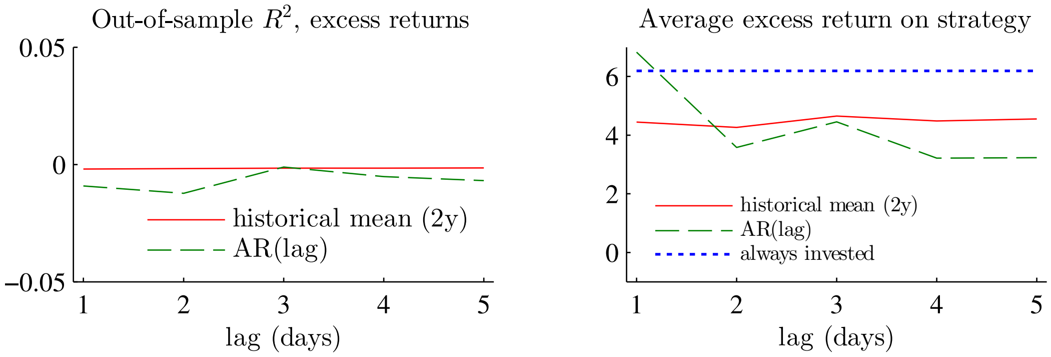


US size deciles (daily) 1979:1-2012:12

Strategies (rebalanced daily):
hold stocks if condition is met;
otherwise, hold T-bills

Figure 8.14: Short-run predictability of US stock returns, out-of-sample

mentum type changes (positive autocorrelation). However, always holding the S&P 500 index seems¨ to dominate both strategies—basically because stocks always outperform T-bills (in this setting). Notice that these strategies assume that you are always invested, in either stocks or the T-bill. In contrast, Figure 8.14 shows that the momentum strategy works reasonably well on small stocks.



S&P 500 daily excess returns, 1979:1-2013:4

The out-of-sample R^2 measures
the fit relative to forecasting 0

The strategies are based on forecasts
of excess returns:
(a) forecast>0: long in stock, short in riskfree
(b) forecast≤ 0: no investment

Figure 8.15: Short-run predictability of US stock returns, out-of-sample

Figure 8.15 shows out-of-sample R^2 and average returns of different strategies. The evidence suggests that an autoregressive model for the daily S&P 500 excess returns performs worse than forecasting zero (and so does using the historical average). In addition, the strategies based on the predicted excess return (from either the AR model or the historical returns) are worse than always being invested into the index. Notice that the strategies here allow for borrowing at the riskfree rate and also for leaving the market, so they are potentially more powerful than in the earlier figures. Figures 8.16 compares the results for small and large stocks—and illustrates that there is more predictability for small stocks.

Figures 8.17–8.19 illustrate the *out-of-sample performance on long-run returns*. Figure 8.17 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. To implement simple strategies based on this observation, 8.18 splits up the observation in (approximately) half: after low and after high p/e values. The results indicate that buying after low p/e ratios is better than after high p/e ratios, but that staying invested in the S&P 500 index all the time is better than sometimes switching over to T-bills. The reason is that even the low stock returns are higher than the interest rate.

Figure 8.19 studies the out-of-sample R^2 for simple forecasting models, and also al-

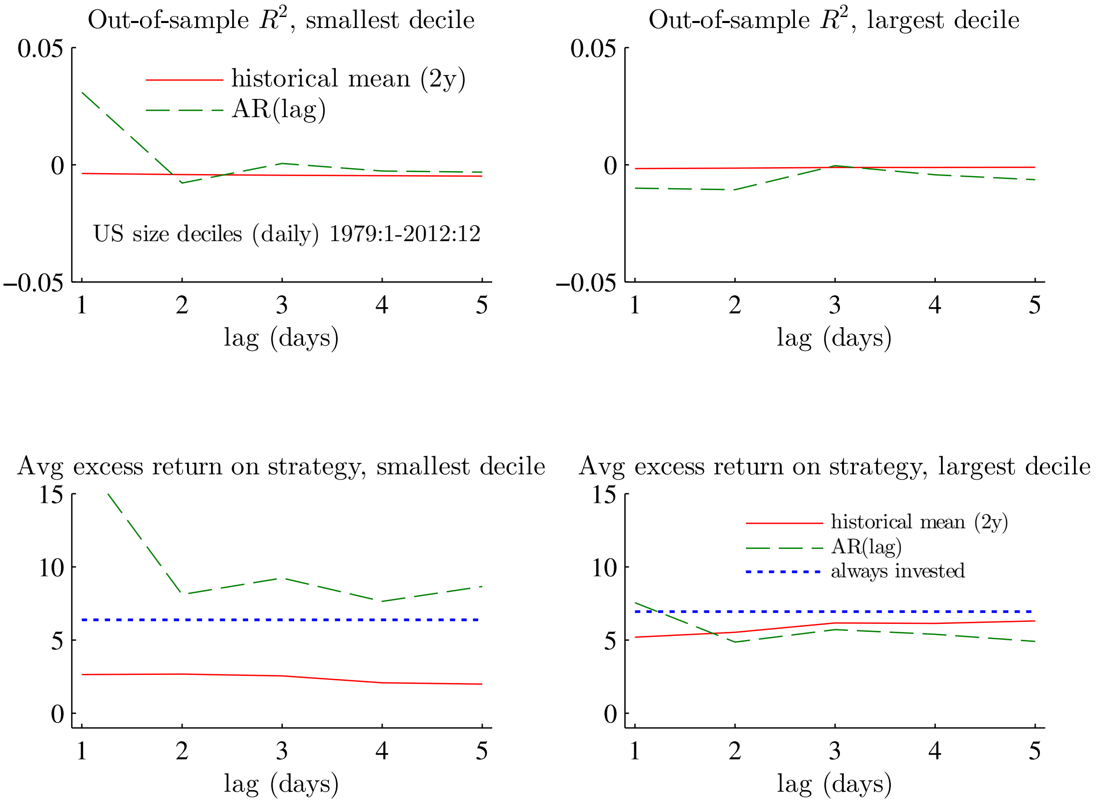


Figure 8.16: Short-run predictability of US stock returns, out-of-sample. See Figure 8.15 for details on the strategies.

lows for somewhat more flexible strategies (where we borrow at the riskfree rate and are allowed to leave the market). The evidence again suggests that it is hard to predict 1-year S&P 500 returns.

8.3.4 Technical Analysis

Main reference: Bodie, Kane, and Marcus (2002) 12.2; 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")

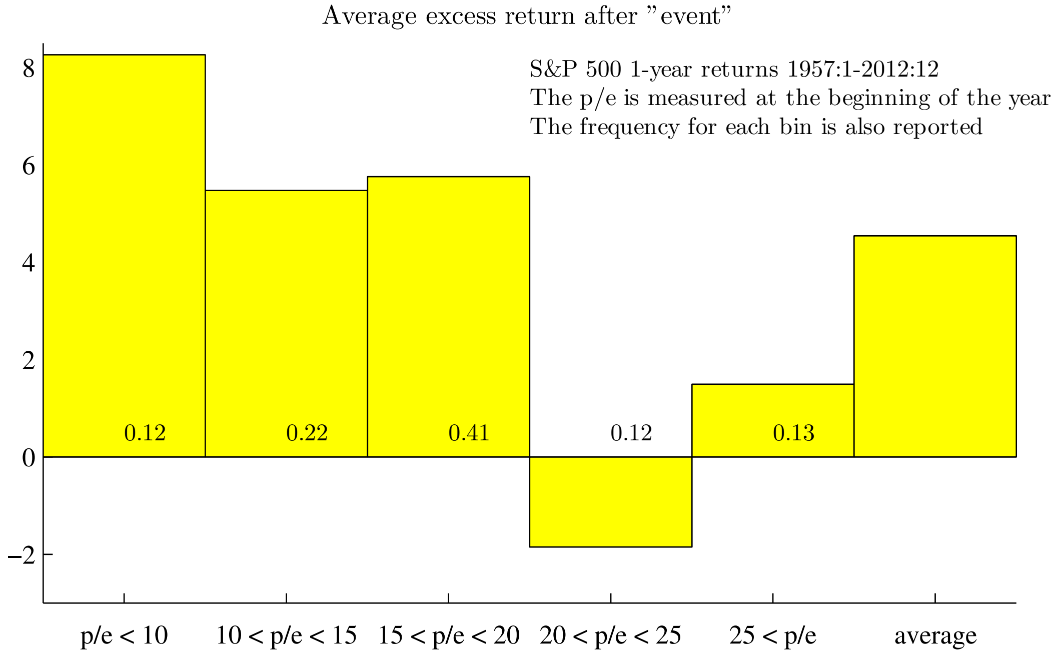


Figure 8.17: Long-run predictability of US stock returns, out-of-sample

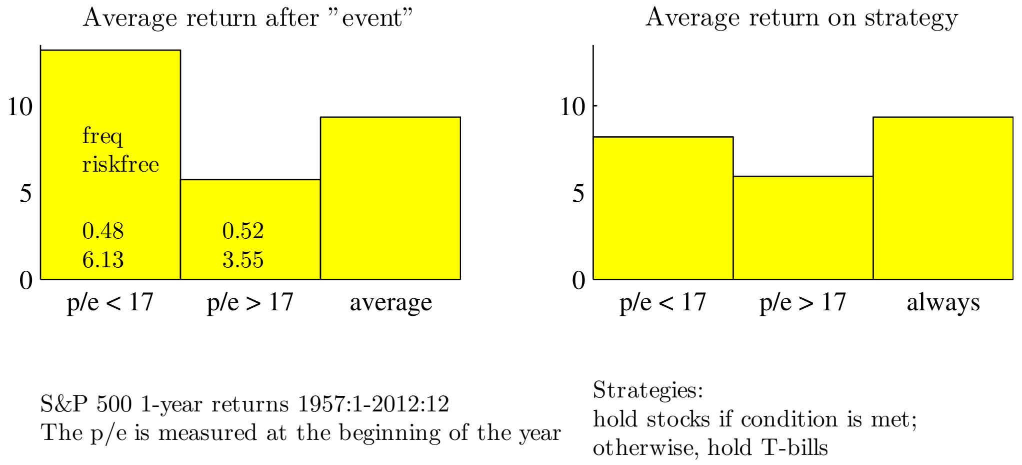


Figure 8.18: Long-run predictability of US stock returns, out-of-sample

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

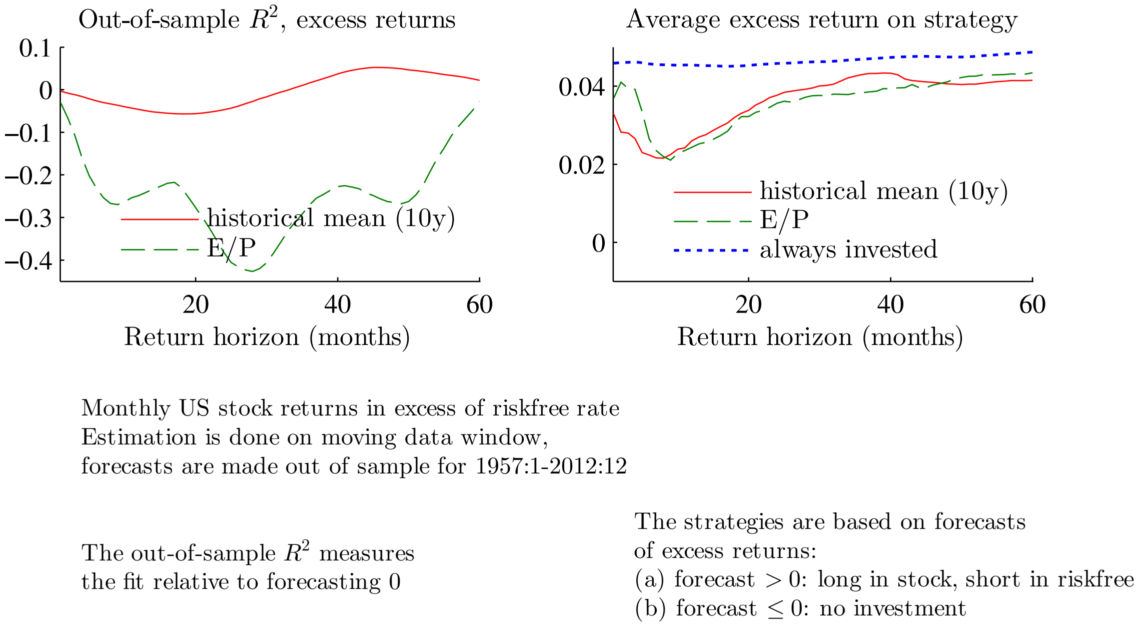


Figure 8.19: Long-run predictability of US stock returns, out-of-sample

efficient: prices react somewhat slowly and predictably to news. The logic is essentially 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, the technical analysis amounts to plotting different transformations (for instance, a moving average) of prices—and to spot known patterns. This section summarizes some simple trading rules that are used.

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 event signals a new upward trend. Let S (L) be the lag order of a short (long) moving average, with

[1] In physics, momentum equals the mass times speed.

$S < L$ and let b be a bandwidth (perhaps 0.01). Then, a MA rule for period t could be

$$\begin{bmatrix} \text{buy in } t \text{ if} & MA_{t-1}(S) > MA_{t-1}(L)(1+b) \\ \text{sell in } t \text{ if} & MA_{t-1}(S) < MA_{t-1}(L)(1-b) \\ \text{no change} & \text{otherwise} \end{bmatrix}, \text{ where} \tag{8.17}$$

$$MA_{t-1}(S) = (p_{t-1} + \ldots + p_{t-S})/S.$$

The difference between the two moving averages is called an *oscillator*

$$\text{oscillator}_t = MA_t(S) - MA_t(L), \tag{8.18}$$

(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 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

[2] Yes, the rumour is true: the tribe of chartists is on the verge of developing their very own language.

[3] Not to be confused with relative strength, which typically refers to the ratio of two different asset prices (for instance, an equity compared to the market).

window of length L. With a bandwidth b (perhaps 0.01), the rule for period t could be

$$\begin{bmatrix} \text{buy in } t \text{ if} & P_t > M_{t-1}(1+b) \\ \text{sell in } t \text{ if} & P_t < m_{t-1}(1-b) \\ \text{no change} & \text{otherwise} \end{bmatrix}, \text{ where} \tag{8.19}$$

$$M_{t-1} = \max(p_{t-1}, \ldots, p_{t-S})$$
$$m_{t-1} = \min(p_{t-1}, \ldots, p_{t-S}).$$

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 8.20 and Table 8.1.

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.

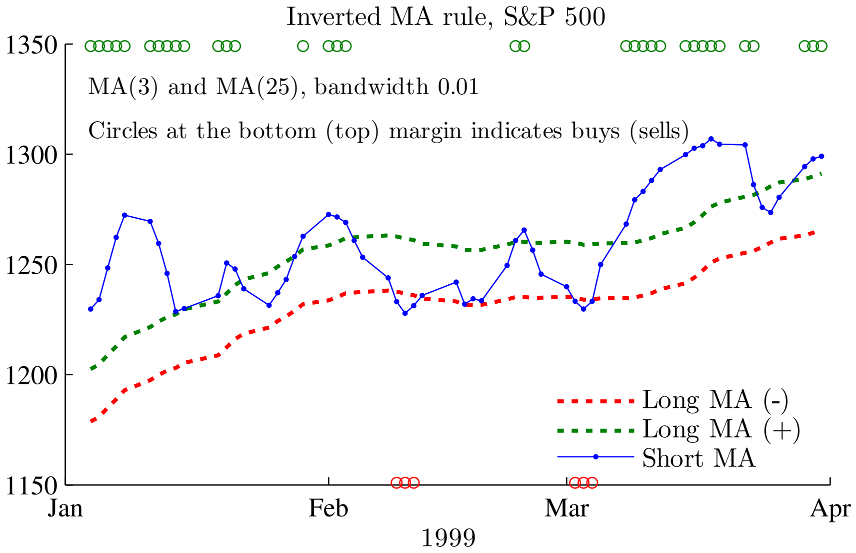


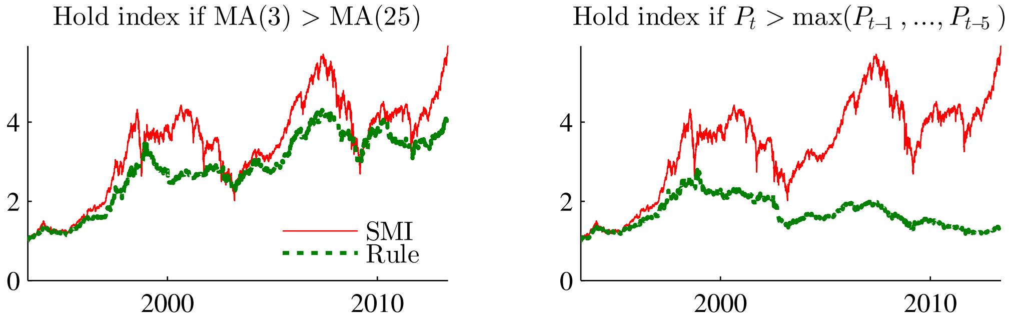
Figure 8.20: Examples of trading rules

	Mean	Std
All days	0.032	1.165
After buy signal	0.054	1.716
After neutral signal	0.047	0.943
After sell signal	0.007	0.903

Table 8.1: Returns (daily, in %) from technical trading rule (Inverted MA rule). Daily S&P 500 data 1990:1-2013:4

8.4 Security Analysts

Reference: Makridakis, Wheelwright, and Hyndman (1998) 10.1 and Elton, Gruber, Brown, and Goetzmann (2010) 26



Daily SMI data
Weekly rebalancing: hold index *or* riskfree

Figure 8.21: Examples of trading rules

8.4.1 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.)

8.4.2 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, for instance no exaggeration).

The main findings are as follows. 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.

8.4.3 High-Frequency Trading Based on Recommendations from Stock Analysts

Barber, Lehavy, McNichols, and Trueman (2001) give a somewhat different picture. They focus on the profitability of a trading strategy based on analyst's recommendations. They use a huge data set (some 360,000 recommendations, US stocks) for the period 1985-1996. They sort stocks in to 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 advise by 30 professional "market timers."

8.4.4 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) show that the economic experts that participate in the semi-annual Livingston survey (mostly bank economists) *(ii)* 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).

8.4.5 The Characteristics of Individual Analysts' Forecasts in Europe

Bolliger (2001) studies the forecast accuracy (earnings per share) of European (13 countries) analysts for the period 1988–1999. In all, some 100,000 forecasts are studied. It is found that the forecast accuracy is positively related to how many times an analyst has forecasted that firm and also (surprisingly) to how many firms he/she forecasts. The accuracy is negatively related to the number of countries an analyst forecasts and also to the size of the brokerage house he/she works for.

8.4.6 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 ever 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.

8.4.7 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.

8.4.8 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.

8.4.9 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.

Bibliography

Ang, J. S., and S. J. Ciccone, 2001, "International differences in analyst forecast properties," mimeo, Florida State University.

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.

Bodie, Z., A. Kane, and A. J. Marcus, 2002, *Investments*, McGraw-Hill/Irwin, Boston, 5th edn.

Bolliger, G., 2001, "The characteristics of individual analysts' forecasts in Europe," mimeo, University of Neuchatel.

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.

Cochrane, J. H., 2001, *Asset pricing*, Princeton University Press, Princeton, New Jersey.

Cuthbertson, K., 1996, *Quantitative financial economics*, Wiley, Chichester, England.

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.

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.

Ferson, W. E., S. Sarkissian, and T. T. Simin, 2003, “Spurious regressions in financial economics,” *Journal of Finance*, 57, 1393–1413.

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.

Granger, C. W. J., 1992, “Forecasting stock market prices: lessons for forecasters,” *International Journal of Forecasting*, 8, 3–13.

Huberman, G., and S. Kandel, 1987, “Mean-variance spanning,” *Journal of Finance*, 42, 873–888.

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.

Makridakis, S., S. C. Wheelwright, and R. J. Hyndman, 1998, *Forecasting: methods and applications*, Wiley, New York, 3rd edn.

Murphy, J. J., 1999, *Technical analysis of the financial markets*, New York Institute of Finance.

Neely, C. J., 1997, "Technical analysis in the foreign exchange market: a layman's guide," *Federal Reserve Bank of St. Louis Review*.

Priestley, M. B., 1981, *Spectral analysis and time series*, Academic Press.

Söderlind, P., 2010, "Predicting stock price movements: regressions versus economists," *Applied Economics Letters*, 17, 869–874.

The Economist, 1993, "Frontiers of finance," pp. 5–20.

9 Maximum Likelihood Estimation

Reference: Verbeek (2008) 2 and 4
More advanced material is denoted by a star (*). It is not required reading.

9.1 Maximum Likelihood

A different route to create a estimate is to maximize the likelihood function.

To understand the principle of maximum likelihood estimation, consider the following examples.

9.1.1 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

$$L = \operatorname{pdf}(x_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x_t-\mu)^2}{\sigma^2}\right], \tag{9.1}$$

where L stands for "likelihood." The basic idea of maximum likelihood estimation (MLE) is to pick model parameters to make the observed data have the highest possible probability. Here this gives $\hat{\mu} = x_t$. This is the maximum likelihood estimator in this example.

What if there are T observations, $x_1, x_2, ... x_T$? 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, $\operatorname{pdf}(x_i, x_j) = \operatorname{pdf}(x_i)\operatorname{pdf}(x_j)$) so

$$\begin{aligned} L &= \operatorname{pdf}(x_1) \times \operatorname{pdf}(x_2) \times ... \times \operatorname{pdf}(x_T) && (9.2)\\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2}\left(\frac{(x_1-\mu)^2}{\sigma^2} + \frac{(x_2-\mu)^2}{\sigma^2} + ... + \frac{(x_T-\mu)^2}{\sigma^2}\right)\right] && (9.3)\end{aligned}$$

Take logs (log likelihood)

$$\ln L = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_T-\mu)^2\right]. \tag{9.4}$$

The derivative with respect to μ is

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\left[(x_1 - \mu) + (x_2 - \mu) + ... + (x_T - \mu)\right]. \tag{9.5}$$

To maximize the likelihood function find the value of $\hat{\mu}$ that makes $\partial \ln L/\partial \mu = 0$, which is the usual sample average

$$\hat{\mu} = (x_1 + x_2 + ... + x_T)/T. \tag{9.6}$$

Remark 9.1 *(Coding the log likelihood function) Many software packages want just the likelihood contribution of data point t (not the full sample). Here it is* $\ln L_t = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_t - \mu)^2$.

9.1.2 Example: Estimating the Variance with ML*

To estimate the variance, use (9.4) and find the value σ^2 that makes $\partial \ln L/\partial \sigma^2 = 0$

$$\begin{aligned} 0 &= \frac{\partial \ln L}{\partial \sigma^2} \\ &= -\frac{T}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{1}{2(\sigma^2)^2}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_T - \mu)^2\right], \end{aligned} \tag{9.7}$$

so

$$\hat{\sigma}^2 = \frac{1}{T}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_T - \mu)^2\right]. \tag{9.8}$$

Notice that we divide by T, not by $T - 1$, so $\hat{\sigma}^2$ must be biased, but the bias disappears as $T \to \infty$

9.1.3 MLE of a Regression

To apply this idea to a (multiple) regression model

$$y_t = \beta' x_t + u_t, \tag{9.9}$$

we could assume that u_t is iid $N(0, \sigma^2)$. The probability density function of u_t is

$$\text{pdf}(u_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}u_t^2/\sigma^2\right). \tag{9.10}$$

Since the errors are independent, we get the joint pdf of the $u_1, u_2, \ldots, u_T$ by multiplying the marginal pdfs of each of the errors

$$\begin{aligned} L &= \text{pdf}\,(u_1) \times \text{pdf}\,(u_2) \times ... \times \text{pdf}\,(u_T) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2}\left(\frac{u_1^2}{\sigma^2} + \frac{u_2^2}{\sigma^2} + ... + \frac{u_T^2}{\sigma^2}\right)\right]. \end{aligned} \tag{9.11}$$

Substitute $y_t - \beta' x_t$ for u_t and take logs to get the log likelihood function of the sample

$$\ln L = \sum\nolimits_{t=1}^{T} \ln L_t, \text{ where} \tag{9.12}$$

$$\ln L_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2}\left(y_t - \beta' x_t\right)^2/\sigma^2. \tag{9.13}$$

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 9.1.

Maximum likelihood estimators have 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).

Example 9.2 *Consider the regression model $y_i = \beta_1 x_i + u_i$, where we (happen to) know that $u_i \sim N(0,1)$.Suppose we have the following data*

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

Suppose $(y_1, x_1) = (-1.5, -1)$. Try different values of β_2 on observation 1

β_2	u_1	*Density function value of u_1*
1.6	$-1.5 - \mathbf{1.6} \times (-1) = 0.1$	0.40
1.8	$-1.5 - \mathbf{1.8} \times (-1) = 0.3$	0.38
2.0	$-1.5 - \mathbf{2.0} \times (-1) = 0.5$	0.35

Observation 1 favours $\beta = 1.6$; see Figure 9.2. Do the same for observations 2 and 3:

β_2	u_2	*Density function value of u_2*
1.6	$-0.6 - \mathbf{1.6} \times 0 = -0.6$	0.33
1.8	$-0.6 - \mathbf{1.8} \times 0 = -0.6$	0.33
2.0	$-0.6 - \mathbf{2.0} \times 0 = -0.6$	0.33

β_2	u_3	*Density function value of u_3*
1.6	$2.1 - \mathbf{1.6} \times 1 = 0.5$	0.35
1.8	$2.1 - \mathbf{1.8} \times 1 = 0.3$	0.38
2.0	$2.1 - \mathbf{2.0} \times 1 = 0.1$	0.40

To sum up, observation 1 favours $\beta = 1.6$, observation 2 is neutral, and observation 3 favours $\beta = 2$. The estimate is a "compromise" that maximises the joint density (the product of the individual densities since the u_i are independent)

β_2	$\text{pdf}(u_1) \times \text{pdf}(u_2) \times \text{pdf}(u_3)$
1.6	$0.40 \times 0.33 \times 0.35 \approx 0.047$
1.8	$0.38 \times 0.33 \times 0.38 \approx 0.048$
2.0	$0.35 \times 0.33 \times 0.40 \approx 0.047$

so 1.8 *has the highest likelihood value of these three alternatives (it is actually the optimum). See Figure 9.2.*

Example 9.3 *Consider the simple regression where we happen to know that the intercept is zero, $y_t = \beta_1 x_t + u_t$. Suppose we have the following data*

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

Suppose $\beta_2 = 2$, then we get the following values for $u_t = y_t - 2x_t$ and its square

$$\begin{bmatrix} -1.5 - 2 \times (-1) \\ -0.6 - 2 \times 0 \\ 2.1 - 2 \times 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ -0.6 \\ 0.1 \end{bmatrix} \text{ with the square } \begin{bmatrix} 0.25 \\ 0.36 \\ 0.01 \end{bmatrix} \text{ with sum } 0.62.$$

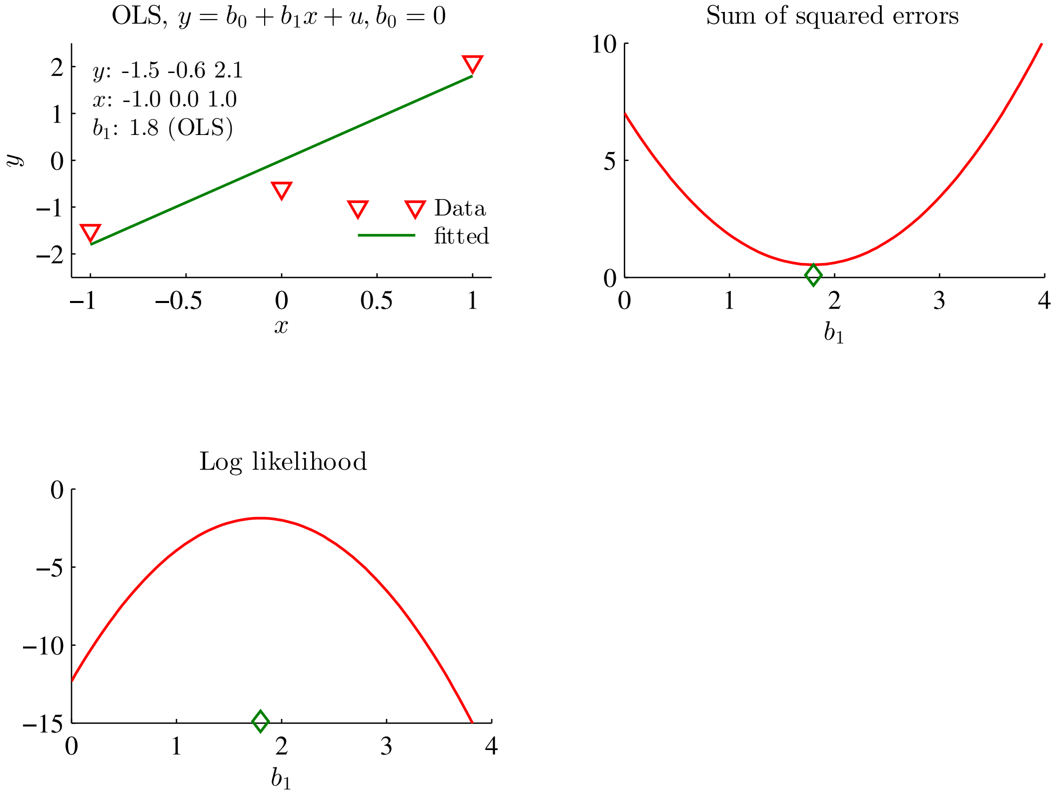


Figure 9.1: Example of OLS and ML estimation

Now, suppose instead that $\beta_2 = 1.8$*, then we get*

$$\begin{bmatrix} -1.5 - 1.8 \times (-1) \\ -0.6 - 1.8 \times 0 \\ 2.1 - 1.8 \times 1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ -0.6 \\ 0.3 \end{bmatrix} \textit{ with the square } \begin{bmatrix} 0.09 \\ 0.36 \\ 0.09 \end{bmatrix} \textit{ with sum } 0.54.$$

The latter choice of β_2 *will certainly give a larger value of the likelihood function (it is actually the optimum). See Figure 9.1.*

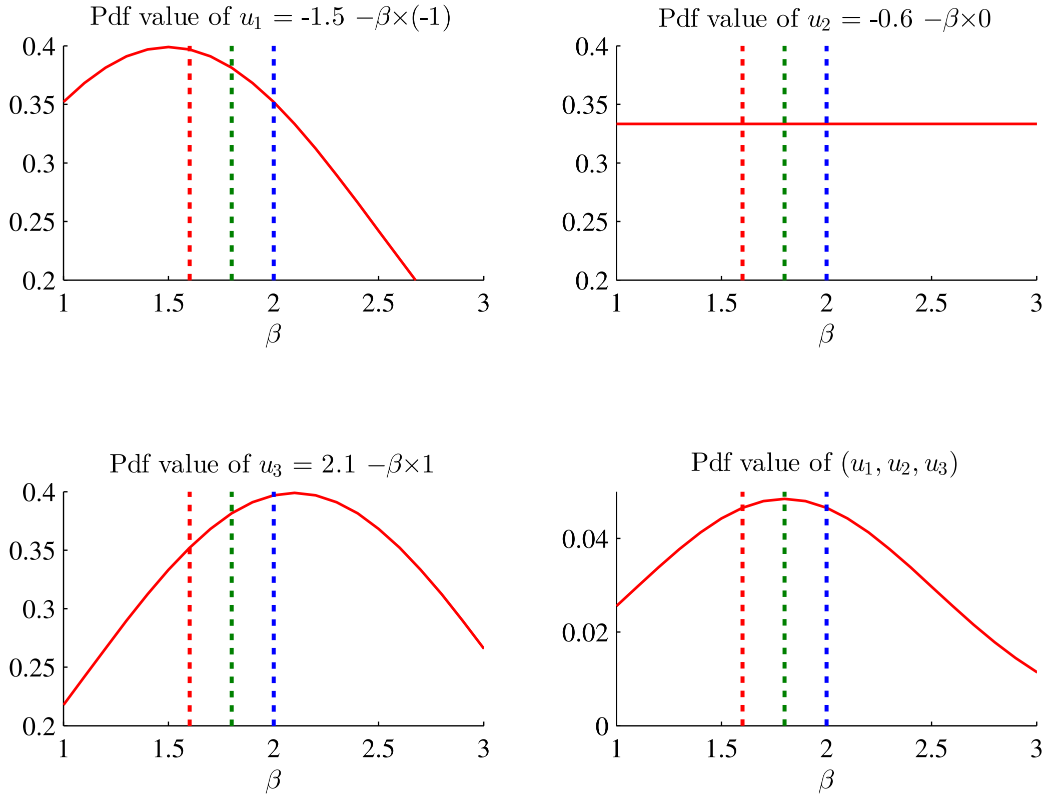


Figure 9.2: Example of OLS and ML estimation

9.1.4 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

$$y_t = b'x_t + u_t, \text{ where } u_t \text{ is } N(0, \sigma_t^2). \tag{9.14}$$

The variance follows the GARCH(1,1) process

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{9.15}$$

(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 optimization routine to maximize the log likelihood function

$$\ln L = \sum_{t=1}^{T} L_t, \text{ where } L_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma_t^2 - \frac{1}{2}\frac{u_t^2}{\sigma_t^2}. \tag{9.16}$$

This means, in effect, that 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.

Remark 9.4 *To perform the estimation, we also need to supply the optimization routine with a starting value for* σ_1^2 *and make sure that the restrictions on the GARCH parameters are fulfilled.*

9.2 Key Properties of MLE

No general results on small-sample properties of MLE: can be biased or not...

MLE have very nice asymptotic (large-sample) properties, provided we maximize the right likelihood function. If so, then

1. MLE is consistent (Pr($|\hat{\beta} - \beta| >$ any number) gets very small as T gets large)

2. MLE is the most efficient/precise estimator, at least asymptotically (efficient = smallest variance)

3. MLE estimates ($\hat{\theta}$) are normally distributed,

$$\sqrt{T}(\hat{\theta} - \theta) \to^d N(0, V), \tag{9.17}$$

$$V = I(\theta)^{-1} \text{ with } I(\theta) = -\operatorname{E}\frac{\partial^2 \ln L}{\partial\theta\partial\theta}/T. \tag{9.18}$$

   ($I(\theta)$ is called the "information matrix"). The information matrix can also be written $I(\theta) = -\operatorname{E}\frac{\partial^2 \log L_t}{\partial\theta\partial\theta}$, where $\ln L_t$ 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).

9.2.1 Example of the Information Matrix

Differentiate (9.5) (and assume we know σ^2) to get

$$\frac{\partial^2 \ln L}{\partial\mu\partial\mu} = -\frac{T}{\sigma^2}. \tag{9.19}$$

The information matrix is

$$I(\theta) = -\mathrm{E}\,\frac{\partial^2 \ln L}{\partial\theta\partial\theta}/T = \frac{1}{\sigma^2}, \tag{9.20}$$

which we combine with (9.17)–(9.18) to get

$$\sqrt{T}(\hat{\mu} - \mu) \to^d N(0, \sigma^2) \text{ or } \hat{\mu} \to^d N(\mu, \sigma^2/T). \tag{9.21}$$

This is the standard expression for the distribution of a sample average.

9.3 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 $\ln L(\hat{\theta}) - \ln L(\text{“}\hat{\theta}$ with H_0 restrictions”$) = 0$. Example: compare the R^2 from a model without and with a restriction that some coefficient equals 1/3

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”

9.4 QMLE*

A MLE based on the wrong likelihood function (distribution) may still be useful.

Suppose we use the likelihood function L and get estimates $\hat{\theta}$ by

$$\frac{\partial \ln L}{\partial\theta} = \mathbf{0} \tag{9.22}$$

If L is wrong, then we are maximizing the wrong thing. With some luck, we still get reasonable (consistent) estimates.

Example 9.5 *(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 $\Sigma_{t=1}^T (y_t - x_t'\hat{\beta})x_t = \mathbf{0}$.*

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

$$\mathrm{E}\,\frac{\partial \ln L}{\partial \theta} = \mathbf{0}, \tag{9.23}$$

then the estimates are still consistent. We are doing *quasi-MLE* (or pseudo-MLE).

With QMLE, $\sqrt{T}(\hat{\theta} - \theta) \to^{d} N(0, V)$, but

$$V = I(\theta)^{-1}\,\mathrm{E}\left[\frac{\partial \ln L_t}{\partial \theta}\left(\frac{\partial \ln L_t}{\partial \theta}\right)'\right] I(\theta)^{-1} \tag{9.24}$$

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.

Bibliography

Verbeek, M., 2008, *A guide to modern econometrics*, Wiley, Chichester, 3rd edn.

10 ARCH and GARCH

Reference: Bodie, Kane, and Marcus (2005) 13.4
Reference (advanced): Taylor (2005) 8–9; Verbeek (2004) 8; Campbell, Lo, and MacKinlay (1997) 12; Franses and van Dijk (2000)

10.1 Heteroskedasticity

10.1.1 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 mean

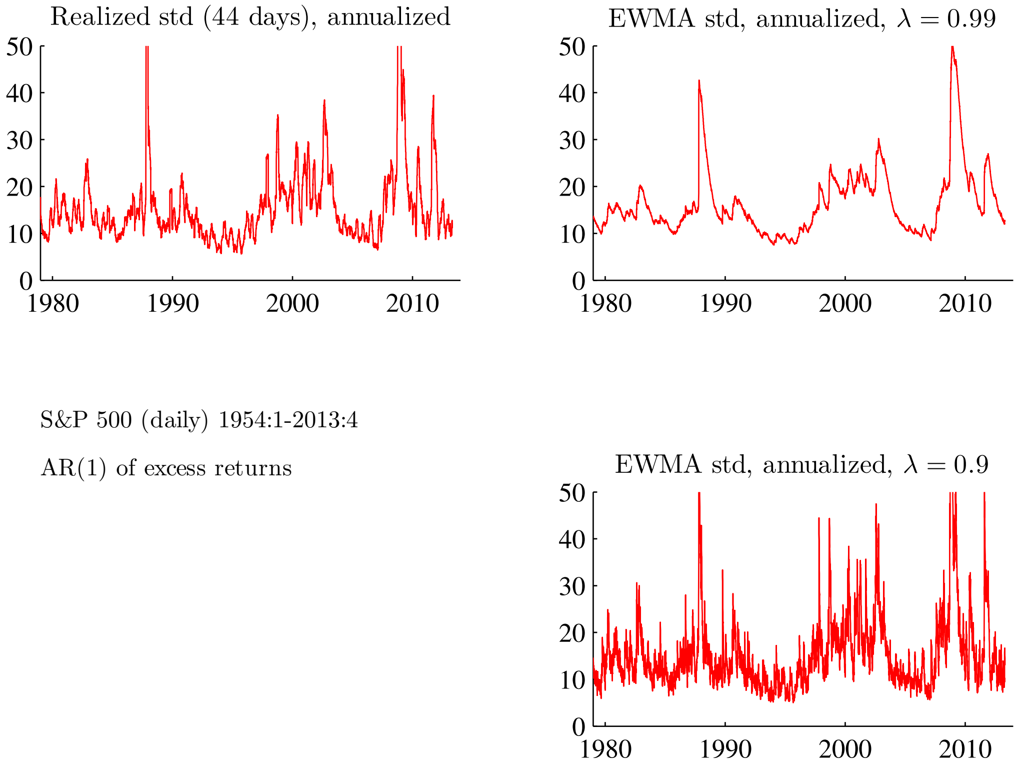
$$\sigma_t^2 = (u_{t-1}^2 + u_{t-2}^2 + \ldots + u_{t-q}^2)/q, \tag{10.1}$$

where the latest q observations are used. 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 both (general) 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 10.1 and 10.2 for examples.

Unfortunately, this method 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 $0 < \lambda < 1$, so

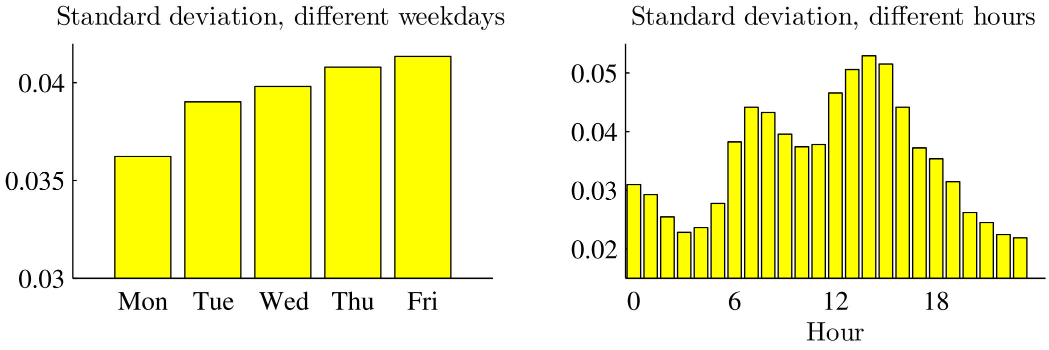
$$\sigma_t^2 = (1-\lambda)(u_{t-1}^2 + \lambda u_{t-2}^2 + \lambda^2 u_{t-3}^2 + \ldots), \tag{10.2}$$



S&P 500 (daily) 1954:1-2013:4

AR(1) of excess returns

Figure 10.1: Standard deviation



5-minute data on EUR/USD changes, 1998:1-2011:11
Sample size: 1045414

Figure 10.2: Standard deviation for EUR/USD exchange rate changes

which can also be calculated in a recursive fashion as

$$\sigma_t^2 = (1-\lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2. \tag{10.3}$$

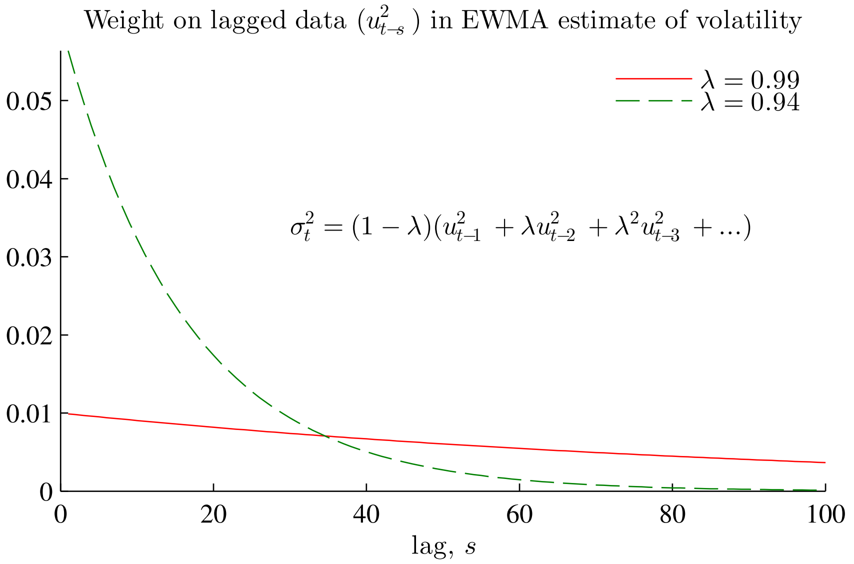


Figure 10.3: Weights on old data in the EMA approach to estimate volatility

The initial value (before the sample) could be assumed to be zero or (perhaps better) the unconditional variance in a historical sample.

This methods 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 $\Sigma_{t=1}^{T}(u_t^2 - \sigma_t^2)^2$.

See Figure 10.3 for an illustration of the weights.

10.1.2 Predicting Realised Volatility

Volatility is often predictable, at least for horizons up to a couple of months. See Tables 10.1–10.2 for examples of very simple prediction equations.

10.1.3 Heteroskedastic Residuals in a Regression

Suppose we have a regression model

$$y_t = b_0 + x_{1t}b_1 + x_{2t}b_2 + \cdots + x_{kt}b_k + \varepsilon_t, \text{ where} \tag{10.4}$$
$$\mathrm{E}\,\varepsilon_t = 0 \text{ and } \mathrm{Cov}(x_{it}, \varepsilon_t) = 0.$$

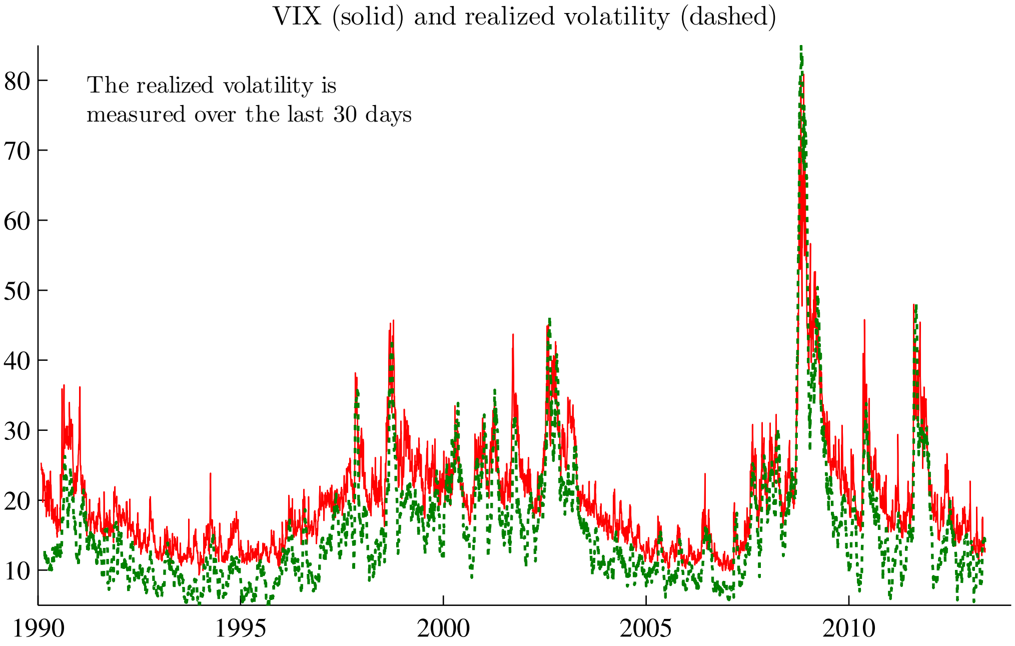


Figure 10.4: VIX and realized volatility (variance)

	(1)	(2)	(3)
lagged RV	0.75		0.26
	(11.02)		(2.20)
lagged VIX		0.91	0.64
		(12.60)	(7.54)
constant	3.97	−2.62	−1.20
	(4.29)	(−2.06)	(−1.55)
R2	0.56	0.61	0.62
obs	5825.00	5845.00	5825.00

Table 10.1: Regression of 22-day realized S&P return volatility 1990:1-2013:5. 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.

In the standard case we assume that ε_t is iid (independently and identically distributed), which rules out 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

	RV(EUR)	RV(GBP)	RV(CHF)	RV(JPY)
lagged RV(EUR)	0.63			
	(6.94)			
lagged RV(GBP)		0.72		
		(10.07)		
lagged RV(CHF)			0.33	
			(2.50)	
lagged RV(JPY)				0.56
				(4.92)
constant	0.06	0.04	0.25	0.13
	(1.71)	(1.40)	(3.36)	(1.93)
D(Tue)	0.12	0.08	0.13	0.11
	(10.60)	(6.01)	(4.00)	(3.37)
D(Wed)	0.11	0.09	0.08	0.13
	(8.60)	(6.56)	(3.64)	(4.09)
D(Thu)	0.12	0.09	0.13	0.15
	(9.25)	(5.38)	(6.26)	(3.40)
D(Fri)	0.13	0.07	0.14	0.12
	(6.13)	(3.83)	(8.47)	(3.81)
R2	0.40	0.52	0.11	0.31
obs	3629.00	3629.00	3629.00	3629.00

Table 10.2: Regression of daily realized variance 1998:1-2011:11. 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.

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 (except in a special case, see below) not correct. This is illustrated in Figure 10.5.

There are two ways to handle this problem. First, we could use some other estimation method than LS that incorporates the structure of the heteroskedasticity. For instance, combining the regression model (10.4) with an ARCH structure of the residuals—and estimate the whole thing with maximum likelihood (MLE) is one way. 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

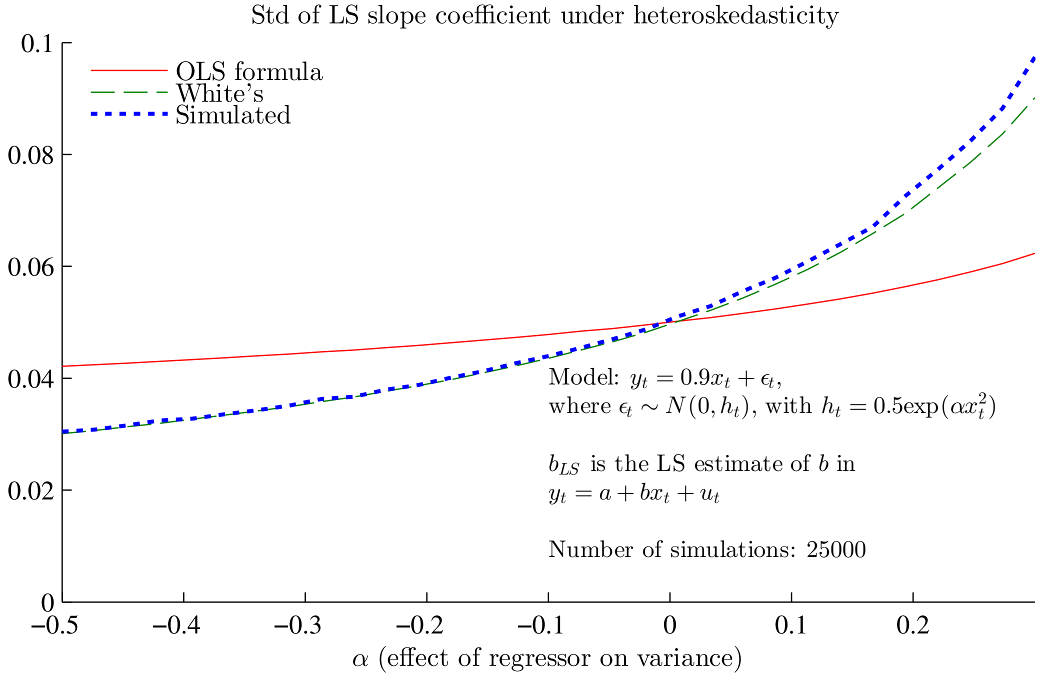


Figure 10.5: Variance of OLS estimator, heteroskedastic errors

covariance matrix" is the most common.

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 residual is uncorrelated with w_t, then the usual LS covariance 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_i be the squares and cross products of the regressors. For instance, if the regressors include $(1, x_{1t}, x_{2t})$ then w_t is the vector $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$—since $(1, x_{1t}, x_{2t}) \times 1$ is $(1, x_{1t}, x_{2t})$ and $1 \times 1 = 1$. The test is then to run a regression of squared fitted residuals on w_t

$$\hat{\varepsilon}_t^2 = w_t'\gamma + v_i, \tag{10.5}$$

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

10.1.4 Autoregressive Conditional Heteroskedasticity (ARCH)

Autoregressive heteroskedasticity is a special form of heteroskedasticity—and it is often found in financial data which shows volatility clustering (calm spells, followed by volatile spells, followed by...).

To test for ARCH features, *Engle's test of ARCH* is perhaps the most straightforward. It amounts to running an AR(q) regression of the squared zero-mean variable (here denoted u_t)

$$u_t^2 = \omega + a_1 u_{t-1}^2 + \ldots + a_q u_{t-q}^2 + v_t, \tag{10.6}$$

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, $TR^2 \sim \chi_q^2$.) This test can also be applied to the fitted residuals from a regression like (10.4). However, in this case, it is not obvious that ARCH effects make the standard expression for the LS covariance matrix invalid—this is tested by White's test as in (10.5).

10.2 ARCH Models

This section discusses the Autoregressive Conditional Heteroskedasticity (ARCH) model. It is a model of how volatility depends on recent volatility.

There are two basic reasons for being interested in an ARCH model. First, if residuals of the regression model (10.4) 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 option pricing.

10.2.1 Properties of ARCH(1)

In the ARCH(1) model the residual in the regression equation (10.4), or some other zero-mean variable, can be written

$$u_t \sim N(0, \sigma_t^2), \text{ with} \tag{10.7}$$

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2, \text{ with } \omega > 0 \text{ and } 0 \leq \alpha < 1. \tag{10.8}$$

The non-negativity restrictions on ω and α are needed in order to guarantee $\sigma_t^2 > 0$. The upper bound $\alpha < 1$ is needed in order to make the conditional variance stationary (more later).

It is clear that the unconditional distribution of u_t is non-normal. While the conditional distribution of u_t is $N(0, \sigma_t^2)$, the unconditional distribution of u_t is a mixture of normal distributions with different (and random) variances. It can be shown that the result is a distribution which has 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 (since σ_{t+1}^2 is known in t.)

$$\mathrm{E}_t\, \sigma_{t+s}^2 = \bar{\sigma}^2 + \alpha^{s-1} \left(\sigma_{t+1}^2 - \bar{\sigma}^2\right), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1-\alpha}, \tag{10.9}$$

where $\bar{\sigma}^2$ is the unconditional variance. The conditional volatility behaves like an AR(1), and $0 \leq \alpha < 1$ is necessary to keep it positive and stationary.

See Figure 10.6 for an illustration of the fitted volatilities.

Proof. (of (10.9)) Notice that $\mathrm{E}_t\, \sigma_{t+2}^2 = \omega + \alpha\, \mathrm{E}_t\, v_{t+1}^2\, \mathrm{E}_t\, \sigma_{t+1}^2$ since v_t is independent of σ_t. Morover, $\mathrm{E}_t\, v_{t+1}^2 = 1$ and $\mathrm{E}_t\, \sigma_{t+1}^2 = \sigma_{t+1}^2$ (known in t). Combine to get $\mathrm{E}_t\, \sigma_{t+2}^2 = \omega + \alpha\sigma_{t+1}^2$. Similarly, $\mathrm{E}_t\, \sigma_{t+3}^2 = \omega + \alpha\, \mathrm{E}_t\, \sigma_{t+2}^2$. Substitute for $\mathrm{E}_t\, \sigma_{t+2}^2$ to get $\mathrm{E}_t\, \sigma_{t+3}^2 = \omega + \alpha(\omega + \alpha\sigma_{t+1}^2)$, which can be written as (10.9). Further periods follow the same pattern. ■

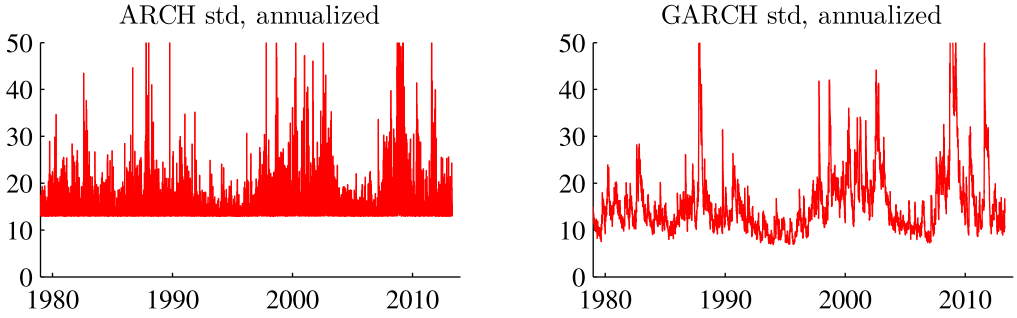
10.2.2 Estimation of the ARCH(1) Model

The most common way to estimate the model is to assume that v_t ~iid $N(0, 1)$ and to set up the likelihood function. The log likelihood is easily found, since the model is conditionally Gaussian. It is

$$\ln L = \sum_{t=1}^{T} L_t, \text{ where } L_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma_t^2 - \frac{1}{2}\frac{u_t^2}{\sigma_t^2}. \tag{10.10}$$

The estimates are found by maximizing the likelihood function (by choosing the parameters). This is done by a numerical optimization routine, which should preferably impose the constraints in (10.8).

If u_t is just a zero-mean variable (no regression equation), then this just amounts to



S&P 500 (daily) 1954:1-2013:4

AR(1) of excess returns
with ARCH(1) or GARCH(1,1) errors

AR(1) coef: 0.09
ARCH coef: 0.31
GARCH coefs: 0.08 0.91

Figure 10.6: ARCH and GARCH estimates

choosing the parameters (ω and α) in (10.8). Instead, if u_t is a residual from a regression equation (10.4), then we instead need to choose both the regression coefficients ($b_0, ..., b_k$) in (10.4) and the parameters (ω and α) in (10.8). In either case, we need a starting value of $\sigma_1^2 = \omega + \alpha u_0^2$. This most common approach is to use the first observation as a "starting point," that is, we actually have a sample from ($t =$) 0 to T, but observation 0 is only used to construct a starting value of σ_1^2, and only observations 1 to T are used in the calculation of the likelihood function value.

Notice that if we estimate a regression function and an ARCH model simultaneous with MLE, then we automatically get the right standard errors of the regression coefficients from the information matrix. There is no need for using any adjusted ("White") values.

Remark 10.1 *(Regression with ARCH(1) residuals) To estimate the full model (10.4) and (10.8) by ML, we can do as follows.*
First, guess values of the parameters $b_0, ..., b_k$, and ω, and α. The guess of $b_0, ..., b_k$ can be taken from an LS estimation of (10.4), and the guess of ω and α from an LS estimation of $\hat{\varepsilon}_t^2 = \omega + \alpha\hat{\varepsilon}_{t-1}^2 + \varepsilon_t$ where $\hat{\varepsilon}_t$ are the fitted residuals from the LS estimation of (10.4).

Second, loop over the sample (first $t = 1$, then $t = 2$, etc.) and calculate $u_t = \hat{\varepsilon}_t$ from (10.4) and σ_t^2 from (10.8). Plug in these numbers in (10.10) 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 10.2 *(Imposing parameter constraints on ARCH(1).) To impose the restrictions in (10.8) when the previous remark is implemented, iterate over values of $(b, \tilde{\omega}, \tilde{\alpha})$ and let $\omega = \tilde{\omega}^2$ and $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha})]$.*

It is sometimes found that the standardized values of u_t, u_t/σ_t, still have too fat tails compared with $N(0, 1)$. This would violate the assumption about a normal distribution in (10.10). 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).

It is straightforward to add more lags to (10.8). For instance, an ARCH(p) would be

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \ldots + \alpha_p u_{t-p}^2. \tag{10.11}$$

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 $\Sigma_{j=1}^{p} \alpha_j \leq 1$.

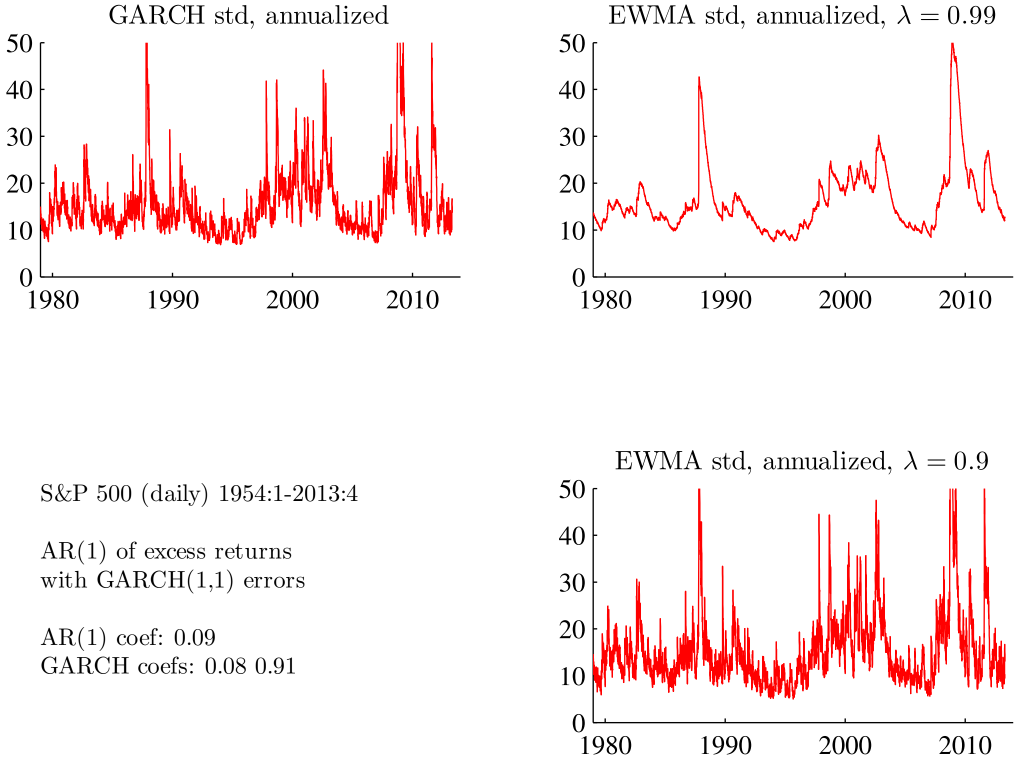
10.3 GARCH Models

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

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2, \text{with} \tag{10.12}$$
$$\omega > 0; \alpha, \beta \geq 0; \text{and } \alpha + \beta < 1.$$

The non-negativity restrictions are needed in order to guarantee that $\sigma_t^2 > 0$ in all periods. The upper bound $\alpha + \beta < 1$ is needed in order to make the σ_t^2 stationary and therefore the unconditional variance finite.

Remark 10.3 *The GARCH(1,1) has many similarities with the exponential moving average estimator of volatility (10.3). The main differences are that the exponential moving*

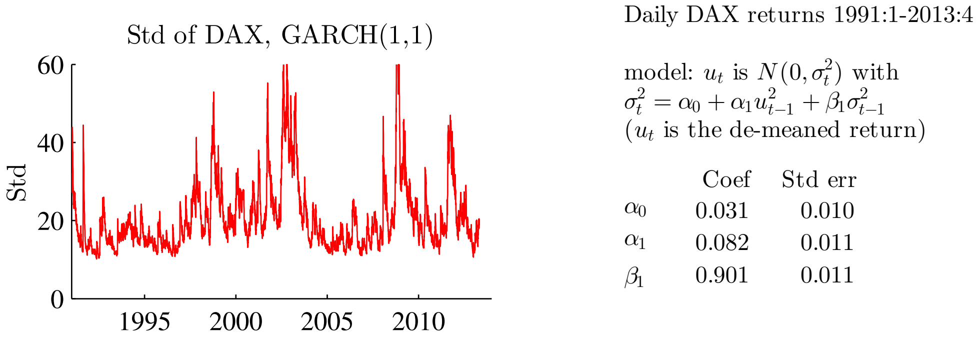


S&P 500 (daily) 1954:1-2013:4

AR(1) of excess returns
with GARCH(1,1) errors

AR(1) coef: 0.09
GARCH coefs: 0.08 0.91

Figure 10.7: Conditional standard deviation, estimated by GARCH(1,1) model



Daily DAX returns 1991:1-2013:4

model: u_t is $N(0, \sigma_t^2)$ with
$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2$
(u_t is the de-meaned return)

	Coef	Std err
α_0	0.031	0.010
α_1	0.082	0.011
β_1	0.901	0.011

Figure 10.8: Results for a univariate GARCH model

average does not have a constant and volatility is non-stationary (the coefficients sum to unity).

See Figure 10.7 for an example.

The GARCH(1,1) corresponds to an ARCH(∞) with geometrically declining weights, which is seen by solving (10.12) recursively by substituting for σ_{t-1}^2 (and then σ_{t-2}^2, σ_{t-3}^2, ...)

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha \sum_{j=0}^{\infty} \beta^j u_{t-1-j}^2. \tag{10.13}$$

This suggests that a GARCH(1,1) might be a reasonable approximation of a high-order ARCH.

Proof. (of (10.13)) Substitute for σ_{t-1}^2 in (10.12), and then for σ_{t-2}^2, etc

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha u_{t-1}^2 + \overbrace{\beta\left(\omega + \alpha u_{t-2}^2 + \beta \sigma_{t-2}^2\right)}^{\sigma_{t-1}^2} \\ &= \omega\left(1+\beta\right) + \alpha u_{t-1}^2 + \beta \alpha u_{t-2}^2 + \beta^2 \sigma_{t-2}^2 \\ &= \vdots\end{aligned}$$

and we get (10.13). ■

Also, the GARCH(1,1) model implies that we in period t can forecast the future conditional variance (σ_{t+s}^2) as

$$\mathrm{E}_t\, \sigma_{t+s}^2 = \bar{\sigma}^2 + (\alpha+\beta)^{s-1}\left(\sigma_{t+1}^2 - \bar{\sigma}^2\right), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}, \tag{10.14}$$

which is of the same form as for the ARCH model (10.9), but where the sum of α and β is like an AR(1) parameter.

Proof. (of (10.14)) Notice that $\mathrm{E}_t\, \sigma_{t+2}^2 = \omega + \alpha\, \mathrm{E}_t\, v_{t+1}^2\, \mathrm{E}_t\, \sigma_{t+1}^2 + \beta \sigma_{t+1}^2$ since v_t is independent of σ_t. Morover, $\mathrm{E}_t\, v_{t+1}^2 = 1$ and $\mathrm{E}_t\, \sigma_{t+1}^2 = \sigma_{t+1}^2$ (known in t). Combine to get $\mathrm{E}_t\, \sigma_{t+2}^2 = \omega + (\alpha+\beta)\sigma_{t+1}^2$. Similarly, $\mathrm{E}_t\, \sigma_{t+3}^2 = \omega + (\alpha+\beta)\, \mathrm{E}_t\, \sigma_{t+2}^2$. Substitute for $\mathrm{E}_t\, \sigma_{t+2}^2$ to get $\mathrm{E}_t\, \sigma_{t+3}^2 = \omega + (\alpha+\beta)[\omega + (\alpha+\beta)\sigma_{t+1}^2]$, which can be written as (10.14). Further periods follow the same pattern. ■

To estimate the model consisting of (10.4) and (10.12) we can still use the likelihood function (10.10) 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 $\sigma_0^2 = \mathrm{Var}(u_t)$.

Remark 10.4 *(Imposing parameter constraints on GARCH(1,1).) To impose the restrictions in (10.12), iterate over values of* $(b, \tilde{\omega}, \tilde{\alpha}, \tilde{\beta})$ *and let* $\omega = \omega^2$*,* $\alpha = \exp(\tilde{\alpha})/[1 +$

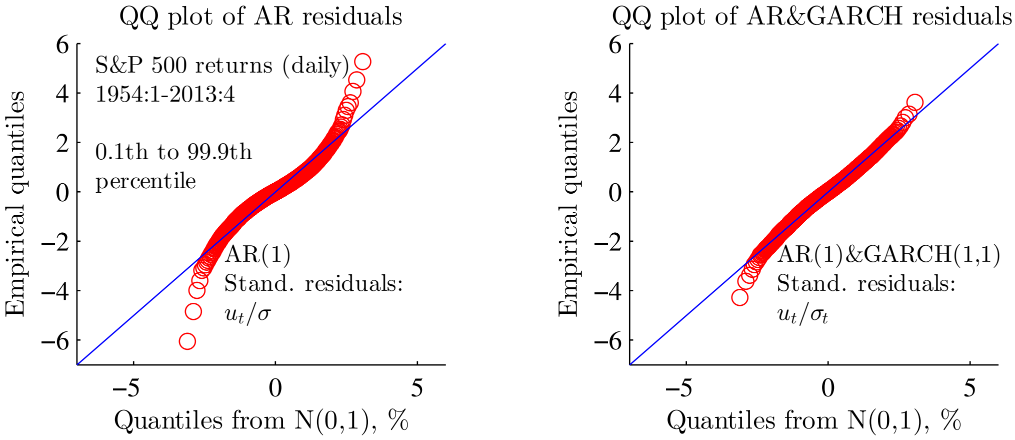


Figure 10.9: QQ-plot of residuals

$\exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$, *and* $\beta = \exp(\tilde{\beta})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$.

See Figure 10.9 for evidence of how the residuals become more normally distributed once the heteroskedasticity is handled.

Remark 10.5 *(Value at Risk) The value at risk (as fraction of the investment) at the* α *level (say,* $\alpha = 0.95$*) is* $VaR_\alpha = -\mathrm{cdf}^{-1}(1-\alpha)$, *where* $\mathrm{cdf}^{-1}()$ *is the inverse of the cdf—so* $\mathrm{cdf}^{-1}(1-\alpha)$ *is the* $1-\alpha$ *quantile of the return distribution. For instance,* $VaR_{0.95} = 0.08$ *says that there is only an 5% chance that the loss will be greater than 8% of the investment. See Figure 10.10 for an illustration. When the return has an* $N(\mu, \sigma^2)$ *distribution, then* $VaR_{95\%} = -(\mu - 1.64\sigma)$. *See Figure 10.11 for an example of time-varying VaR, based on a GARCH model.*

10.4 Non-Linear Extensions

A very large number of extensions have been suggested. I summarize a few of them, which can be estimated by using the likelihood function (10.10) to do a MLE.

An asymmetric GARCH (Glosten, Jagannathan, and Runkle (1993)) can be con-

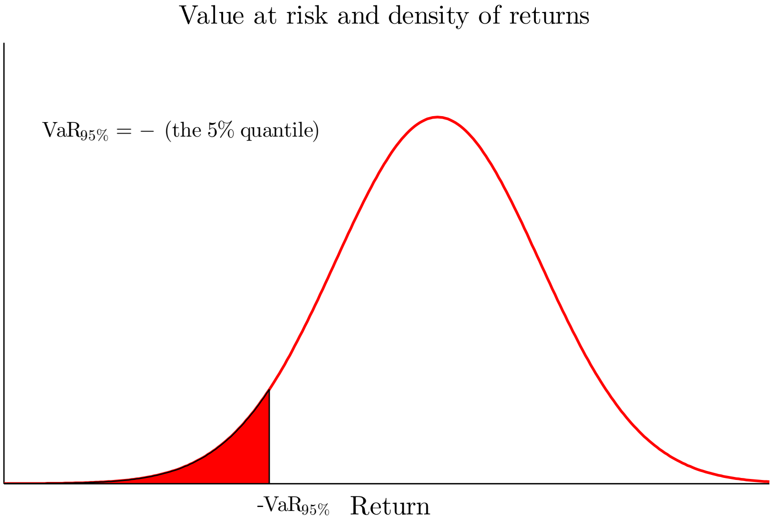


Figure 10.10: Value at risk

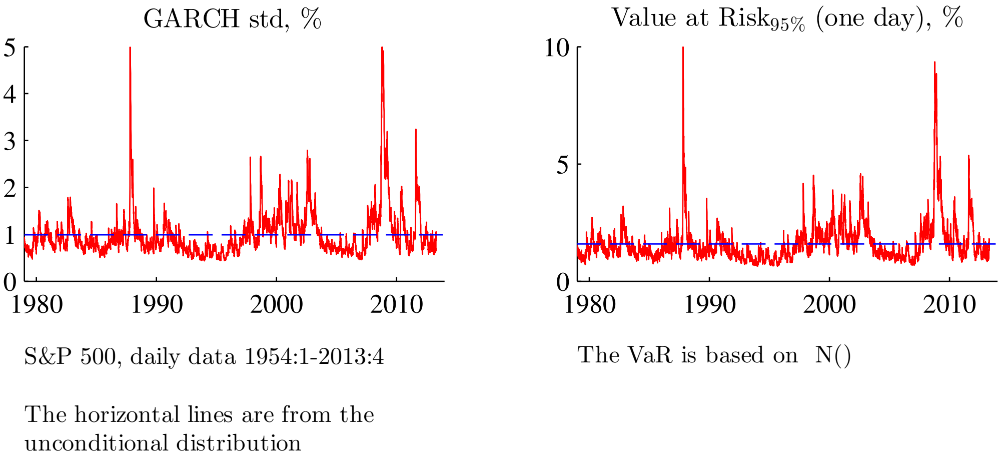


Figure 10.11: Conditional volatility and VaR

structed as

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \delta(u_{t-1} > 0) u_{t-1}^2, \text{ where} \tag{10.15}$$

$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

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

$$\ln \sigma_t^2 = \omega + \alpha \frac{|u_{t-1}|}{\sigma_{t-1}} + \beta \ln \sigma_{t-1}^2 + \gamma \frac{u_{t-1}}{\sigma_{t-1}} \tag{10.16}$$

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, 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}.

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 $|u_{t-1} - b|^{3/2}$.

10.5 (G)ARCH-M

It can make sense to let the conditional volatility enter the mean equation—for instance, as a proxy for risk which may influence the expected return.

We modify the "mean equation" (10.4) to include the conditional variance σ_t^2 (taken from any of the models for heteroskedasticity) as a regressor

$$y_t = x_t' b + \varphi \sigma_t^2 + u_t. \tag{10.17}$$

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 (10.10) to do MLE.

Remark 10.6 *(Coding of (G)ARCH-M) We can use the same approach as in Remark 10.1, except that we use (10.17) instead of (10.4) to calculate the residuals (and that we obviously also need a guess of φ).*

Example 10.7 *(Theoretical motivation of GARCH-M) A mean variance investor solves*

$$\begin{aligned} &\max_{\alpha} \mathrm{E}\, R_p - \sigma_p^2 k/2, \text{ subject to} \\ &R_p = \alpha R_m + (1-\alpha) R_f, \end{aligned}$$

where R_m is the return on the risky asset (the market index) and R_f is the riskfree return. The solution is

$$\alpha = \frac{1}{k} \frac{\mathrm{E}(R_m - R_f)}{\sigma_m^2}.$$

In equilibrium, this weight is one (since the net supply of bonds is zero), so we get

$$\mathrm{E}(R_m - R_f) = k\sigma_m^2,$$

which says that the expected excess return is increasing in both the market volatility and risk aversion (k).

10.6 Multivariate (G)ARCH

10.6.1 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

$$\Sigma_t = \mathrm{E}_{t-1}\, u_t u_t'. \tag{10.18}$$

Remark 10.8 *(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. Although we only need to care about the unique elements of Σ_t, that is, vech(Σ_t), this still gives very many parameters

$$\text{vech}(\Sigma_t) = C + A\text{vech}(u_{t-1}u'_{t-1}) + B\text{vech}(\Sigma_{t-1}). \tag{10.19}$$

For instance, with $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = C + A \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1}u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + B \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \tag{10.20}$$

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 (compare the restrictions of positive coefficients in (10.12)).

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

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1}u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \tag{10.21}$$

which gives $3 + 3 + 3 = 9$ parameters (in C, A, and B, respectively). 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* assumes that every variance follows a univariate GARCH process and that the conditional correlations are constant. With $n = 2$ the covariance matrix is

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \tag{10.22}$$

and each of σ_{11t} and σ_{22t} follows a GARCH process. Assuming a GARCH(1,1) as in (10.12) 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 10.9 *(Estimating the constant correlation model) A quick (and dirty) method for estimating 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*. It is slightly harder to estimate.

See Figure 10.12 for an illustration and Figure 10.13 for a comparison with the EWMA approach.

Bibliography

Bodie, Z., A. Kane, and A. J. Marcus, 2005, *Investments*, McGraw-Hill, Boston, 6th edn.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The econometrics of financial markets*, Princeton University Press, Princeton, New Jersey.

Franses, P. H., and D. van Dijk, 2000, *Non-linear time series models in empirical finance*, Cambridge University Press.

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.

Hentschel, L., 1995, "All in the family: nesting symmetric and asymmetric GARCH models," *Journal of Financial Economics*, 39, 71–104.

Nelson, D. B., 1991, "Conditional heteroskedasticity in asset returns," *Econometrica*, 59, 347–370.

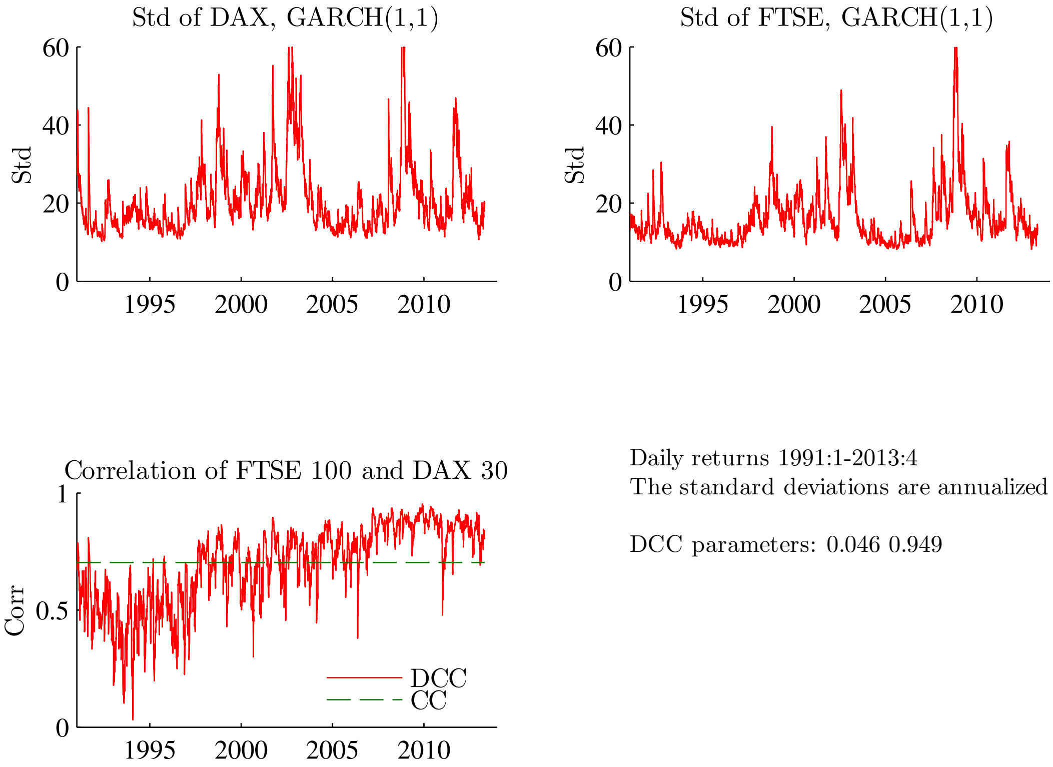


Figure 10.12: Results for multivariate GARCH models

Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.

Verbeek, M., 2004, *A guide to modern econometrics*, Wiley, Chichester, 2nd edn.

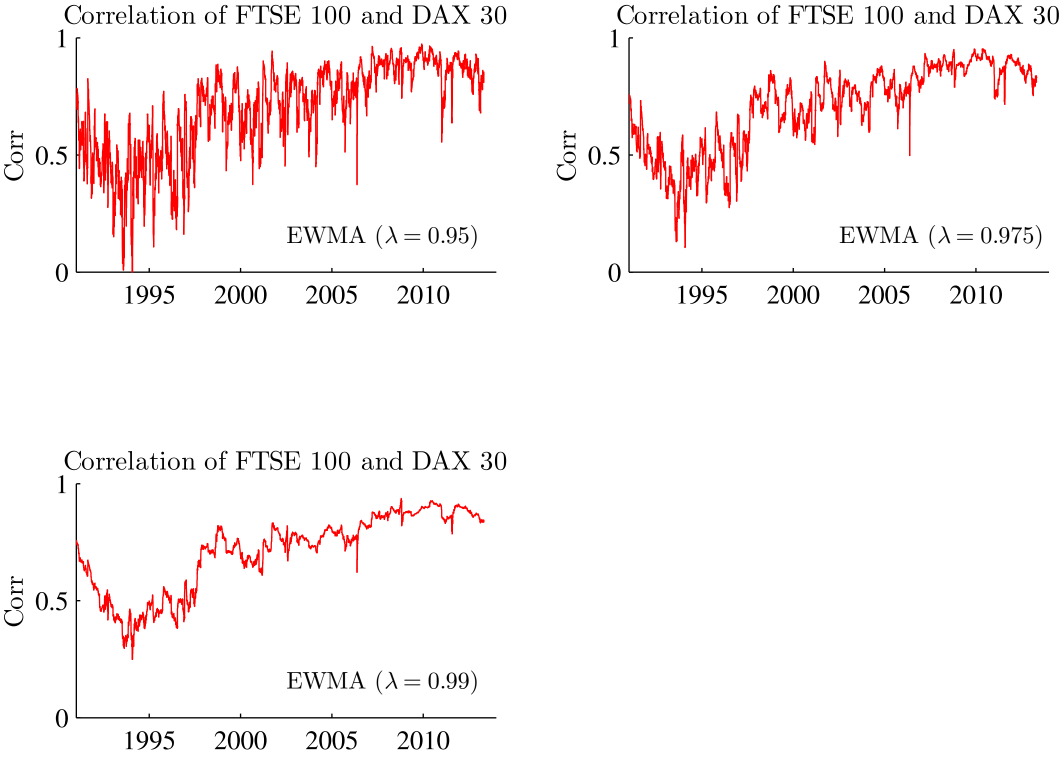


Figure 10.13: Time-varying correlations (different EWMA estimates)

11 Risk Measures

Reference: Hull (2006) 18; McDonald (2006) 25; Fabozzi, Focardi, and Kolm (2006) 4–5; McNeil, Frey, and Embrechts (2005); Alexander (2008)

11.1 Value at Risk

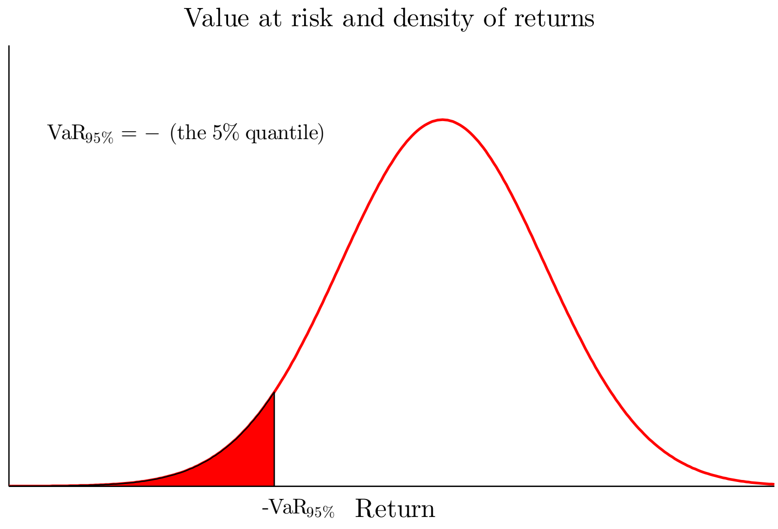


Figure 11.1: Value at risk

The mean-variance framework is often criticized for failing to distinguish between downside (considered to be risk) and upside (considered to be potential).

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 that $\text{VaR}_{95\%}$

$$\Pr(-R \geq \text{VaR}_{95\%}) = 5\%. \tag{11.1}$$

Here, 95% is the confidence level of the VaR. More generally, a there is only a $1 - \alpha$

chance that the loss ($-R$) is larger that VaR_α (the confidence level is α)

$$\Pr(-R \geq \mathrm{VaR}_\alpha) = 1 - \alpha. \tag{11.2}$$

Clearly, $-R$ $\geq$VaR$_\alpha$ is true when (and only when) $R \leq -\mathrm{VaR}_\alpha$, so (11.2) can also be expressed as

$$\Pr(R \leq -\mathrm{VaR}_\alpha) = \mathrm{cdf}_R(-\mathrm{VaR}_\alpha) = 1 - \alpha, \tag{11.3}$$

where $\mathrm{cdf}_R()$ is the cumulative distribution function of the returns. This says that $-\mathrm{VaR}_\alpha$ is a number such that there is only a $1-\alpha$ (5%, say) chance that the return is below it. See Figures 11.1–11.2 for illustrations. Using (11.3) allows us to work directly with the return distribution (not the loss distribution), which is often convenient.

Example 11.1 *(Quantile of a distribution) The 0.05 quantile is the value such that there is only a 5% probability of a lower number,* $\Pr(R \leq quantile_{0.05}) = 0.05$.

We can solve (11.3) for the value at risk, VaR_α, as

$$\mathrm{VaR}_\alpha = -\mathrm{cdf}_R^{-1}(1-\alpha), \tag{11.4}$$

where $\mathrm{cdf}_R^{-1}()$ is the inverse of the cumulative distribution function of the returns, so $\mathrm{cdf}_R^{-1}(1-\alpha)$ is the $1-\alpha$ quantile (or "critical value") of the return distribution. For instance, $\mathrm{VaR}_{95\%}$ is the (negative of the) 0.05 quantile of the return distribution.

To convert the value at risk into value terms (CHF, say), just multiply the VaR for returns with the value of the investment (portfolio).

If the return is normally distributed, $R \sim N(\mu, \sigma^2)$ and $c_{1-\alpha}$ is the $1-\alpha$ quantile of a N(0,1) distribution (for instance, -1.64 for $1-\alpha = 0.05$), then

$$\mathrm{VaR}_\alpha = -(\mu + c_{1-\alpha}\sigma). \tag{11.5}$$

This is illustrated in Figure 11.4.

Remark 11.2 *(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 $R \leq \mu - 2.33\sigma$.*

Example 11.3 *(VaR with $R \sim N(\mu, \sigma^2)$) If daily returns have $\mu = 8\%$ and $\sigma = 16\%$, then the 1-day $VaR_{95\%}$* $= -(0.08 - 1.64 \times 0.16) \approx 0.18$*; we are 95% sure that we will not*

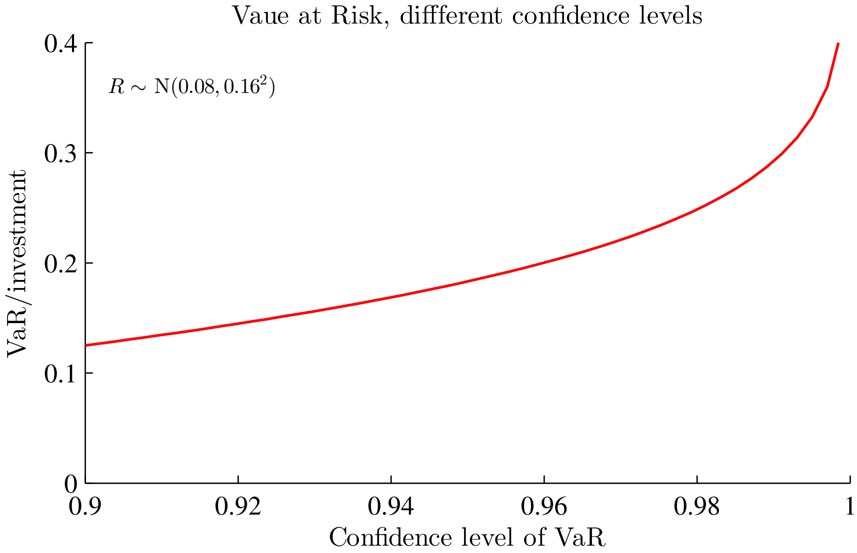


Figure 11.2: Value at risk, different probability levels

loose more than 18% *of the investment over one day, that is, $VaR_{95\%} = 0.18$. Similarly, $VaR_{97.5\%} = -(0.08 - 1.96 \times 0.16) \approx 0.24$.*

Figure 11.3 shows the distribution and VaRs (for different probability levels) for the daily S&P 500 returns. Two different VaRs are shown: based on a normal distribution and as the empirical VaR (from the empirical quantiles of the distribution). While these results are interesting, they are just time-averages in the sense of being calculated from the unconditional distribution: time-variation in the distribution is not accounted for.

Figure 11.5 illustrates the VaR calculated from a time series model (to be precise, an AR(1)+GARCH(1,1) model) for daily S&P returns. In this case, the VaR changes from day to day as both the mean return (the forecast) as well as the standard error (of the forecast error) do. Since the volatility clearly changes over time, this is crucial for a reliable VaR model.

Notice that the value at risk in (11.5), that is, when the return is normally distributed, is a strictly increasing function of the standard deviation (and the variance). This follows from the fact that $c_{1-\alpha} < 0$ (provided $1-\alpha < 50\%$, which is the relevant case). Minimizing the VaR 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, the portfolio choice will be different (and perhaps complicated to perform).

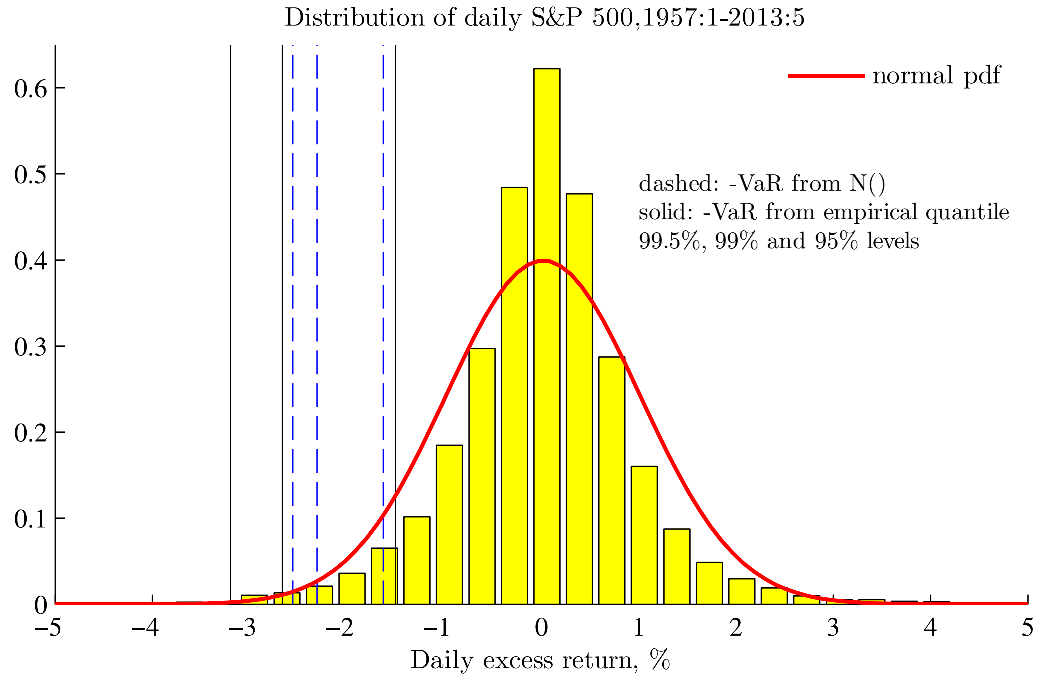


Figure 11.3: Return distribution and VaR for S&P 500

Example 11.4 *(VaR and regulation of bank capital) Bank regulations have used 3 times the 99% VaR for 10-day returns as the required bank capital.*

Notice that the return distribution depends on the investment horizon, so a value at risk measure 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 11.5 *(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.*

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

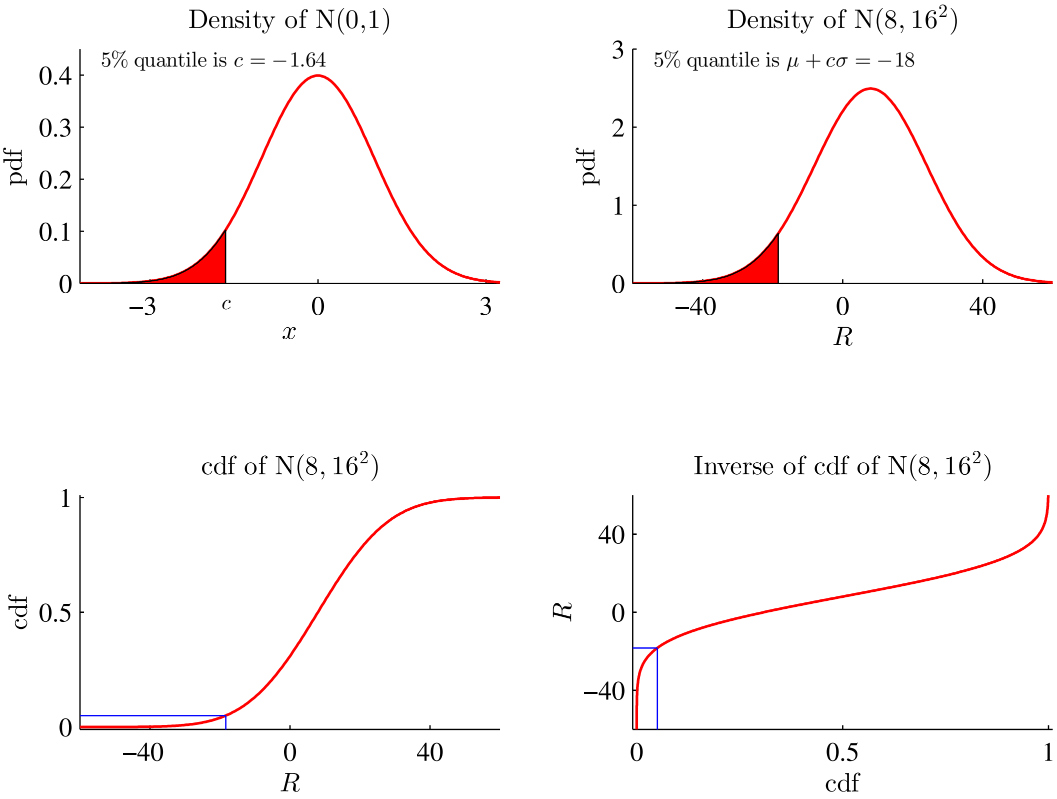
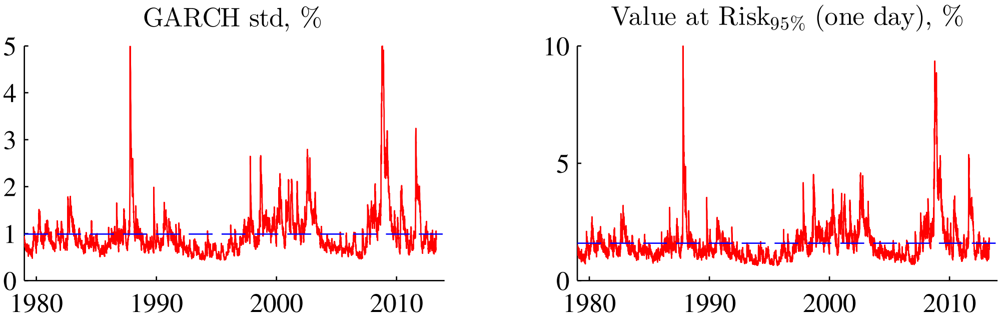


Figure 11.4: Finding critical value of N(μ,σ^2) distribution

negative of) this number. Figures 11.6–11.7 show results from backtesting a VaR model where the volatility follows a GARCH process. It suggests that a GARCH model (to capture the time varying volatility), combined with the assumption that the return is normally distributed (but with time-varying parameters), works relatively well.

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.)

See Table 11.1 for an empirical comparison of the VaR with some alternative downside risk measures (discussed below).

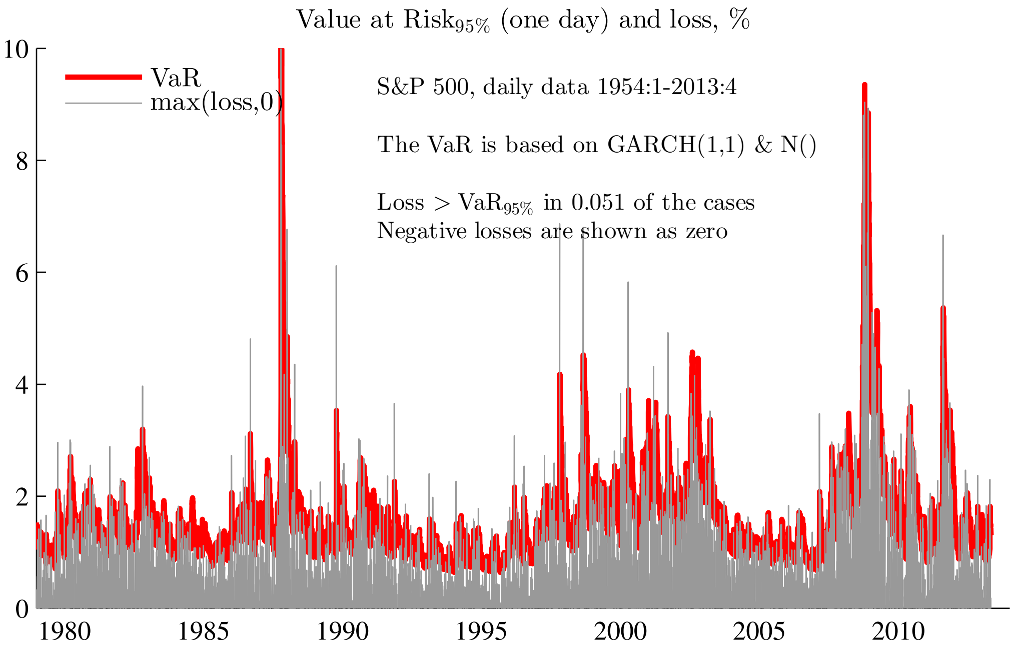


S&P 500, daily data 1954:1-2013:4

The horizontal lines are from the unconditional distribution

The VaR is based on N()

Figure 11.5: Conditional volatility and VaR



S&P 500, daily data 1954:1-2013:4

The VaR is based on GARCH(1,1) & N()

Loss > VaR$_{95\%}$ in 0.051 of the cases
Negative losses are shown as zero

Figure 11.6: Backtesting VaR from a GARCH model, assuming normally distributed shocks

11.1.1 Value at Risk of a Portfolio*

If the return distribution is normal with a zero mean, then the value at risk for asset i is

$$\text{VaR}_i = 1.64\sigma_i. \tag{11.6}$$

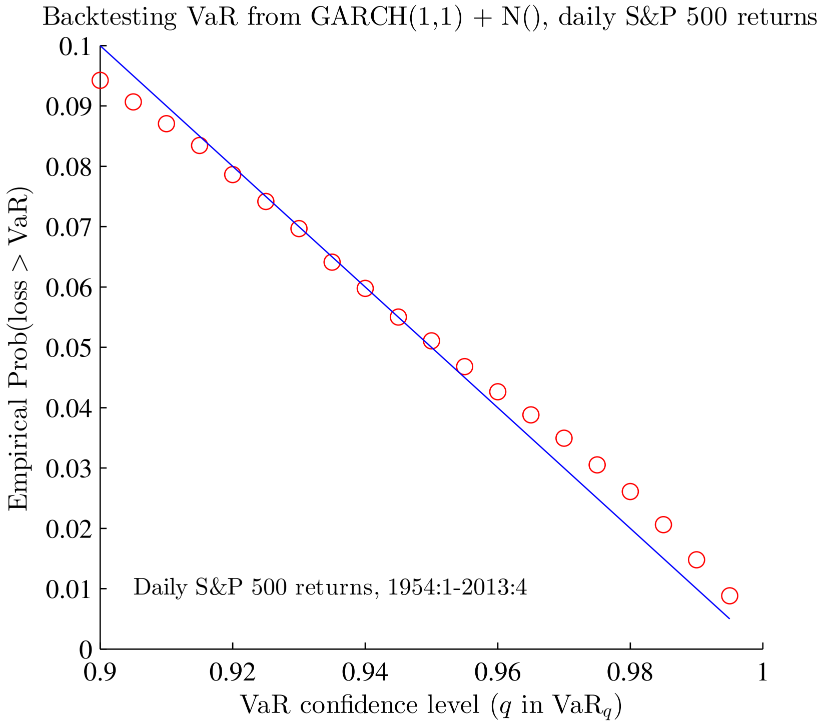


Figure 11.7: Backtesting VaR from a GARCH model, assuming normally distributed shocks

	Small growth	Large value
Std	8.0	5.0
VaR (95%)	12.3	8.3
ES (95%)	17.2	10.8
SemiStd	5.5	3.4
Drawdown	79.7	52.3

Table 11.1: Risk measures of monthly returns of two stock indices (%), US data 1957:1-2012:12.

It is then straightfoward to show that the VaR for a portfortfolio

$$R_p = w_1 R_1 + w_2 R_2, \tag{11.7}$$

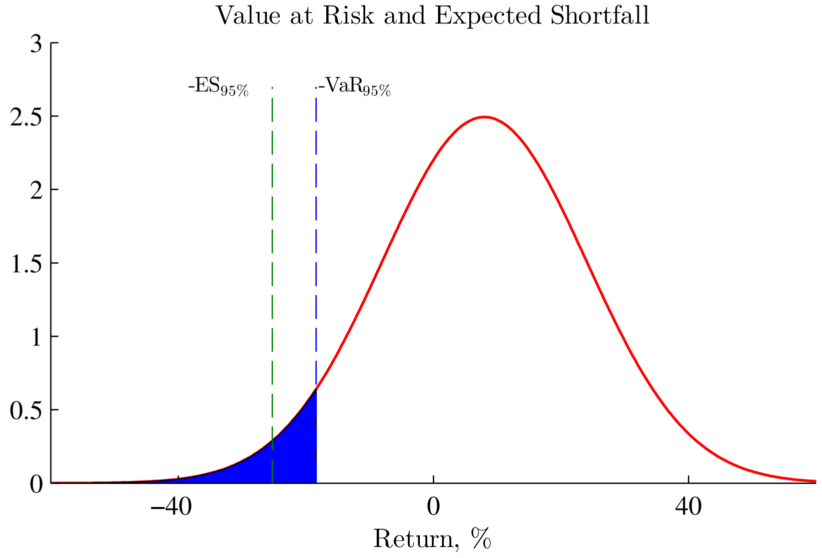


Figure 11.8: Value at risk and expected shortfall

where $w_1 + w_2 = 1$ can be written

$$\text{VaR}_p = \left(\begin{bmatrix} w_1\text{Var}_1 & w_2\text{Var}_2 \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \begin{bmatrix} w_1\text{Var}_1 \\ w_2\text{Var}_2 \end{bmatrix} \right)^{1/2}, \tag{11.8}$$

where ρ_{12} is the correlation of R_1 and R_2. The extension to n (instead of 2) assets is straightforward.

This expression highlights the importance of both the individual VaR$_i$ values and the correlation. Clearly, a worst case scenario is when the portfolio is long in all assets ($w_i > 0$) and the correlation turns out to be perfect ($\rho_{12} = 1$).

Proof. (of (11.8)) Recall that VaR$_p$ = 1.64σ_p, and that

$$\sigma_p^2 = w_1^2\sigma_{11} + w_2^2\sigma_{22} + 2w_1w_2\rho_{12}\sigma_1\sigma_2.$$

Use (11.6) to substitute as σ_i =VaR$_i$/1.64

$$\sigma_p^2 = w_1^2\text{VaR}_1^2/1.64^2 + w_2^2\text{VaR}_2^2/1.64^2 + 2w_1w_2\rho_{12} \times \text{VaR}_1 \times \text{VaR}_2/1.64^2.$$

Multiply both sides by 1.64^2 and take the square root to get (11.8). ■

11.1.2 Index Models for Calculating the Value at Risk*

Consider a multi-index model

$$\begin{aligned} R &= a + b_1 I_1 + b_2 I_2 + \ldots + b_k I_k + e, \text{ or} \qquad (11.9)\\ &= a + b'I + e, \end{aligned}$$

where b is a $k \times 1$ vector of the b_i coefficients and I is also a $k \times 1$ vector of the I_i indices. As usual, we assume $\mathrm{E}(e) = 0$ and $\mathrm{Cov}\,(e, I_i) = 0$. This model can be used to generate the inputs to a VaR model. For instance, the mean and standard deviation of the return are

$$\begin{aligned} \mu &= a + b'\,\mathrm{E}\,I\\ \sigma &= \sqrt{b'\,\mathrm{Cov}(I)b + \mathrm{Var}(e)}, \qquad (11.10) \end{aligned}$$

which can be used in (11.5), that is, an assumption of a normal return distribution. If the return is of a well diversified portfolio and the indices include the key stock indices, then the idiosyncratic risk Var(e) is close to zero. The RiskMetrics approach is to make this assumption.

Stand-alone VaR is a way to assess the contribution of different factors (indices). For instance, the indices in (11.9) could include: an equity indices, interest rates, exchange rates and perhaps also a few commodity indices. Then, an *equity VaR* is calculated by setting all elements in b, except those for the equity indices, to zero. Often, the intercept, a, is also set to zero. Similarly, an *interest rate VaR* is calculated by setting all elements in b, except referring to the interest rates, to zero. And so forth for an *FX VaR* and a *commodity VaR*. Clearly, these different VaRs do not add up to the total VaR, but they still give an indication of where the main risk comes from.

If an asset or a portfolio is a non-linear function of the indices, then (11.9) can be thought of as a first-order Taylor approximation where b_i represents the partial derivative of the asset return with respect to index i. For instance, an option is a non-linear function of the underlying asset value and its volatility (as well as the time to expiration and the interest rate). This approach, when combined with the normal assumption in (11.5), is called the *delta-normal method*.

11.2 Expected Shortfall

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,

$$\text{ES}_\alpha = -\text{E}(R|R \leq -\text{VaR}_\alpha). \tag{11.11}$$

This might be more informative than the VaR_α, which is the *minimum loss* that will happen with a $1 - \alpha$ probability.

For a normally distributed return $R \sim N(\mu, \sigma^2)$ we have

$$\text{ES}_\alpha = -\mu + \sigma\frac{\phi(c_{1-\alpha})}{1-\alpha}, \tag{11.12}$$

where $\phi()$ is the pdf or 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 (11.12)) If $x \sim N(\mu, \sigma^2)$, then $\text{E}(x|x \leq b) = \mu - \sigma\phi(b_0)/\Phi(b_0)$ 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 11.6 *(ES) If $\mu = 8\%$ and $\sigma = 16\%$, the 95% expected shortfall is $ES_{95\%} = -0.08 + 0.16\phi(-1.64)/0.05 \approx 0.25$ 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 (11.12) 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, the portfolio choice will be different (and perhaps complicated to perform).

11.3 Target Semivariance (Lower Partial 2nd Moment) and Max Drawdown

Reference: Bawa and Lindenberg (1977) and Nantell and Price (1979)

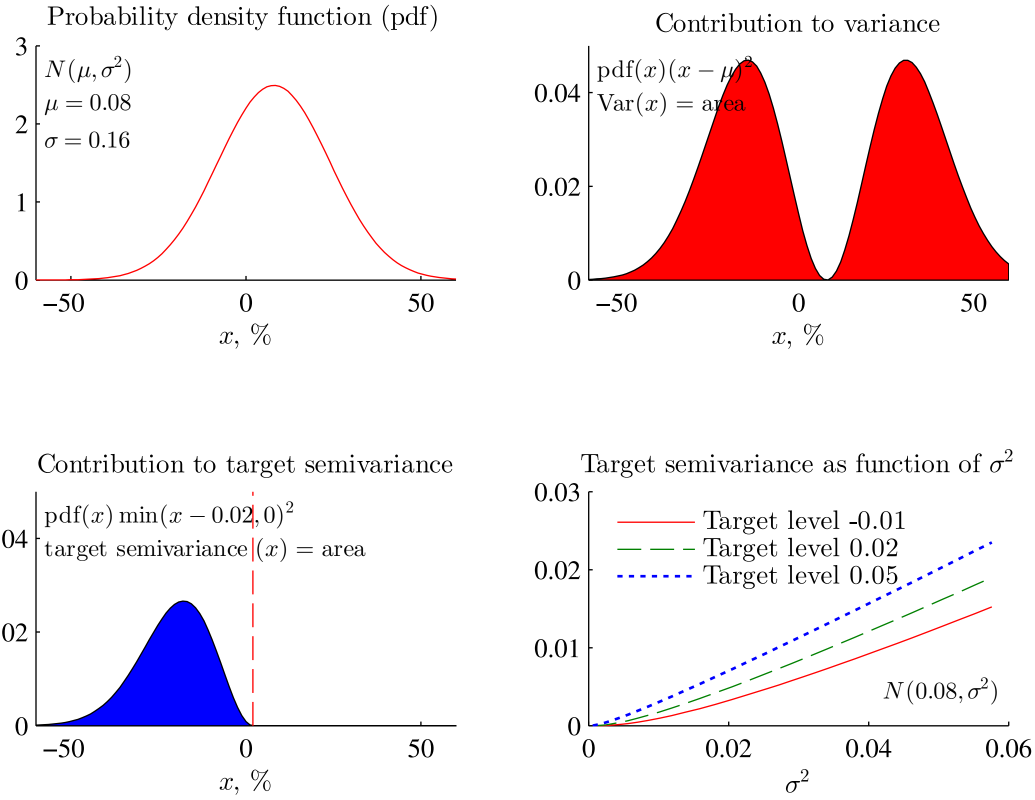


Figure 11.9: Target semivariance as a function of mean and standard deviation for a N(μ,σ^2) variable

Using the variance (or standard deviation) as a measure of portfolio risk (as a mean-variance investor does) fails to distinguish between the downside and upside. As an alternative, one could consider using a target semivariance (lower partial 2nd moment) instead. It is defined as

$$\lambda_p(h) = \text{E}[\min(R_p - h, 0)^2], \tag{11.13}$$

where h is a "target level" chosen by the investor. In the subsequent analysis it will be set equal to the riskfree rate. (It can clearly also be written $\lambda_p(h) = \int_{-\infty}^{h}(R_p - h)^2 f(R_p)dR_p$, where $f()$ is the pdf of the portfolio return.)

In comparison with a variance

$$\sigma_p^2 = \text{E}(R_p - \mu_p)^2, \tag{11.14}$$

the target semivariance differs on two accounts: *(i)* it uses the target level h as a reference point instead of the mean μ_p: and *(ii)* only negative deviations from the reference point are given any weight. See Figure 11.9 for an illustration (based on a normally distributed variable).

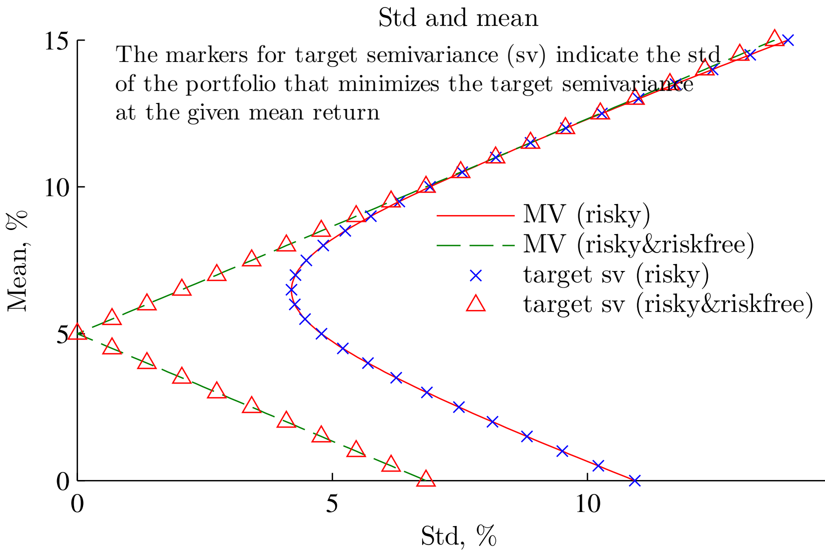


Figure 11.10: Standard deviation and expected returns

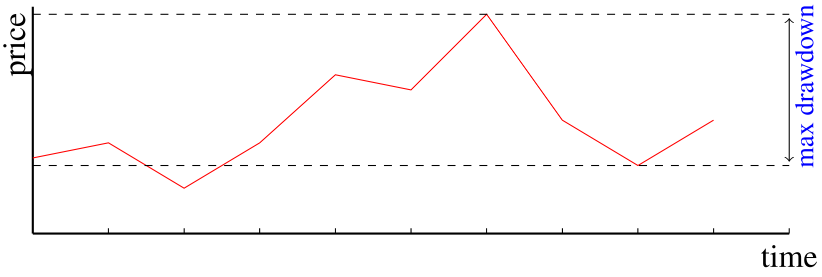


Figure 11.11: Max drawdown

For a normally distributed variable, the target semivariance $\lambda_p(h)$ is increasing in the standard deviation (for a given mean)—see Remark 11.7. See also Figure 11.9 for an illustration. This means that minimizing $\lambda_p(h)$ at a given mean return gives the same solution (portfolio weights) as minimizing σ_p (or σ_p^2) at the same given mean return. As

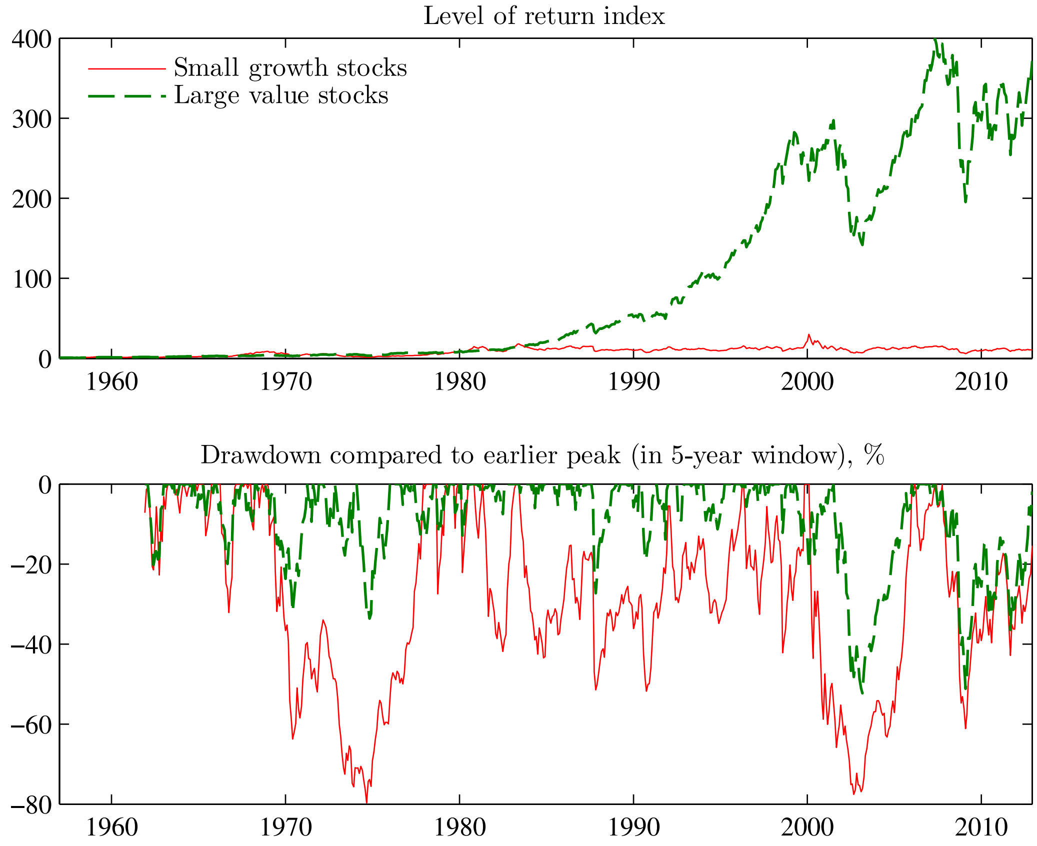


Figure 11.12: Drawdown

a result, with normally distributed returns, an investor who wants to minimize the lower partial 2nd moment (at a given mean return) is behaving just like a mean-variance investor. In other cases, the portfolio choice will be different (and perhaps complicated to perform).

See Figure 11.10 for an illustration.

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 11.11. 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 11.12–11.13 for an illustration of max drawdown.

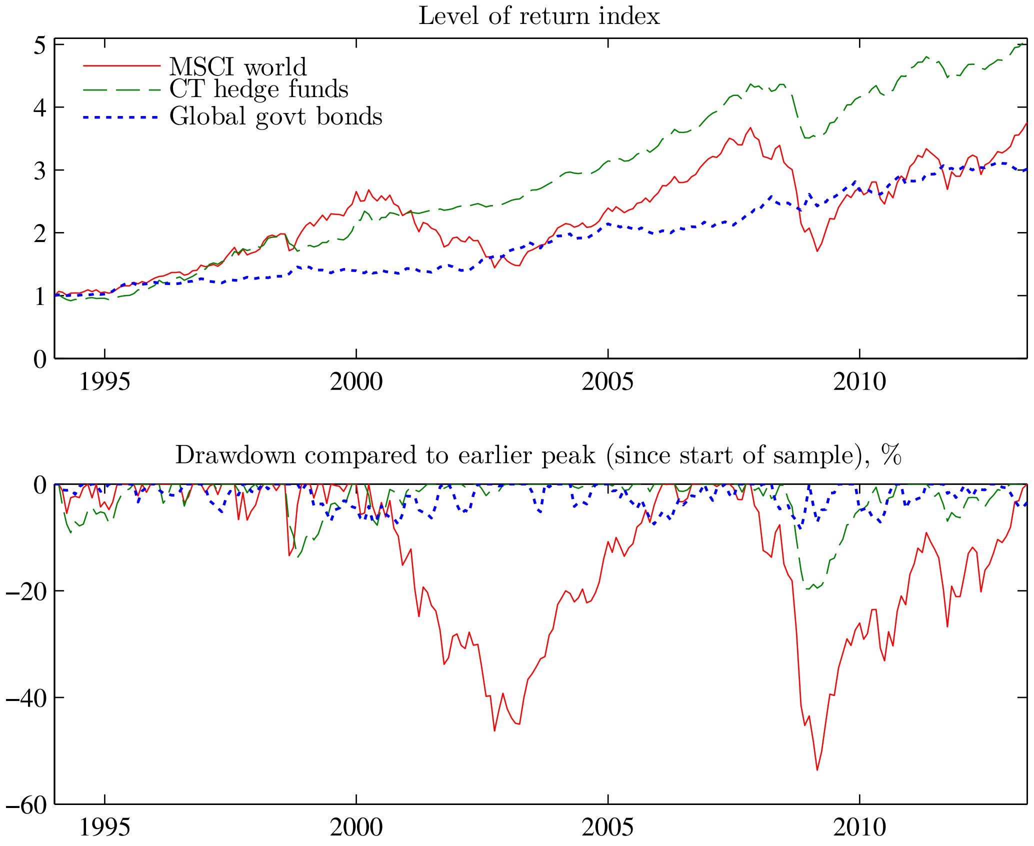


Figure 11.13: Drawdown

Remark 11.7 *(Target semivariance calculation for normally distributed variable*) For an $N(\mu, \sigma^2)$ variable, target semivariance around the target level h is*

$$\lambda_p(h) = \sigma^2 a\phi(a) + \sigma^2(a^2+1)\Phi(a), \text{ where } a = (h-\mu)/\sigma,$$

where $\phi()$ and $\Phi()$ are the pdf and cdf of a $N(0,1)$ variable respectively. Notice that $\lambda_p(h) = \sigma^2/2$ for $h = \mu$. See Figure 11.9 for a numerical illustration. It is straightforward (but a bit tedious) to show that

$$\frac{\partial \lambda_p(h)}{\partial \sigma} = 2\sigma\Phi(a),$$

so the target semivariance is a strictly increasing function of the standard deviation.

See Table 11.2 for an empirical comparison of the different risk measures.

	Std	VaR (95%)	ES (95%)	SemiStd	Drawdown
Std	1.00	0.94	0.98	0.97	0.68
VaR (95%)	0.94	1.00	0.94	0.95	0.72
ES (95%)	0.98	0.94	1.00	0.98	0.67
SemiStd	0.97	0.95	0.98	1.00	0.68
Drawdown	0.68	0.72	0.67	0.68	1.00

Table 11.2: Correlation of rank of risk measures across the 25 FF portfolios (%), US data 1957:1-2012:12.

Bibliography

Alexander, C., 2008, *Market Risk Analysis: Value at Risk Models*, Wiley.

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.

Fabozzi, F. J., S. M. Focardi, and P. N. Kolm, 2006, *Financial modeling of the equity market*, Wiley Finance.

Hull, J. C., 2006, *Options, futures, and other derivatives*, Prentice-Hall, Upper Saddle River, NJ, 6th 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.

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.

12 Return Distributions (Univariate)

Sections denoted by a star (*) is not required reading.

12.1 Estimating and Testing Distributions

Reference: Harvey (1989) 260, Davidson and MacKinnon (1993) 267, Silverman (1986); Mittelhammer (1996), DeGroot (1986)

12.1.1 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,

$$u_i = F_i(x_i) = \Pr(X_i \leq x_i). \tag{12.1}$$

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

$$F_i(x_i) = \int_{s=-\infty}^{x_i} f_i(s)ds. \tag{12.2}$$

(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 12.1 *(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 12.1 for an illustration.*

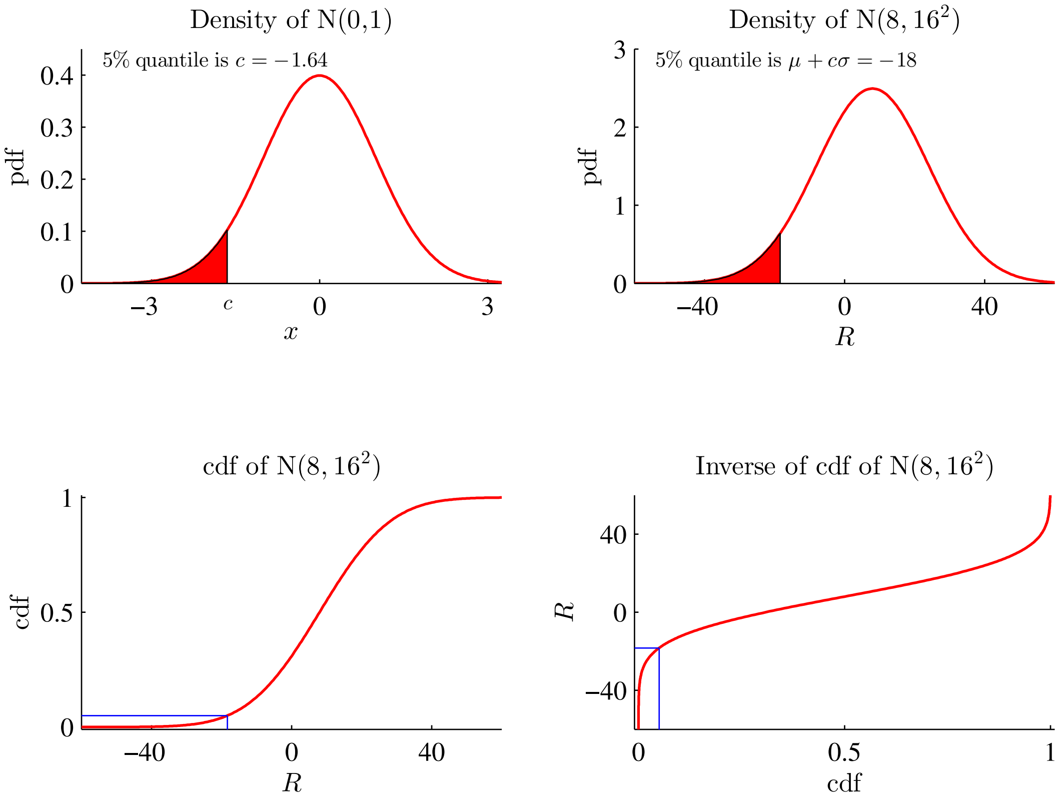


Figure 12.1: Finding quantiles of a N(μ,σ^2) distribution

12.1.2 QQ Plots

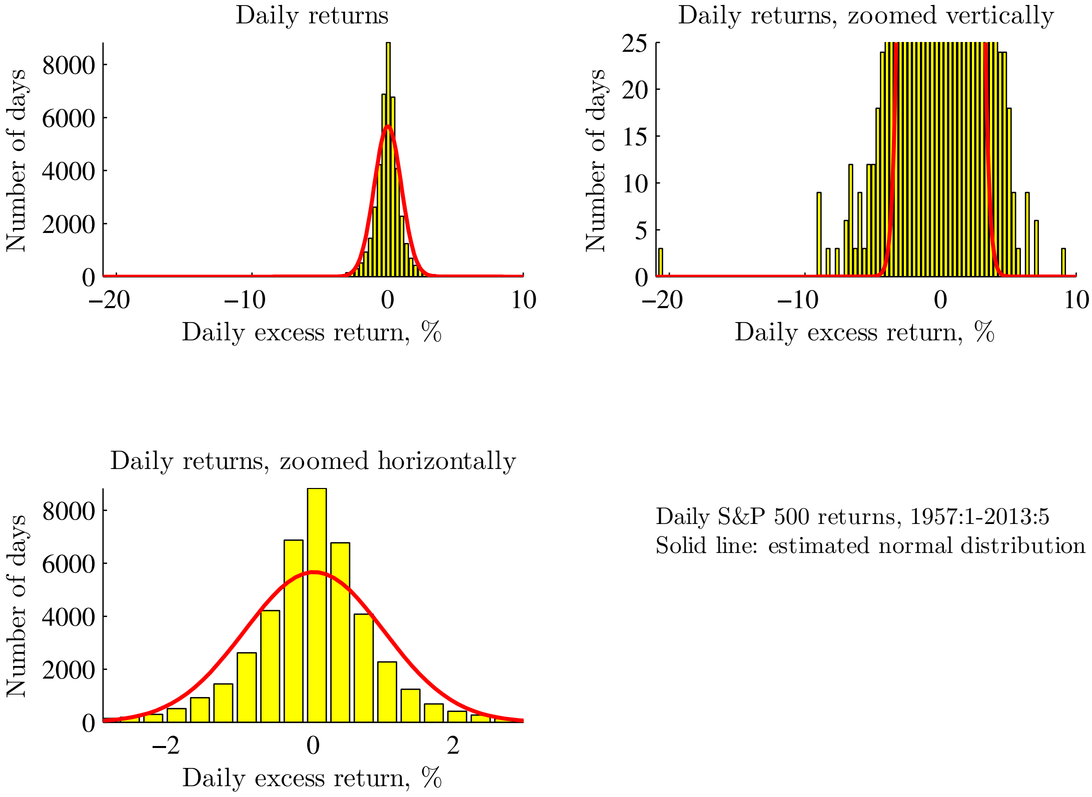
Are returns normally distributed? Mostly not, but it depends on the asset type and on the data frequency. Options returns typically have very non-normal distributions (in particular, since the return is -100% on many expiration days). Stock returns are typically distinctly non-linear at short horizons, 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 12.2–12.4 for illustrations.

Remark 12.2 *(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. Each point in the QQ plot shows a specific percentile (quantile) according to the empiri-*

cal 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 fairly different left tails.

There is one caveat to this way of studying data: it only provides evidence on the unconditional distribution. For instance, nothing rules out the possibility that we could estimate a model for time-varying volatility (for instance, a GARCH model) of the returns and thus generate a description for how the VaR changes over time. However, data with time varying volatility will typically not have an unconditional normal distribution.



Daily S&P 500 returns, 1957:1-2013:5
Solid line: estimated normal distribution

Figure 12.2: Distribution of daily S&P returns

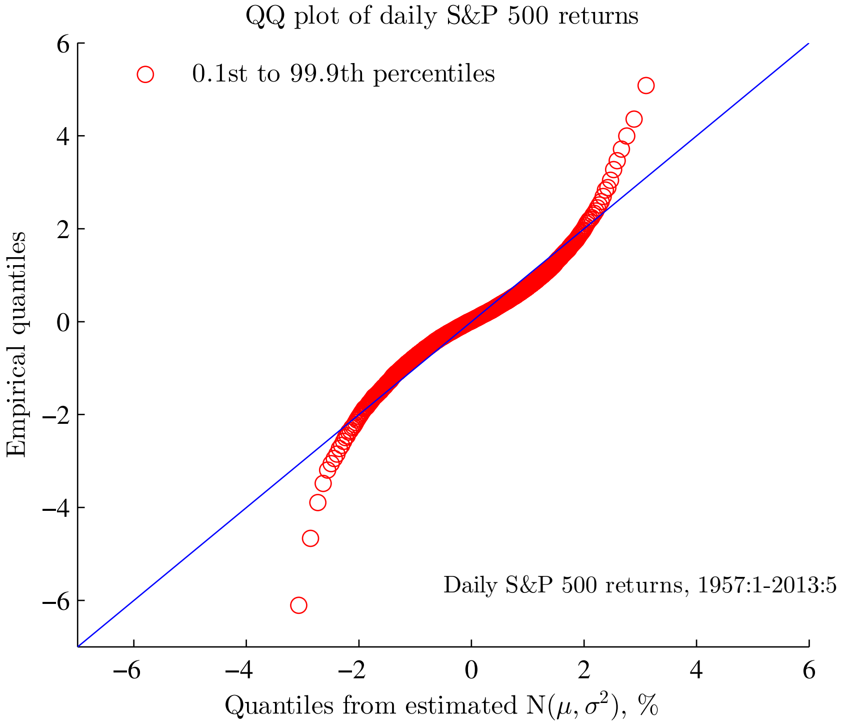


Figure 12.3: Quantiles of daily S&P returns

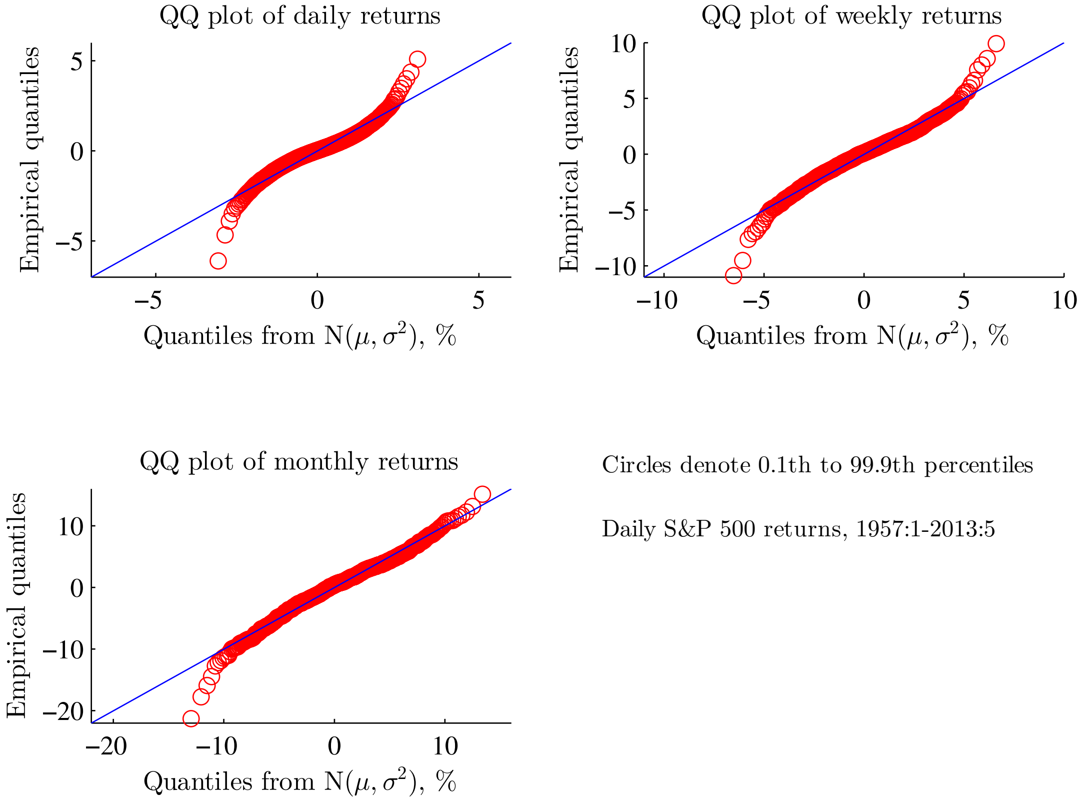
12.1.3 Parametric Tests of Normal Distribution

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

$$\begin{array}{lcll} & & \text{Test statistic} & \text{Distribution} \\ \text{skewness} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^3 & N\left(0,6/T\right) \\ \text{kurtosis} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^4 & N\left(3,24/T\right) \\ \text{Bera-Jarque} & = & \frac{T}{6}\text{skewness}^2+\frac{T}{24}\left(\text{kurtosis}-3\right)^2 & \chi^2_2. \end{array} \tag{12.3}$$

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (12.3) are under the null hypothesis that x_t is iid $N\left(\mu,\sigma^2\right)$. 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



Circles denote 0.1th to 99.9th percentiles

Daily S&P 500 returns, 1957:1-2013:5

Figure 12.4: Distribution of S&P returns (different horizons)

sum of the square of two uncorrelated $N(0,1)$ variables, which has a χ_2^2 distribution.

12.1.4 Nonparametric Tests of General Distributions

The *Kolmogorov-Smirnov* test is designed to test if an empirical distribution function, EDF(x), conforms with a theoretical cdf, $F(x)$. The empirical distribution function is defined as the fraction of observations which are less or equal to x, that is,

$$\text{EDF}(x) = \frac{1}{T}\sum_{t=1}^{T}\delta(x_t \leq x), \text{ where} \tag{12.4}$$

$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

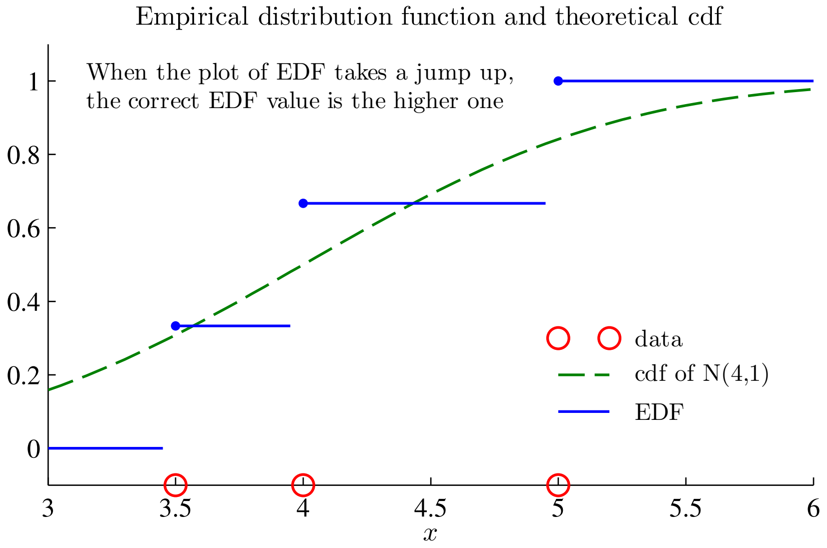


Figure 12.5: Example of empirical distribution function

The EDF(x_t) and $F(x_t)$ are often plotted against the sorted (in ascending order) sample $\{x_t\}_{t=1}^{T}$.

See Figure 12.5 for an illustration.

Example 12.3 *(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 12.5.*

Define the absolute value of the maximum distance

$$D_T = \max_{x_t} |\text{EDF}(x_t) - F(x_t)|. \tag{12.5}$$

Example 12.4 *(Kolmogorov-Smirnov test statistic) Figure 12.5 also shows the cumulative distribution function (cdf) of a normally distributed variable. The test statistic (12.5) is then the largest difference (in absolute terms) of the EDF and the cdf—among the observed values of x_t.*

We reject the null hypothesis that EDF$(x) = F(x)$ if $\sqrt{T} D_t > c$, where c is a critical value which can be calculated from

$$\lim_{T\to\infty} \Pr\left(\sqrt{T} D_T \leq c\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2c^2}. \tag{12.6}$$

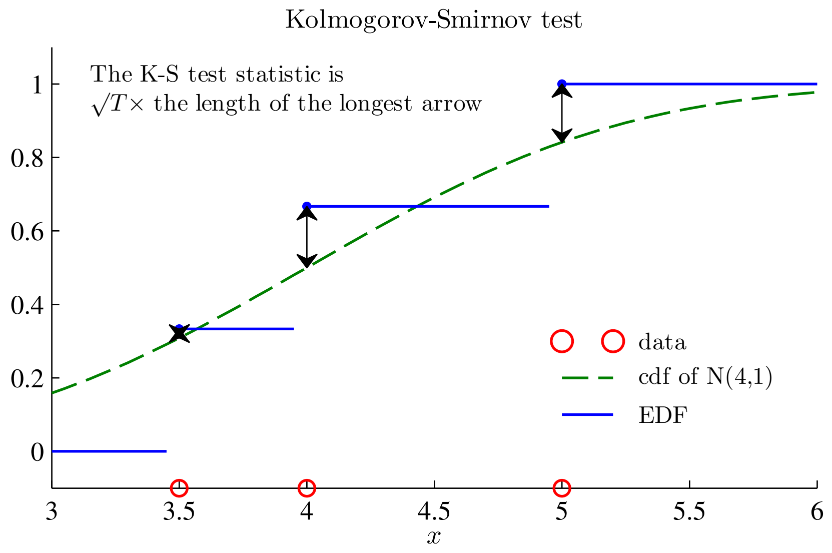


Figure 12.6: K-S test

It can be approximated by replacing ∞ with a large number (for instance, 100). For instance, $c = 1.35$ provides a 5% critical value. See Figure 12.7. There is a corresponding test for comparing two empirical cdfs.

Pearson's χ^2 *test* does the same thing as the K-S test but for a discrete distribution. Suppose you have K categories with N_i values in category i. The theoretical distribution predicts that the fraction p_i should be in category i, with $\sum_{i=1}^{K} p_i = 1$. Then

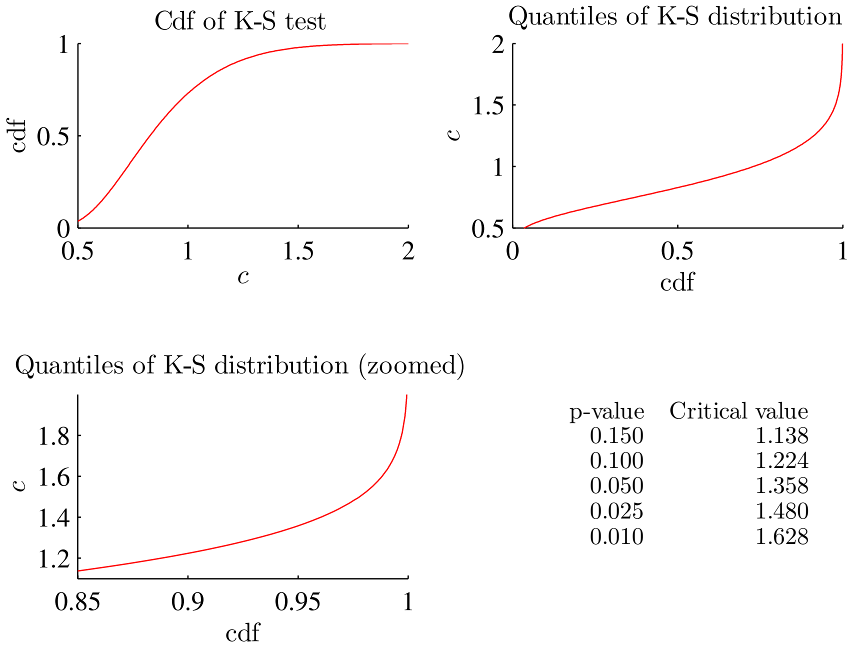
$$\sum_{i=1}^{K} \frac{(N_i - Tp_i)^2}{Tp_i} \sim \chi^2_{K-1}. \tag{12.7}$$

There is a corresponding test for comparing two empirical distributions.

12.1.5 Fitting a Mixture Normal Distribution to Data*

Reference: Hastie, Tibshirani, and Friedman (2001) 8.5

A normal distribution often fits returns poorly. If we need a distribution, then a mixture



p-value	Critical value
0.150	1.138
0.100	1.224
0.050	1.358
0.025	1.480
0.010	1.628

Figure 12.7: Distribution of the Kolmogorov-Smirnov test statistics, $\sqrt{T} D_T$

of two normals is typically much better, and still fairly simple.

The pdf of this distribution is just a weighted average of two different (bell shaped) pdfs of normal distributions (also called mixture components)

$$f(x_t; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi) = (1 - \pi)\phi(x_t; \mu_1, \sigma_1^2) + \pi\phi(x_t; \mu_2, \sigma_2^2), \tag{12.8}$$

where $\phi(x; \mu_i, \sigma_i^2)$ is the pdf of a normal distribution with mean μ_i and variance σ_i^2. It thus contains five parameters: the means and the variances of the two components and their relative weight (π).

See Figures 12.8–12.10 for an illustration.

Remark 12.5 *(Estimation of the mixture normal pdf) With 2 mixture components, the log likelihood is just*

$$LL = \sum\nolimits_{t=1}^{T} \ln f(x_t; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi),$$

where $f()$ is the pdf in (12.8) A numerical optimization method could be used to maximize this likelihood function. However, this is tricky so an alternative approach is often used. This is an iterative approach in three steps:

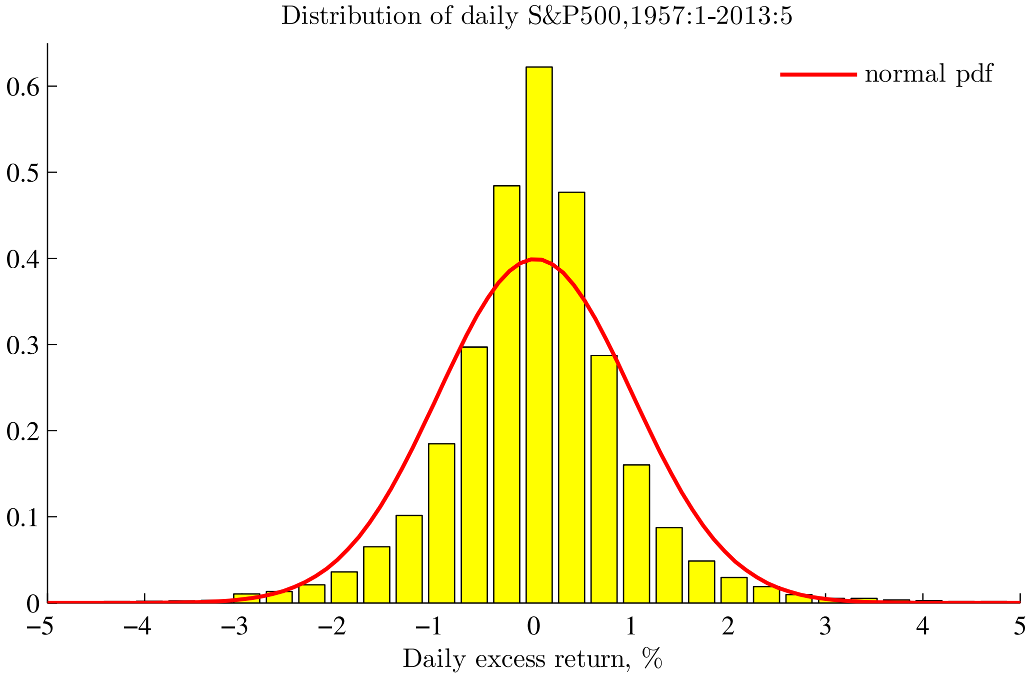


Figure 12.8: Histogram of returns and a fitted normal distribution

(1) Guess values of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ *and* π*. For instance, pick* $\mu_1 = x_1$*,* $\mu_2 = x_2$*,* $\sigma_1^2 = \sigma_2^2 = \operatorname{Var}(x_t)$ *and* $\pi = 0.5$*.*

(2) Calculate

$$\gamma_t = \frac{\pi\phi(x_t;\mu_2,\sigma_2^2)}{(1-\pi)\phi(x_t;\mu_1,\sigma_1^2)+\pi\phi(x_t;\mu_2,\sigma_2^2)} \text{ for } t = 1,\ldots,T.$$

(3) Calculate (in this order)

$$\begin{aligned}
\mu_1 &= \frac{\sum_{t=1}^T (1-\gamma_t)x_t}{\sum_{t=1}^T (1-\gamma_t)},\ \sigma_1^2 = \frac{\sum_{t=1}^T (1-\gamma_t)(x_t-\mu_1)^2}{\sum_{t=1}^T (1-\gamma_t)},\\
\mu_2 &= \frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t},\ \sigma_2^2 = \frac{\sum_{t=1}^T \gamma_t(x_t-\mu_2)^2}{\sum_{t=1}^T \gamma_t}, \text{ and}\\
\pi &= \sum\nolimits_{t=1}^T \gamma_t/T.
\end{aligned}$$

Iterate over (2) and (3) until the parameter values converge. (This is an example of the EM algorithm.) Notice that the calculation of σ_i^2 *uses* μ_i *from the same (not the previous)*

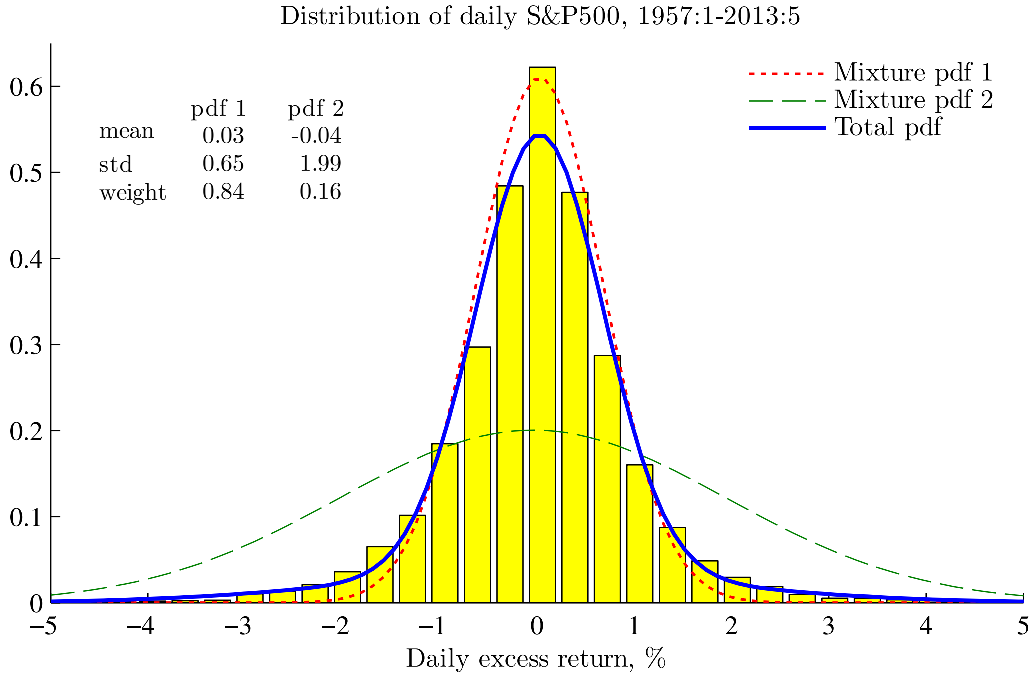


Figure 12.9: Histogram of returns and a fitted mixture normal distribution

iteration.

12.1.6 Kernel Density Estimation

Reference: Silverman (1986)

A histogram is just a count of the relative number of observations that fall in (prespecified) non-overlapping intervals. If we also divide by the width of the interval, then the area under the histogram is unity, so the scaled histogram can be interpreted as a density function. For instance, if the intervals ("bins") are h wide, then the scaled histogram at the point x (say, $x = 2.3$) can be defined as

$$g(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{h}\delta(x_t \text{ is in bin}_i), \text{ where} \tag{12.9}$$

$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

Note that the area under $g(x)$ indeed integrates to unity.

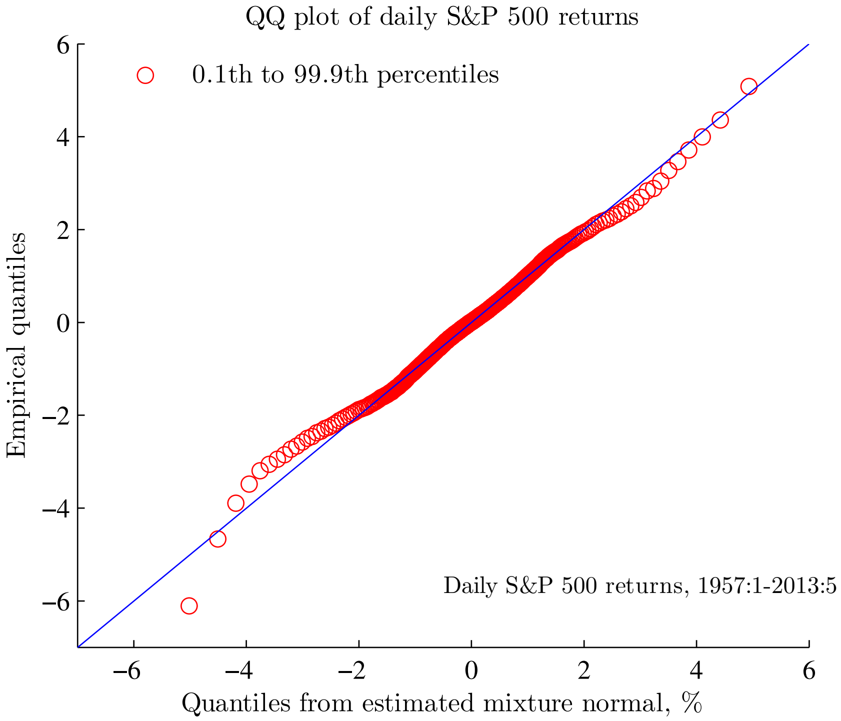


Figure 12.10: Quantiles of daily S&P returns

We can gain efficiency by using a more sophisticated estimator. In particular, using a pdf instead of the binary function is often both convenient and more efficient. To develop that method, we first show an alternative way of constructing a histogram. First, let a bin be defined as symmetric interval around a point x: $x - h/2$ to $x + h/2$. (We can vary the value of x to define other bins.) Second, notice that the histogram value at point x can be written

$$g(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{h}\delta\left(\left|\frac{x_t - x}{h}\right| \leq 1/2\right). \tag{12.10}$$

In fact, $\frac{1}{h}\delta(|x_t - x| \leq h/2)$ 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 that the uniform. In particular, using a density function that tapers off continuously instead of suddenly dropping to zero (as the uniform density

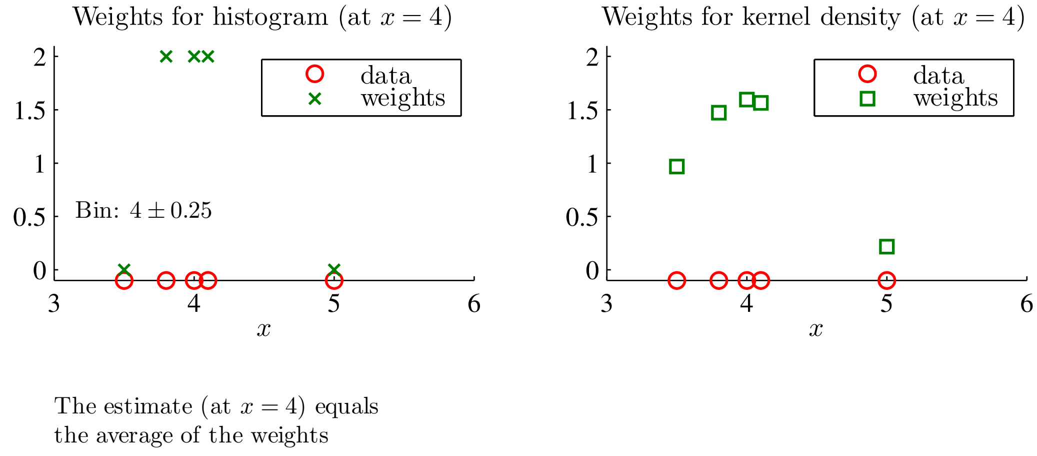


Figure 12.11: Calculation of the pdf at $x = 4$

does) improves the properties. In fact, the $N(0, h^2)$ is often used. The kernel density estimator of the pdf at some point x is then

$$\hat{f}(x) = \frac{1}{T} \sum\nolimits_{t=1}^{T} \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - x}{h}\right)^2\right] \tag{12.11}$$

Notice that the function in the summation is the density function of a $N(x, h^2)$ distribution.

The value $h = \text{Std}(x_t)1.06T^{-1/5}$ is sometimes recommended, since it can be shown to be the optimal choice (in MSE sense) if data is normally distributed. The bandwidth h could also be chosen by a leave-one-out cross-validation technique.

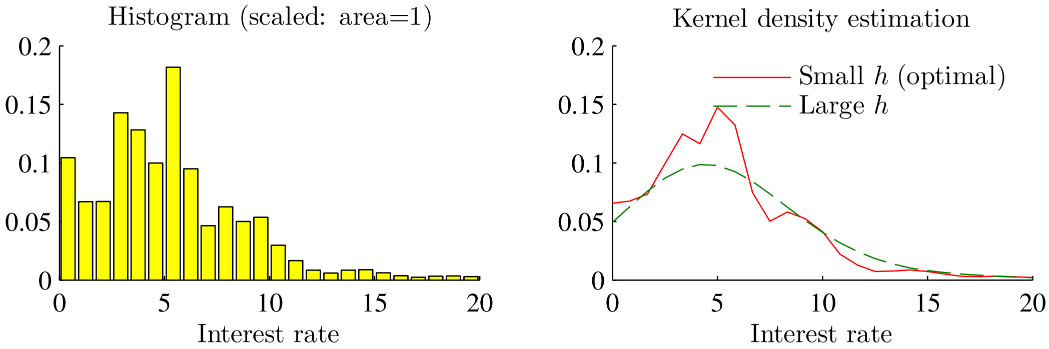
See Figure 12.12 for an example and Figure 12.13 for a QQ plot which is a good way to visualize the difference between the empirical and a given theoretical distribution.

It can be shown that (with iid data and a Gaussian kernel) the asymptotic distribution is

$$\sqrt{Th}\left[\hat{f}(x) - \mathrm{E}\,\hat{f}(x)\right] \to^{d} N\left[0, \frac{1}{2\sqrt{\pi}} f(x)\right], \tag{12.12}$$

The easiest way to handle a bounded support of x is to transform the variable into one with an unbounded support, estimate the pdf for this variable, and then use the "change of variable" technique to transform to the pdf of the original variable.

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



Daily federal funds rates 1954:7-2013:4
K-S (against N(μ,σ^2)) : $\sqrt{T}D = 14.1$
Skewness: 1.1
kurtosis: 4.9
Bera-Jarque: 7244.7

Figure 12.12: Federal funds rate

Gaussian kernel as

$$\hat{f}(x) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \frac{1}{(2\pi)^{d/2}|H^2\hat{\Omega}|^{1/2}} \exp\left[-\frac{1}{2}(x_t - x)'(H^2\hat{\Omega})^{-1}(x_t - x)\right]. \quad (12.13)$$

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 12.6 *((12.13) with $d = 1$) With just one variable, (12.13) becomes*

$$\hat{f}(x) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \frac{1}{H\,\mathrm{Std}(x_t)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - x}{H\,\mathrm{Std}(x_t)}\right)^2\right],$$

which is the same as (12.11) if $h = H\,\mathrm{Std}(x_t)$.

12.1.7 "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" (TA). This paper uses kernel regressions to identify and implement some technical trading rules, and then tests if the distribution (of the return) after a signal is the same as the unconditional distribution

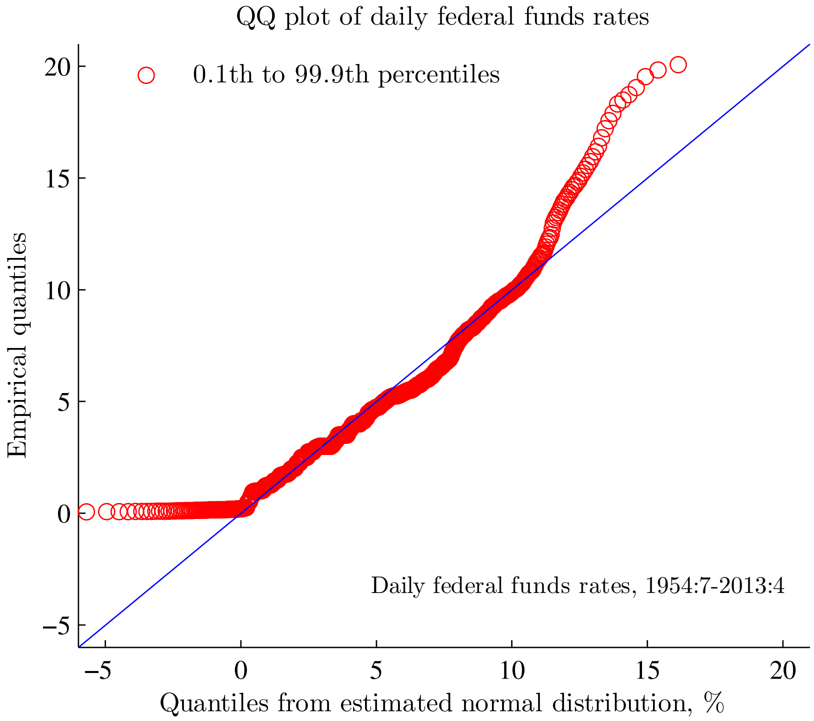


Figure 12.13: Federal funds rate

(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 12.14–12.15 for an illustration.

12.2 Tail Distribution

Reference: McNeil, Frey, and Embrechts (2005) 7, Alexander (2008) 3

In risk control, the focus is the distribution of losses beyond some threshold level. This has three direct implications. First, the object under study is the loss

$$X = -R, \tag{12.14}$$

that is, the negative of the return. Second, the attention is on how the distribution looks like beyond a threshold and also on the the probability of exceeding this threshold. In contrast, the exact shape of the distribution below that point is typically disregarded. Third,

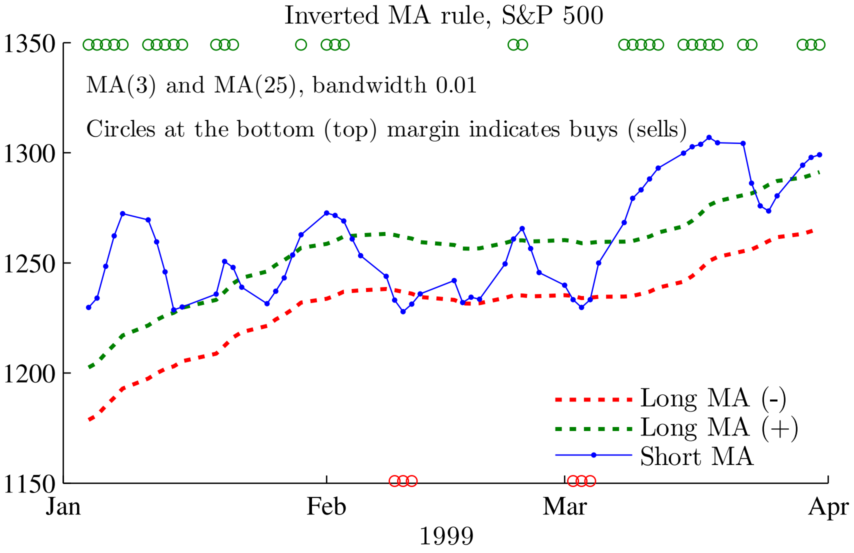


Figure 12.14: Examples of trading rules

modelling the tail of the distribution is best done by using a distribution that allows for a much heavier tail that suggested by a normal distribution. The generalized Pareto (GP) distribution is often used.

12.2.1 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 12.16 for an illustration.

Remark 12.7 *(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*

$$G(z) = \begin{cases} 1 - (1 + \xi z/\beta)^{-1/\xi} & \textit{if } \xi \neq 0 \\ 1 - \exp(-z/\beta) & \xi = 0, \end{cases}$$

for $0 \leq z$ if $\xi \geq 0$ and $z \leq -\beta/\xi$ in case $\xi < 0$. The pdf is therefore

$$g(z) = \begin{cases} \frac{1}{\beta}(1 + \xi z/\beta)^{-1/\xi - 1} & \textit{if } \xi \neq 0 \\ \frac{1}{\beta}\exp(-z/\beta) & \xi = 0. \end{cases}$$

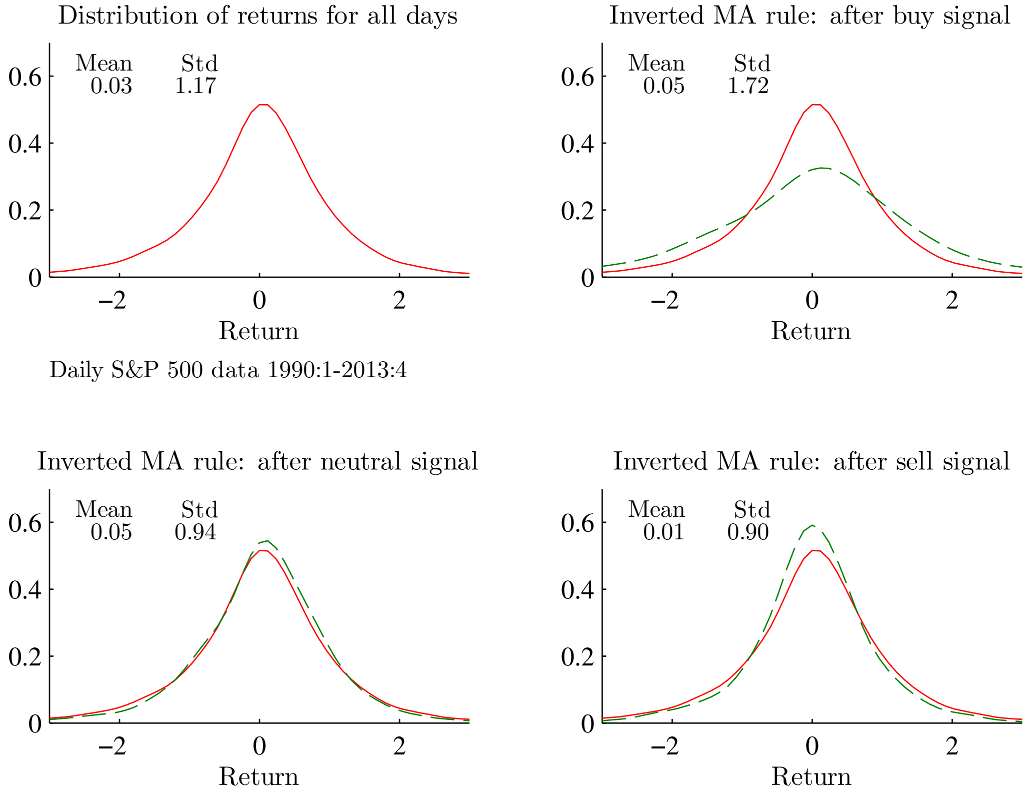


Figure 12.15: Examples of trading rules

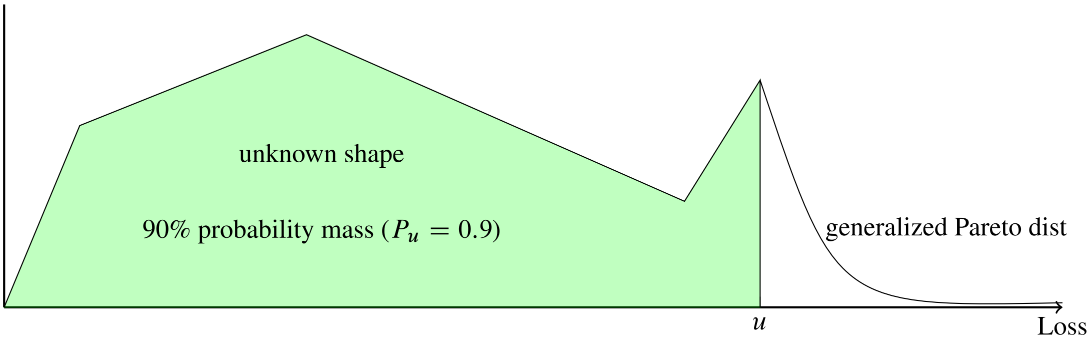


Figure 12.16: Loss distribution

The mean is defined (finite) if $\xi < 1$ *and is then* $\mathrm{E}(z) = \beta/(1-\xi)$*. Similarly, the variance is finite if* $\xi < 1/2$ *and is then* $\mathrm{Var}(z) = \beta^2/[(1-\xi)^2(1-2\xi)]$*. See Figure 12.17 for an illustration.*

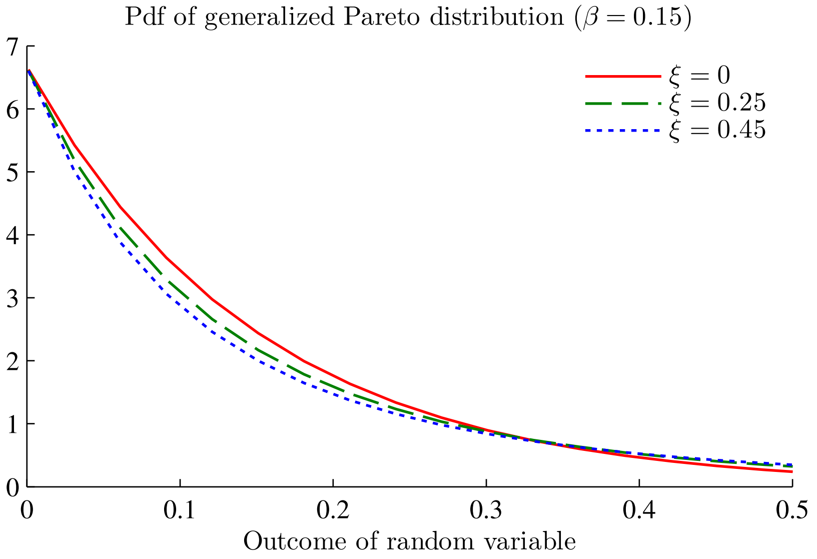


Figure 12.17: Generalized Pareto distributions

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, $X \leq u$. Then, the cdf of the loss for values greater than the threshold (Pr$(X \leq x)$ for $x > u$) can be written

$$\Pr(X \leq x) = F(x) = P_u + G(x - u)(1 - P_u), \text{ for } x > u, \tag{12.15}$$

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

$$f(x) = g(x - u)(1 - P_u), \text{ for } x > u, \tag{12.16}$$

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 12.16 and 12.19 for illustrations.

Remark 12.8 *(Random number from a generalized Pareto distribution*) By inverting the cdf of the generalized Pareto distribution, we can notice that if u is uniformly distributed* *on* $(0, 1]$, *then we can construct random variables with a GP distribution by*

$$\begin{array}{ll} z = \frac{\beta}{\xi}[(1-u)^{-\xi} - 1] & \textit{if } \xi \neq 0 \\ z = -\ln(1-u)\beta & \xi = 0. \end{array}$$

In addition, if z *is the threshold exceedance* ($z = X - u$), *then the loss (conditional on being above the threshold* u) *can be generated as* $X = u + z$.

It is often useful to calculate the *tail probability* $\Pr(X > x)$, which in the case of the cdf in (12.15) is

$$\Pr(X > x) = 1 - F(x) = (1 - P_u)[1 - G(x - u)], \tag{12.17}$$

where $G(z)$ is the cdf of the generalized Pareto distribution.

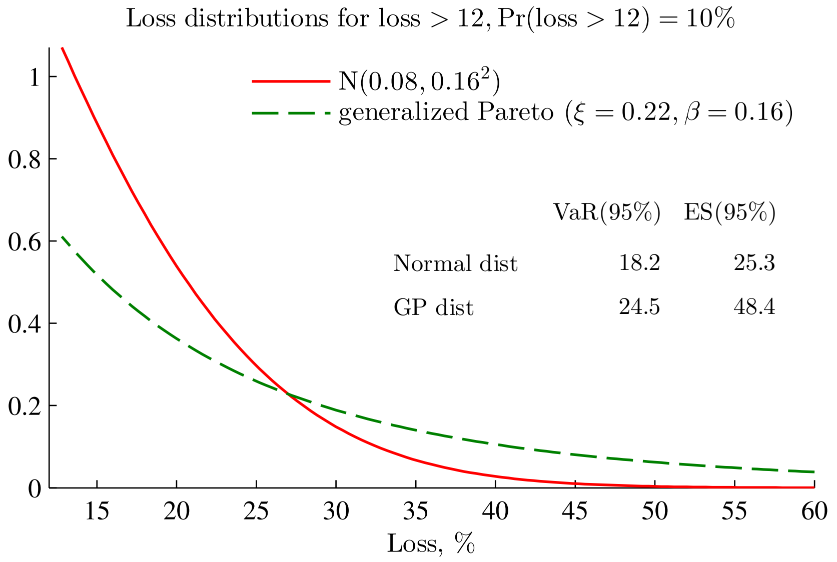


Figure 12.18: Comparison of a normal and a generalized Pareto distribution for the tail of losses

12.2.2 VaR and Expected Shortfall of a GP Distribution

The *value at risk*, VaR_α (say, $\alpha = 0.95$), is the α-th quantile of the loss distribution

$$\text{VaR}_\alpha = \text{cdf}_X^{-1}(\alpha), \tag{12.18}$$

where $\text{cdf}_X^{-1}()$ is the inverse cumulative distribution function of the losses, so $\text{cdf}_X^{-1}(\alpha)$ 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

$$\Pr(X \leq \text{VaR}_\alpha) = \alpha. \tag{12.19}$$

(Equivalently, the $\Pr(X > \text{VaR}_\alpha) = 1 - \alpha$.)

Assuming α is higher than P_u (so $\text{VaR}_\alpha \geq u$), the cdf (12.15) together with the form of the generalized Pareto distribution give the VaR

$$\text{VaR}_\alpha = \begin{cases} u + \frac{\beta}{\xi}\left[\left(\frac{1-\alpha}{1-P_u}\right)^{-\xi} - 1\right] & \text{if } \xi \neq 0 \\ u - \beta \ln\left(\frac{1-\alpha}{1-P_u}\right) & \xi = 0 \end{cases}, \text{ for } \alpha \geq P_u. \tag{12.20}$$

Proof. (of (12.20)) Set $F(x) = \alpha$ in (12.15) and use $z = x - u$ in the cdf from Remark 12.7 and solve for x. ■

If we assume $\xi < 1$ (to make sure that the mean is finite), then straightforward integration using (12.16) shows that the *expected shortfall* is

$$\begin{aligned} \text{ES}_\alpha &= \text{E}(X|X \geq \text{VaR}_\alpha) \\ &= \frac{\text{VaR}_a}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \text{ for } \alpha > P_u \text{ and } \xi < 1. \end{aligned} \tag{12.21}$$

12.2.3 Expected Exceedance of a GP Distribution

To locate the cut-off level where the tail "starts," that is, to choose the value of u. It often helps to study the *expected exceedance*.

The average exceedance (in data) over some threshold level υ is the mean of $X_t - \upsilon$ for those observations where $X_t > \upsilon$

$$\hat{e}(\upsilon) = \frac{\sum_{t=1}^{T}(X_t - \upsilon)\delta(X_t > \upsilon)}{\sum_{t=1}^{T}(X_t > \upsilon)}, \text{ where} \tag{12.22}$$
$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

The expected exceedance of a GD distribution is easily found by letting $\upsilon = \text{VaR}_\alpha$ in the expected shortfall (12.21) and then subtract υ from both sides to get the expected exceedance of the loss over another threshold $\upsilon > u$

$$\begin{aligned} e(\upsilon) &= \mathrm{E}\left(X - \upsilon | X > \upsilon\right) \\ &= \frac{\xi\upsilon}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \text{ for } \upsilon > u \text{ and } \xi < 1. \end{aligned} \tag{12.23}$$

The expected exceedance of a generalized Pareto distribution (with $\xi > 0$) is increasing with the threshold level υ. This indicates that the tail of the distribution is very long. In contrast, a normal distribution would typically show a negative relation (see Figure 12.19for an illustration). This provides a way of assessing which distribution that 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 υ) after which the average exceedance in data (12.22) appears to be a linear function of the threshold.

Remark 12.9 *(Expected exceedance from a normal distribution) If $X \sim N(\mu, \sigma^2)$, then*

$$\begin{aligned} \mathrm{E}(X - \upsilon | X > \upsilon) &= \mu + \sigma \frac{\phi(\upsilon_0)}{1 - \Phi(\upsilon_0)} - \upsilon, \\ \textit{with } \upsilon_0 &= (\upsilon - \mu)/\sigma \end{aligned}$$

where $\phi()$ and Φ are the pdf and cdf of a $N(0,1)$ variable respectively.

12.2.4 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 (12.22) with the theoretical (12.23) can help. Suppose we calculate the empirical exceedance for different values of the threshold level (denoted υ_i—all large enough so the relation looks linear), then we can estimate (by LS)

$$\hat{e}(\upsilon_i) = a + b\upsilon_i + \varepsilon_i. \tag{12.24}$$

Then, the theoretical exceedance (12.23) for a given starting point of the GP distribution (u) is related to this regression according to

$$a = \frac{\beta - \xi u}{1 - \xi} \text{ and } b = \frac{\xi}{1 - \xi}\text{, or}$$
$$\xi = \frac{b}{1 + b} \text{ and } \beta = a(1 - \xi) + \xi u. \qquad (12.25)$$

See Figure 12.20 for an illustration.

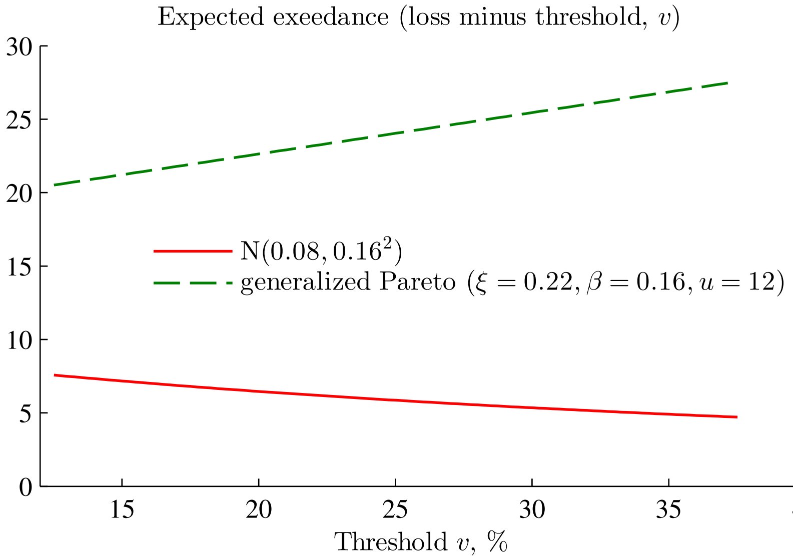


Figure 12.19: Expected exceedance, normal and generalized Pareto distribution

Remark 12.10 *(Log likelihood function of the loss distribution) Since we have assumed that the threshold exceedance $(X - u)$ has a generalized Pareto distribution, Remark 12.7 shows that the log likelihood for the observation of the loss above the threshold $(X_t > u)$ is*

$$L = \sum_{t \text{ st. } X_t > u} L_t$$
$$\ln L_t = \begin{cases} -\ln\beta - (1/\xi + 1)\ln\left[1 + \xi\left(X_t - u\right)/\beta\right] & \text{if } \xi \neq 0 \\ -\ln\beta - \left(X_t - u\right)/\beta & \xi = 0. \end{cases}$$

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).

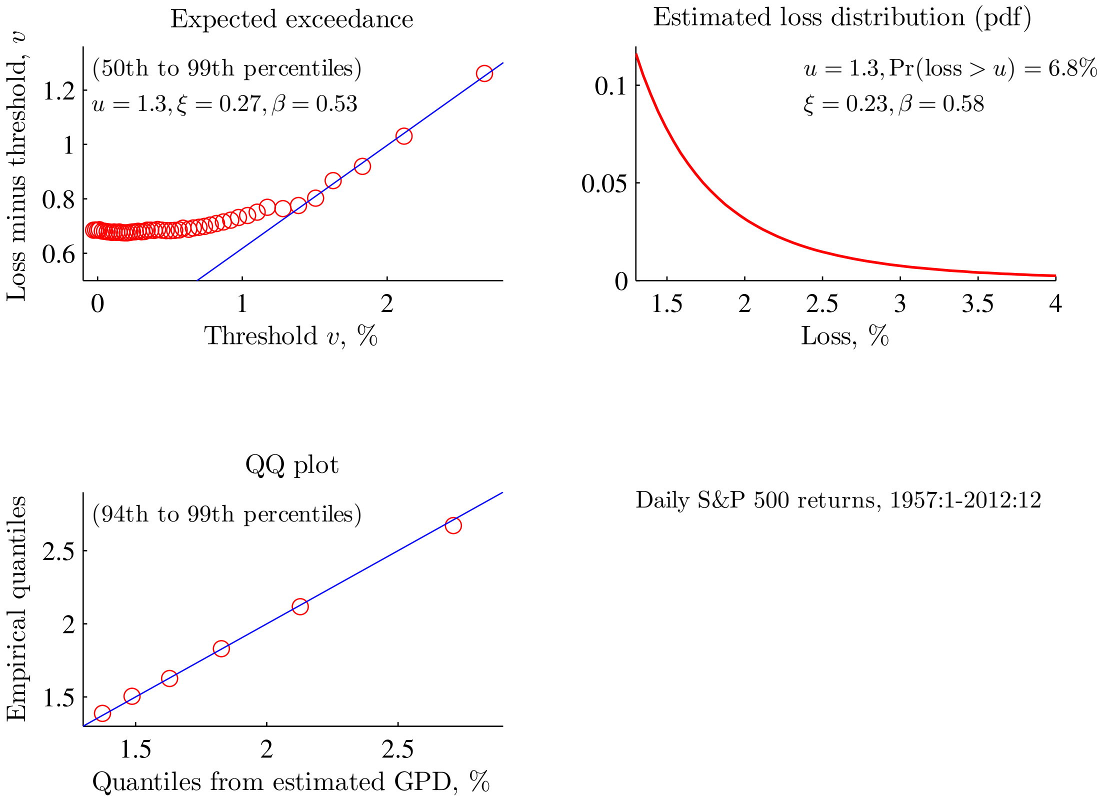


Figure 12.20: Results from S&P 500 data

Example 12.11 *(Estimation of the generalized Pareto distribution on S&P daily returns). Figure 12.20 (upper left panel) shows that it may be reasonable to fit a GP distribution with a threshold $u = 1.3$. The upper right panel illustrates the estimated distribution, while the lower left panel shows that the highest quantiles are well captured by estimated distribution.*

Bibliography

Alexander, C., 2008, *Market Risk Analysis: Value at Risk Models*, Wiley.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and inference in econometrics*, Oxford University Press, Oxford.

DeGroot, M. H., 1986, *Probability and statistics*, Addison-Wesley, Reading, Massachusetts.

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.

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.

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.

Silverman, B. W., 1986, *Density estimation for statistics and data analysis*, Chapman and Hall, London.

13 Return Distributions (Multivariate)*

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

13.1 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,

$$u_i = F_i(x_i) = \Pr(X_i \leq x_i). \tag{13.1}$$

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

$$F_i(x_i) = \int_{s=-\infty}^{x_i} f_i(s)ds. \tag{13.2}$$

(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 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.*

13.2 Exceedance Correlations

Reference: Ang and Chen (2002)

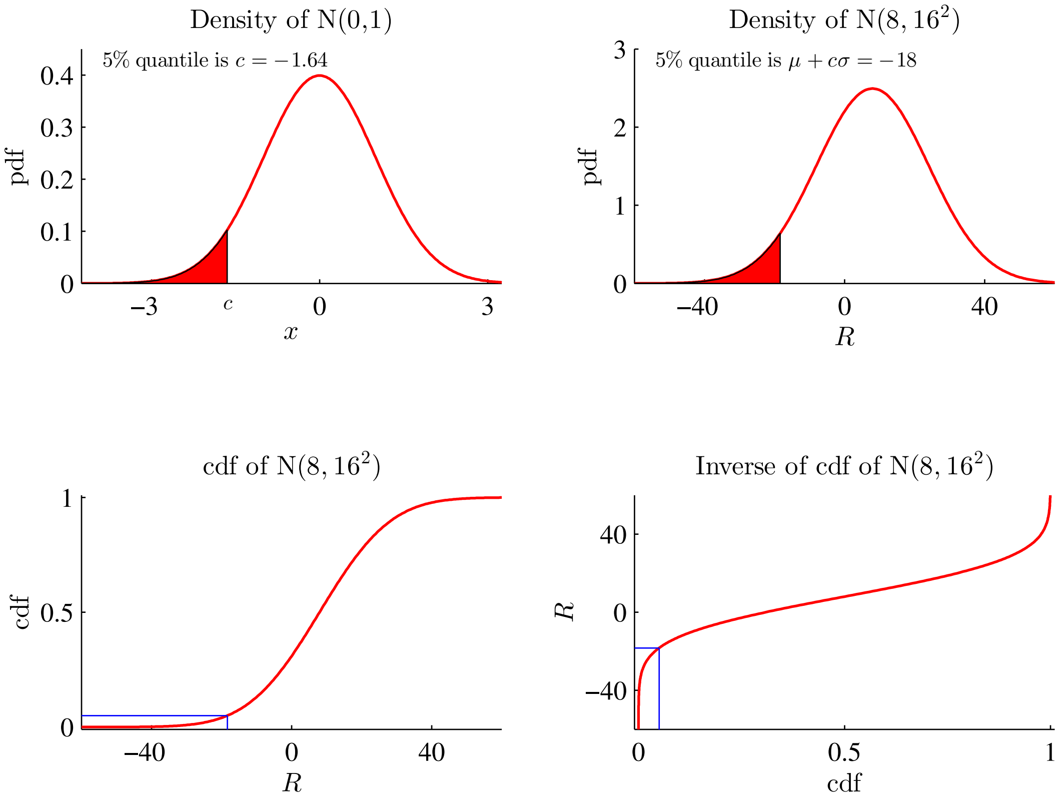


Figure 13.1: Finding quantiles of a N(μ,σ^2) distribution

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

$$\text{Corr}(x_t, y_t | h_1 < x_t \leq h_2, k_1 < y_t \leq k_2). \tag{13.3}$$

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 very little probability mass at low returns, which leads to the correlation being squeezed towards zero as we only consider data far out in the tail. In short, the tail correlation of a normal distribution is always closer to zero than the correlation for all data points. This is illustrated in Figure 13.2.

In contrast, Figures 13.3–13.4 suggest (for two US portfolios) that the correlation in the lower tail is almost as high as for all the data and considerably higher than for the upper tail. 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.

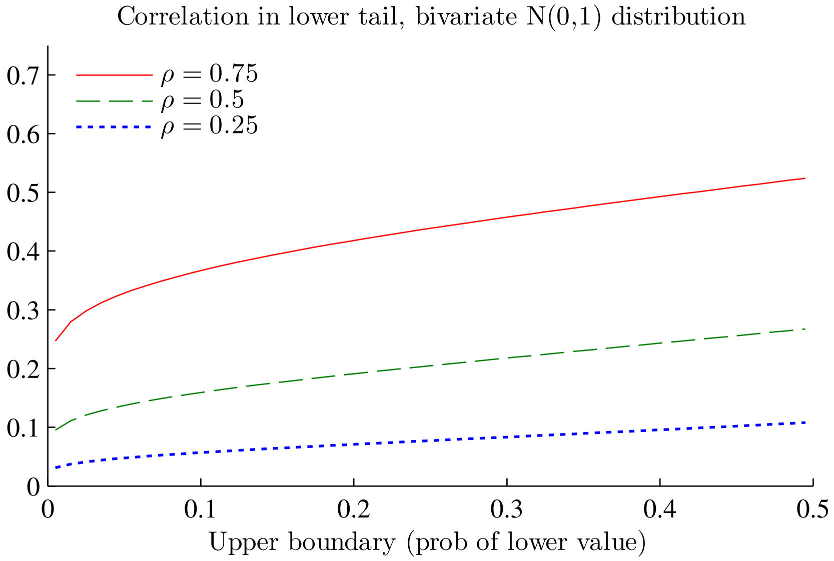


Figure 13.2: Correlation in lower tail when data is drawn from a normal distribution with correlation ρ

13.3 Beyond (Linear) Correlations

Reference: Alexander (2008) 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 go beyond the correlation—to capture non-linear

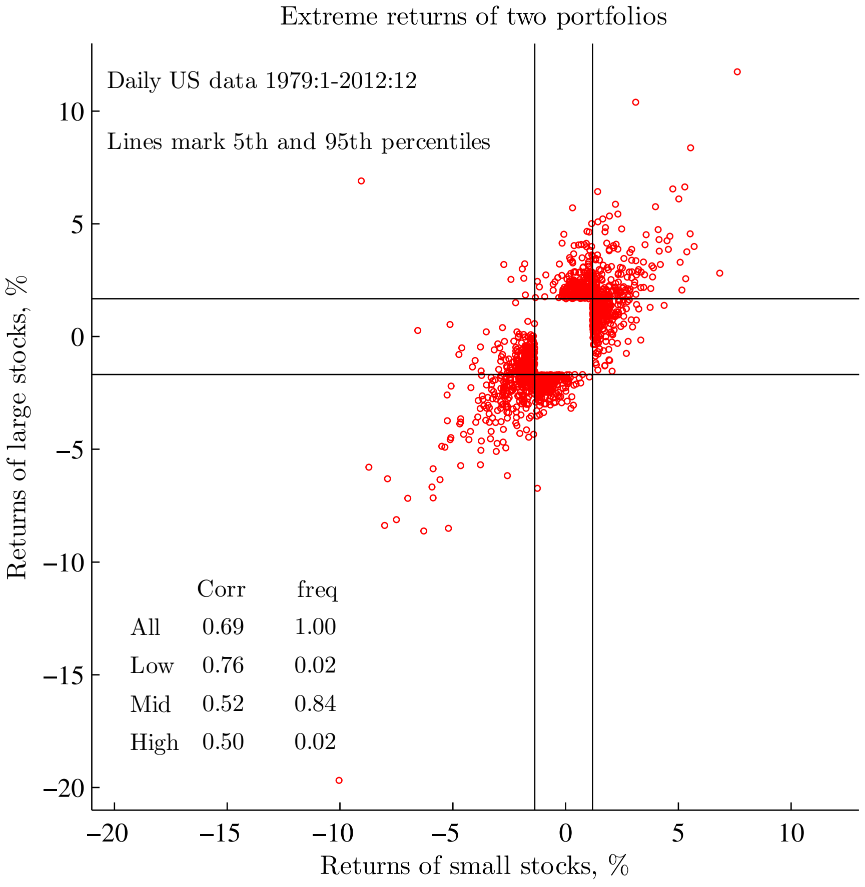


Figure 13.3: Correlation of two portfolios

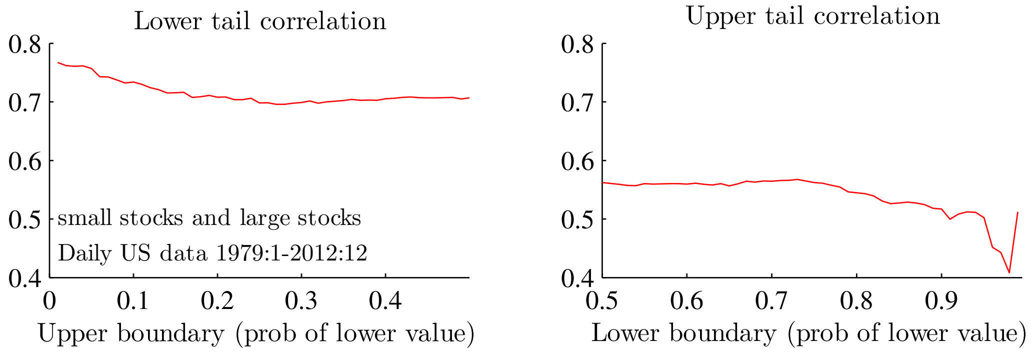


Figure 13.4: Correlation in the tails for two portfolios

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) of two variables measures to what degree their relation is monotonic: 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.

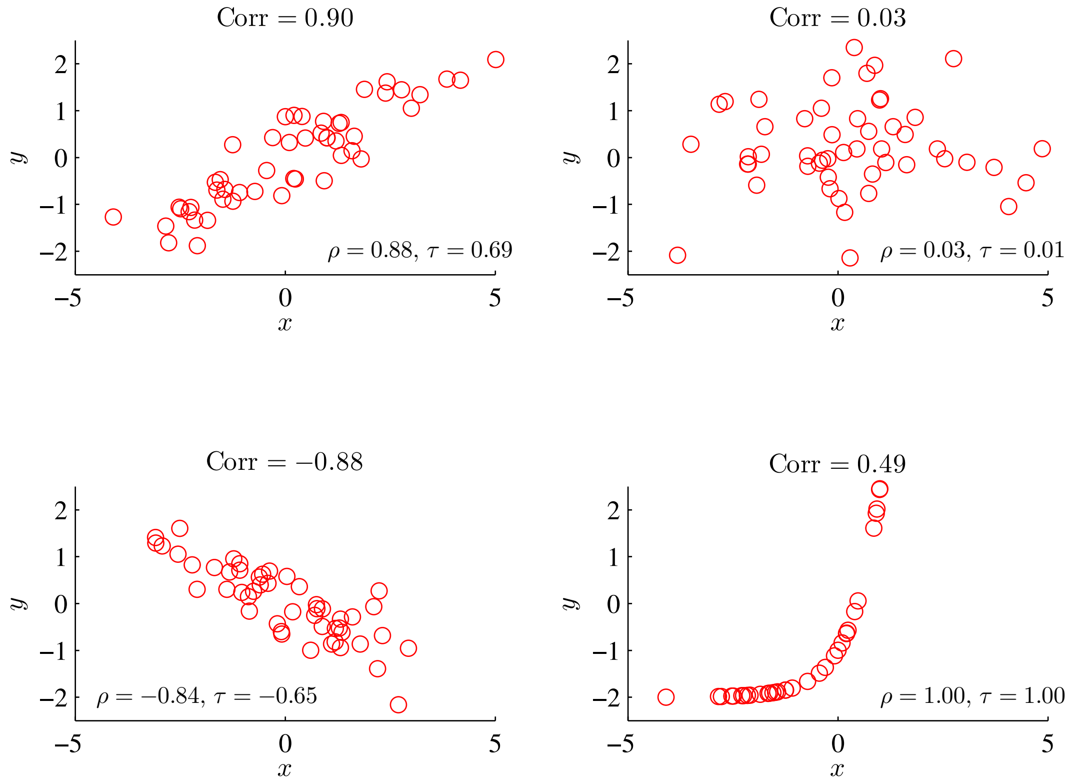


Figure 13.5: Illustration of correlation and rank correlation

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

(13.4)

In the second step, simply estimate the correlation of the ranks of two variables

$$\text{Spearman's } \rho = \text{Corr}[\text{rank}(x_t), \text{rank}(y_t)]. \tag{13.5}$$

Clearly, this correlation is between -1 and 1. (There is an alternative way of calculating the rank correlation based on the difference of the ranks, d_t =rank(x_t)−rank(y_t), $\rho = 1 - 6\Sigma_{t=1}^{T} d_t^2/(T^3 - T)$. It gives the same result if there are no tied ranks.) See Figure 13.5 for an illustration.

The rank correlation can be tested by using the fact that under the null hypothesis the rank correlation is zero. We then get

$$\sqrt{T-1}\hat{\rho} \to^{d} N(0,1). \tag{13.6}$$

(For samples of 20 to 40 observations, it is often recommended to use $\sqrt{(T-2)/(1-\hat{\rho}^2)}\hat{\rho}$ which has an t_{T-2} distribution.)

Remark 13.2 *(Spearman's ρ for a distribution*) If we have specified the joint distribution of the random variables X and Y, then we can also calculate the implied Spearman's ρ (sometimes only numerically) as* Corr$[F_X(X), F_Y(Y)]$ *where $F_X(X)$ is the cdf of X and $F_Y(Y)$ of Y.*

Kendall's rank correlation (called Kendall's τ) is similar, but is based on comparing changes of x_t (compared to $x_1, \ldots x_{t-1}$) 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
$x_2 - x_1$	$y_2 - y_1$
$x_3 - x_1$	$y_3 - y_1$
$x_3 - x_2$	$y_3 - y_2$,

(13.7)

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

$$\begin{aligned} ij \text{ is concordant if } (x_j - x_i)(y_j - y_i) &> 0 \\ ij \text{ is discordant if } (x_j - x_i)(y_j - y_i) &< 0. \end{aligned} \tag{13.8}$$

Finally, we count the number of concordant (T_c) and discordant (T_d) pairs and calculate Kendall's tau as

$$\text{Kendall's } \tau = \frac{T_c - T_d}{T(T-1)/2}. \tag{13.9}$$

It can be shown that

$$\text{Kendall's } \tau \to^d N\left(0, \frac{4T+10}{9T(T-1)}\right), \tag{13.10}$$

so it is straightforward to test τ by a t-test.

Example 13.3 *(Kendall's tau) Suppose the data is*

x	y
2	7
10	9
−3	10.

We then get the following changes

Changes of x	*Changes of y*	
$x_2 - x_1 = 10 - 2 = 8$	$y_2 - y_1 = 9 - 7 = 2$	*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

$$\text{Spearman's rho} = \frac{6}{\pi}\arcsin(\rho/2) \approx \rho \tag{13.11}$$

$$\text{Kendall's tau} = \frac{2}{\pi}\arcsin(\rho). \tag{13.12}$$

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 13.6. Clearly, when data is not normally distributed, then these measures can give distinctly different answers.

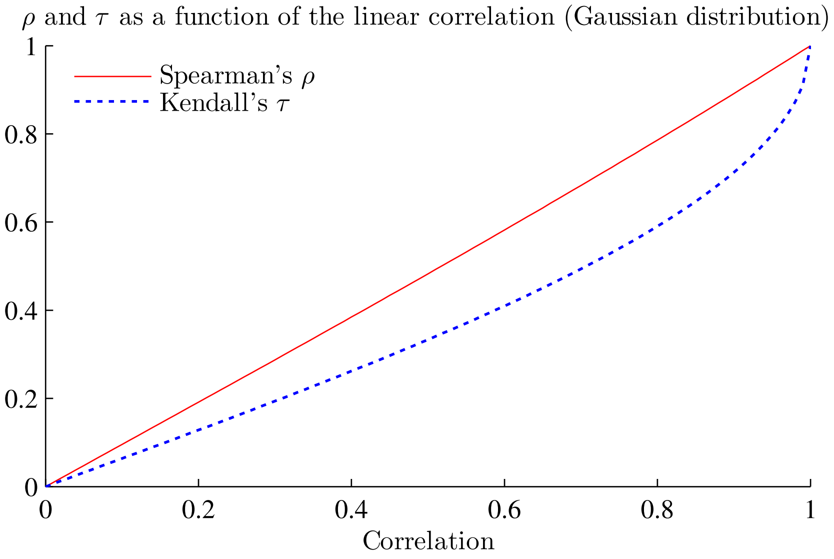


Figure 13.6: Spearman's rho and Kendall's tau if data has a bivariate normal distribution

A *joint α-quantile exceedance probability* measures how often two random variables (x and y, say) are both above their α quantile. Similarly, we can also define the probability that they are both below their α quantile

$$G_\alpha = \Pr(x \leq \xi_{x,\alpha}, y \leq \xi_{y,\alpha}), \tag{13.13}$$

$\xi_{x,\alpha}$ 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

$$\hat{G}_\alpha = \frac{1}{T}\sum\nolimits_{t=1}^{T} \delta_t, \text{ where } \delta_t = \begin{cases} 1 \text{ if } x_t \leq \hat{\xi}_{x,\alpha} \text{ and } y_t \leq \hat{\xi}_{y,\alpha} \\ 0 \text{ otherwise.} \end{cases} \tag{13.14}$$

See Figure 13.7 for an illustration based on a joint normal distribution.

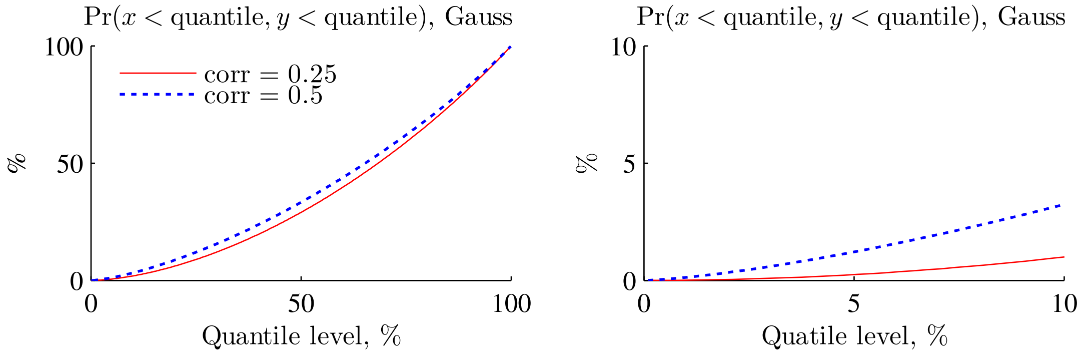


Figure 13.7: Probability of joint low returns, bivariate normal distribution

13.4 Copulas

Reference: McNeil, Frey, and Embrechts (2005), Alexander (2008) 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 (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 returns separately, then we model the comovements by a copula.

13.4.1 Multivariate Distributions and Copulas

Any pdf can also be written as

$$f_{1,2}(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2), \text{ with } u_i = F_i(x_i), \tag{13.15}$$

where $c()$ is a *copula density* function and $u_i = F_i(x_i)$ is the cdf value as in (13.1). The extension to three or more random variables is straightforward.

Equation (13.15) 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 must be. Alternatively, if we know the pdfs $f_1(x_1)$ and $f_2(x_2)$—and thus also the cdfs $F_1(x_1)$ and $F_2(x_2)$—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.

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 13.4 *(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 13.5 *(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 13.8 for an illustration.

Remark 13.6 *(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 $f_1(x_1) = \int_{x_2=-\infty}^{\infty} f_{1,2}(x_1, x_2) dx_2$.*

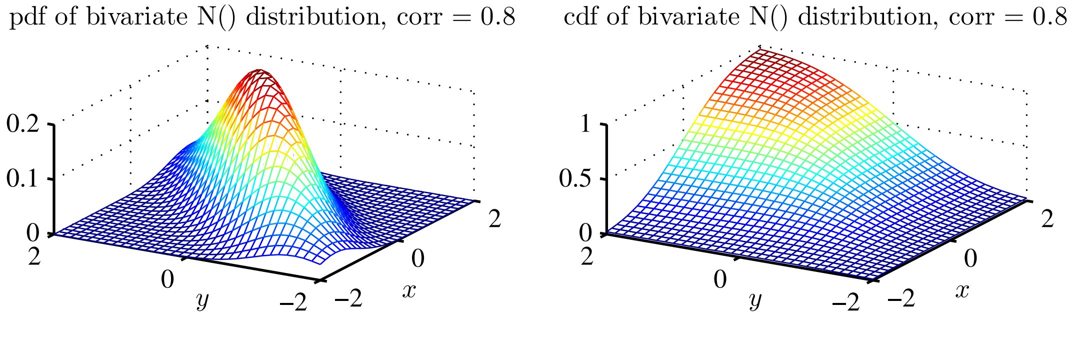


Figure 13.8: Bivariate normal distributions

Remark 13.7 *(Joint pdf and copula density, n variables) For n variables (13.15) generalizes to*

$$f_{1,2,\ldots,n}(x_1, x_2, \ldots, x_n) = c(u_1, u_2, \ldots, u_n) f_1(x_1) f_2(x_2) \ldots f_n(x_n), \text{ with } u_i = F_i(x_i),$$

Remark 13.8 *(Cdfs and copulas*) The joint cdf can be written as*

$$F_{1,2}(x_1, x_2) = C[F_1(x_1), F_2(x_2)],$$

where $C()$ is the unique copula function. Taking derivatives gives (13.15) 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 $C()$.

13.4.2 The Gaussian and Other Copula Densities

The *Gaussian copula density function* is

$$c(u_1, u_2) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2\xi_1^2 - 2\rho\xi_1\xi_2 + \rho^2\xi_2^2}{2(1-\rho^2)}\right), \text{ with } \xi_i = \Phi^{-1}(u_i), \quad (13.16)$$

where $\Phi^{-1}()$ is the inverse of an $N(0,1)$ distribution. Notice that when using this function in (13.15) to construct the joint pdf, we have to first calculate the cdf values $u_i = F_i(x_i)$ from the univariate distribution of x_i (which may be non-normal) and then calculate the quantiles of those according to a standard normal distribution $\xi_i = \Phi^{-1}(u_i) = \Phi^{-1}[F_i(x_i)]$.

It can be shown that assuming that the marginal pdfs ($f_1(x_1)$ and $f_2(x_2)$) are normal and then combining with the Gaussian copula density recovers a bivariate normal distribution. However, the way we typically use copulas is to assume (and estimate) some other type of univariate distribution, for instance, with fat tails—and then combine with a (Gaussian) copula density to create the joint distribution. See Figure 13.9 for an illustration.

A zero correlation ($\rho = 0$) makes the copula density (13.16) 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

$$\rho = \text{Spearman's rho}, \tag{13.17}$$

as suggested by (13.11).

Alternatively, the ρ parameter can calibrated to give a joint probability of both x_1 and x_2 being lower than some quantile as to match data: see (13.14). The values of this probability (according to a copula) is easily calculated by finding the copula function (essentially the cdf) corresponding to a copula density. Some results are given in remarks below. See Figure 13.7 for results from a Gaussian copula. This 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).

Remark 13.9 *(The Gaussian copula function*) The distribution function corresponding to the Gaussian copula density (13.16) is obtained by integrating over both u_1 and u_2 and the value is $C(u_1, u_2; \rho) = \Phi_\rho(\xi_1, \xi_2)$ where ξ_i is defined in (13.16) and Φ_ρ is the bivariate normal cdf for $N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$. Most statistical software contains numerical returns for calculating this cdf.*

Remark 13.10 *(Multivariate Gaussian copula density*) The Gaussian copula density for n variables is*

$$c(u) = \frac{1}{\sqrt{|R|}} \exp\left[-\frac{1}{2}\xi'(R^{-1} - I_n)\xi\right],$$

where R is the correlation matrix with determinant $|R|$ and ξ is a column vector with $\xi_i = \Phi^{-1}(u_i)$ as the ith element.

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

$$c(u_1, u_2) = (-1 + u_1^{-\alpha} + u_2^{-\alpha})^{-2-1/\alpha}(u_1 u_2)^{-\alpha-1}(1 + \alpha), \tag{13.18}$$

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 the tail).

See Figure 13.9 for an illustration.

For the Clayton copula we have

$$\text{Kendall's } \tau = \frac{\alpha}{\alpha+2}, \text{ so} \tag{13.19}$$

$$\alpha = \frac{2\tau}{1-\tau}. \tag{13.20}$$

The easiest way to calibrate a Clayton copula is therefore to set the parameter α according to (13.20).

Figure 13.10 illustrates how the probability of both variables to be below their respective quantiles depend on the α parameter. These parameters are comparable to the those for the correlations in Figure 13.7 for the Gaussian copula, see (13.11)–(13.12). The figure 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 left tail. This could be interpreted as if the correlation is higher in market crashes than during normal times.

Remark 13.11 *(Multivariate Clayton copula density*) The Clayton copula density for n variables is*

$$c(u) = \left(1 - n + \textstyle\sum_{i=1}^{n} u_i^{-\alpha}\right)^{-n-1/\alpha} \left(\textstyle\prod_{i=1}^{n} u_i\right)^{-\alpha-1} \left(\textstyle\prod_{i=1}^{n} [1 + (i-1)\alpha]\right).$$

Remark 13.12 *(Clayton copula function*) The copula function (the cdf) corresponding to (13.18) is*

$$C(u_1, u_2) = (-1 + u_1^{-\alpha} + u_2^{-\alpha})^{-1/\alpha}.$$

The following steps summarize how the copula is used to construct the multivariate distribution.

1. Construct the marginal pdfs $f_i(x_i)$ and thus also the marginal cdfs $F_i(x_i)$. For instance, this could be done by fitting a distribution with a fat tail. With this, calculate the cdf values for the data $u_i = F_i(x_i)$ as in (13.1).

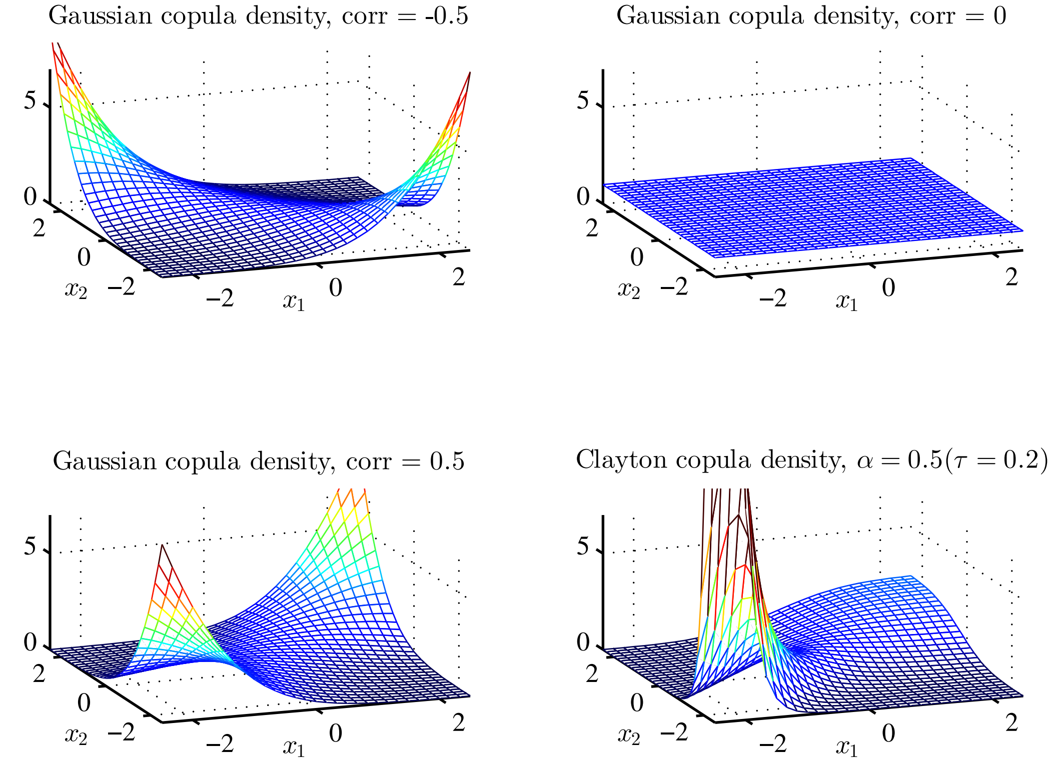


Figure 13.9: Copula densities (as functions of x_i)

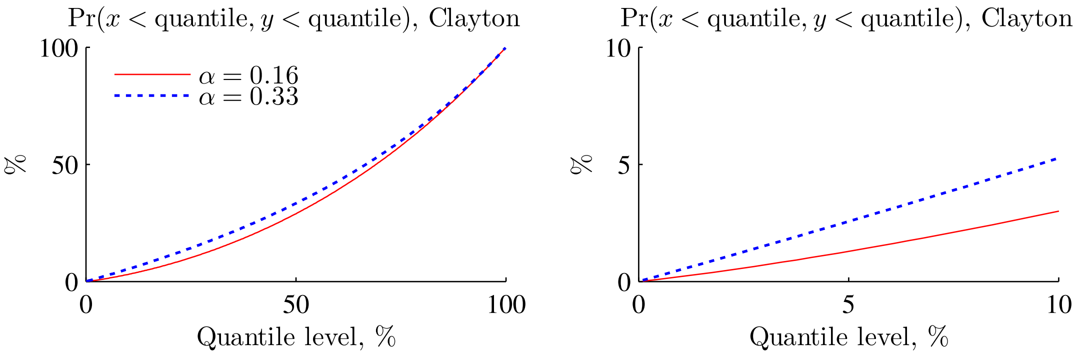


Figure 13.10: Probability of joint low returns, Clayton copula

2. Calculate the copula density as follows (for the Gaussian or Clayton copulas, respectively):

(a) for the Gaussian copula (13.16)

i. assume (or estimate/calibrate) a correlation ρ to use in the Gaussian copula

ii. calculate $\xi_i = \Phi^{-1}(u_i)$, where $\Phi^{-1}()$ is the inverse of a $N(0, 1)$ distribution

iii. combine to get the copula density value $c(u_1, u_2)$

(b) for the Clayton copula (13.18)

i. assume (or estimate/calibrate) an α to use in the Clayton copula (typically based on Kendall's τ as in (13.20))

ii. calculate the copula density value $c(u_1, u_2)$

3. Combine the marginal pdfs and the copula density as in (13.15), $f_{1,2}(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2)$, where $u_i = F_i(x_i)$ is the cdf value according to the marginal distribution of variable i.

See Figures 13.11–13.12 for illustrations.

Remark 13.13 *(Tail Dependence*) The measure of* lower tail dependence *starts by finding the probability that X_1 is lower than its qth quantile ($X_1 \leq F_1^{-1}(q)$) given that X_2 is lower than its qth 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

$$\lambda_l = \lim_{q \to 0} \Pr[X_1 \leq F_1^{-1}(q) | X_2 \leq F_2^{-1}(q)].$$

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.

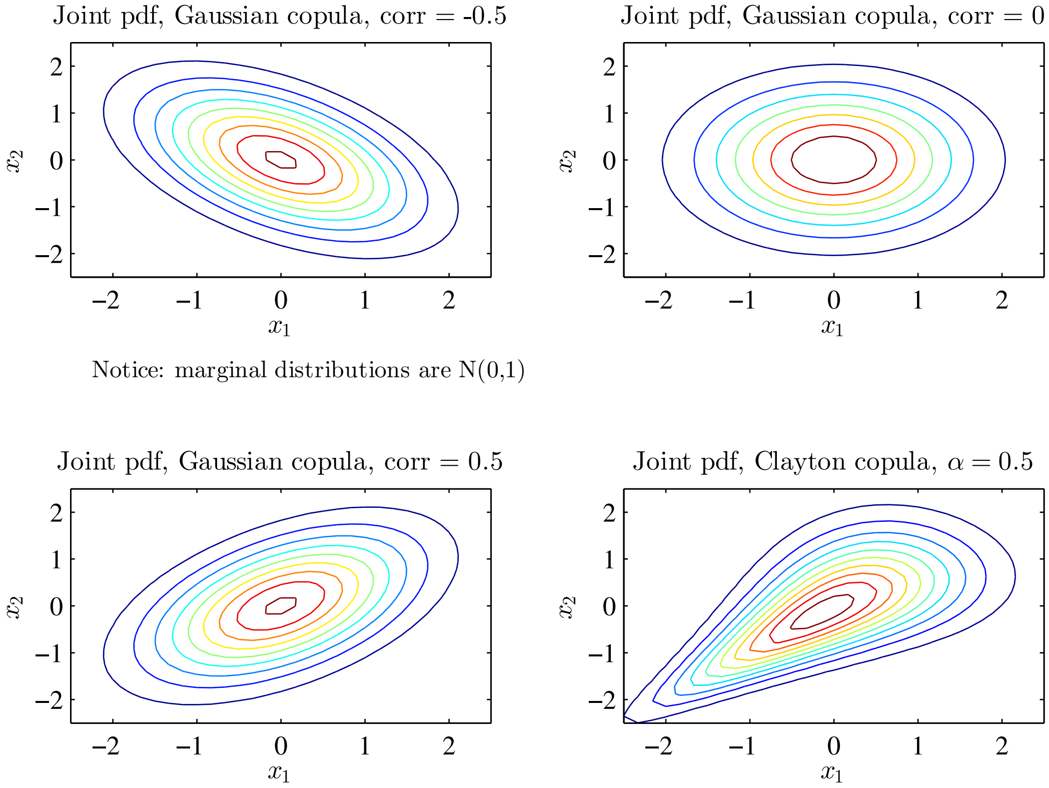


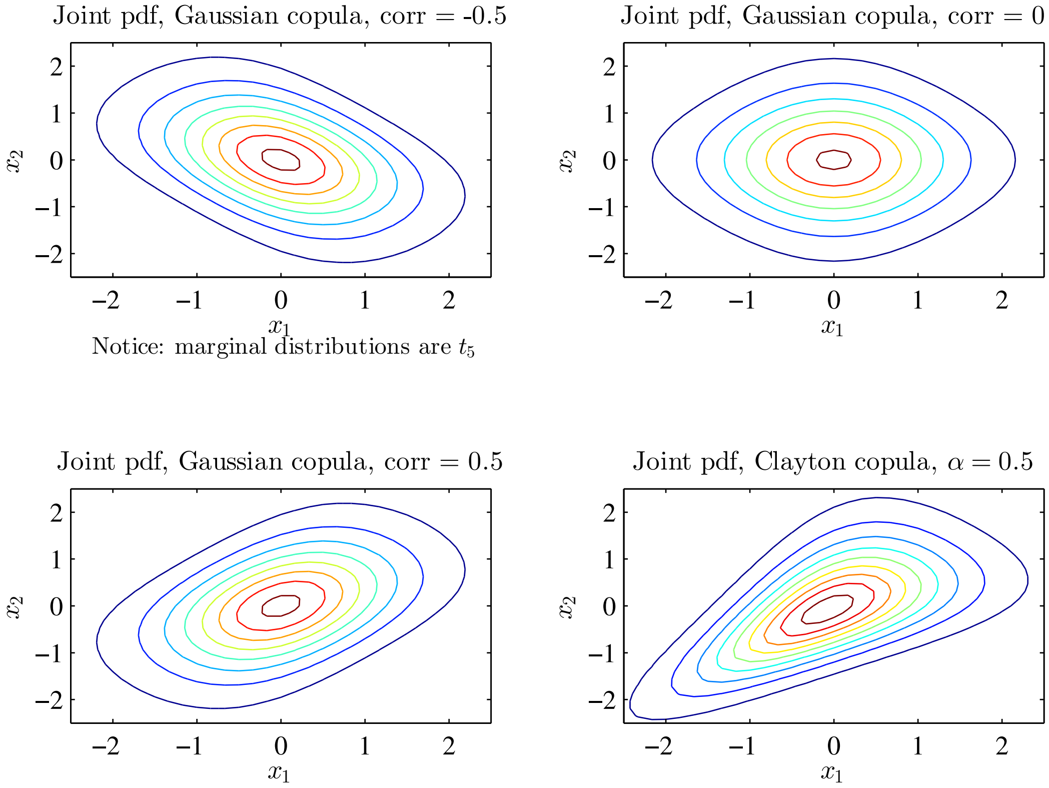
Figure 13.11: Contours of bivariate pdfs

13.5 Joint Tail Distribution

The methods for estimating the (marginal, that is, for one variable at a time) distribution of the lower tail can be combined with a copula to model the joint tail distribution. In particular, combining the generalized Pareto distribution (GPD) with the Clayton copula provides a flexible way.

This can be done by first modelling the loss ($X_t = -R_t$) beyond some threshold (u), that is, the variable $X_t - u$ with the GDP. To get a distribution of the return, we simply use the fact that $\text{pdf}_R(-z) = \text{pdf}_X(z)$ for any value z. Then, in a second step we calibrate the copula by using Kendall's τ for the subsample when both returns are less than u. Figures 13.13–13.15 provide an illustration.

Remark 13.14 *Figure 13.13 suggests that the joint occurrence (of these two assets) of really negative returns happens more often than the estimated normal distribution would*



Joint pdf, Gaussian copula, corr = -0.5

Joint pdf, Gaussian copula, corr = 0

Notice: marginal distributions are t_5

Joint pdf, Gaussian copula, corr = 0.5

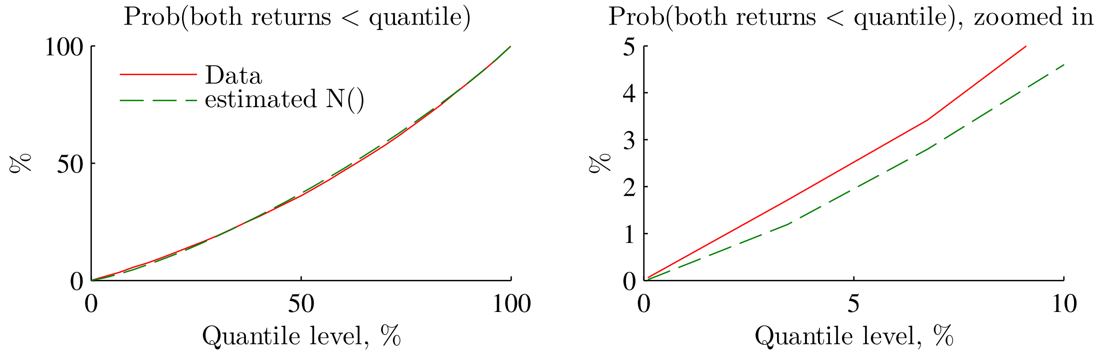
Joint pdf, Clayton copula, $\alpha = 0.5$

Figure 13.12: Contours of bivariate pdfs

suggest. For that reason, the joint distribution is estimated by first fitting generalized Pareto distributions to each of the series and then these are combined with a copula as in (13.15) to generate the joint distribution. In particular, the Clayton copula seems to give a long joint negative tail.

To find the implication for a portfolio of several assets with a given joint tail distribution, 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 very hard to actually calculate the distribution of the portfolio by using mathematics, so we have to rely on raw number crunching.

The approach proceeds in two steps. First, draw n values for the copula ($u_i, i = 1, \ldots, n$). Second, calculate the random number ("return") by inverting the cdf $u_i =$



Daily US data 1979:1-2012:12
small stocks and large stocks

Figure 13.13: Probability of joint low returns

$F_i(x_i)$ in (13.15) as

$$x_i = F_i^{-1}(u_i), \tag{13.21}$$

where $F_i^{-1}()$ is the inverse of the cdf.

Remark 13.15 *(To draw n random numbers from a Gaussian copula) First, draw n numbers from an $N(0, R)$ distribution, where R is the correlations matrix. Second, calculate $u_i = \Phi(x_i)$, where Φ is the cdf of a standard normal distribution.*

Remark 13.16 *(To draw n random numbers from a Clayton copula) First, draw x_i for $i = 1, \ldots, n$ from a uniform distribution (between 0 and 1). Second, draw v from a gamma($1/\alpha, 1$) distribution. Third, calculate $u_i = [1 - \ln(x_i)/v]^{-1/\alpha}$ for $i = 1, \ldots, n$. These u_i values are the marginal cdf values.*

Remark 13.17 *(Inverting a normal and a generalised Pareto cdf) Must numerical software packages contain a routine for investing a normal cdf. My lecture notes on the Generalised Pareto distribution shows how to invert that distribution.*

Such simulations can be used to quickly calculate the VaR and other risk measures for different portfolios. A Clayton copula with a high α parameter (and hence a high Kendall's τ) has long lower tail with highly correlated returns: when asset takes a dive,

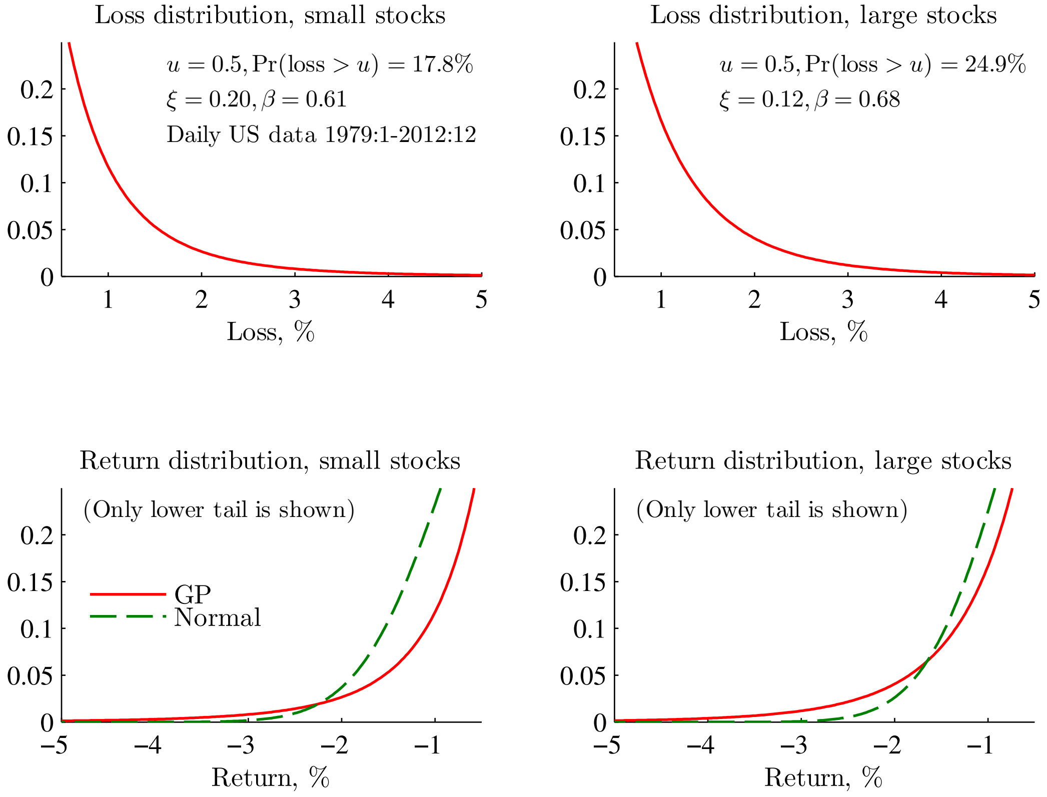


Figure 13.14: Estimation of marginal loss distributions

other assets are also likely to decrease. That is, the correlation in the lower tail of the return distribution is high, which will make the VaR high.

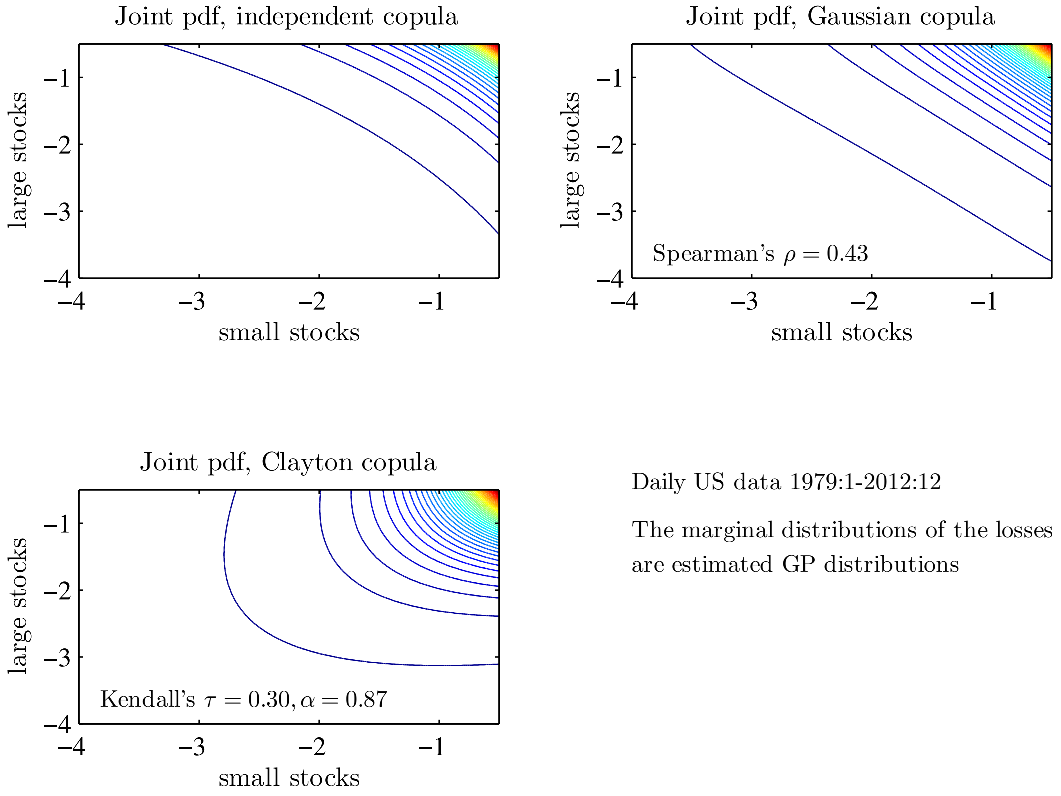
Figures 13.16–13.17 give an illustration of how the movements in the lower get more synchronised as the α parameter in the Clayton copula increases.

Bibliography

Alexander, C., 2008, *Market Risk Analysis: Practical Financial Econometrics*, Wiley.

Ang, A., and J. Chen, 2002, "Asymmetric correlations of equity portfolios," *Journal of Financial Economics*, 63, 443–494.

Jondeau, E., S.-H. Poon, and M. Rockinger, 2007, *Financial Modeling under Non-Gaussian Distributions*, Springer.


Joint pdf, independent copula

Joint pdf, Gaussian copula

Joint pdf, Clayton copula

Daily US data 1979:1-2012:12

The marginal distributions of the losses are estimated GP distributions

Figure 13.15: Joint pdfs with different copulas

McNeil, A. J., R. Frey, and P. Embrechts, 2005, *Quantitative risk management*, Princeton University Press.

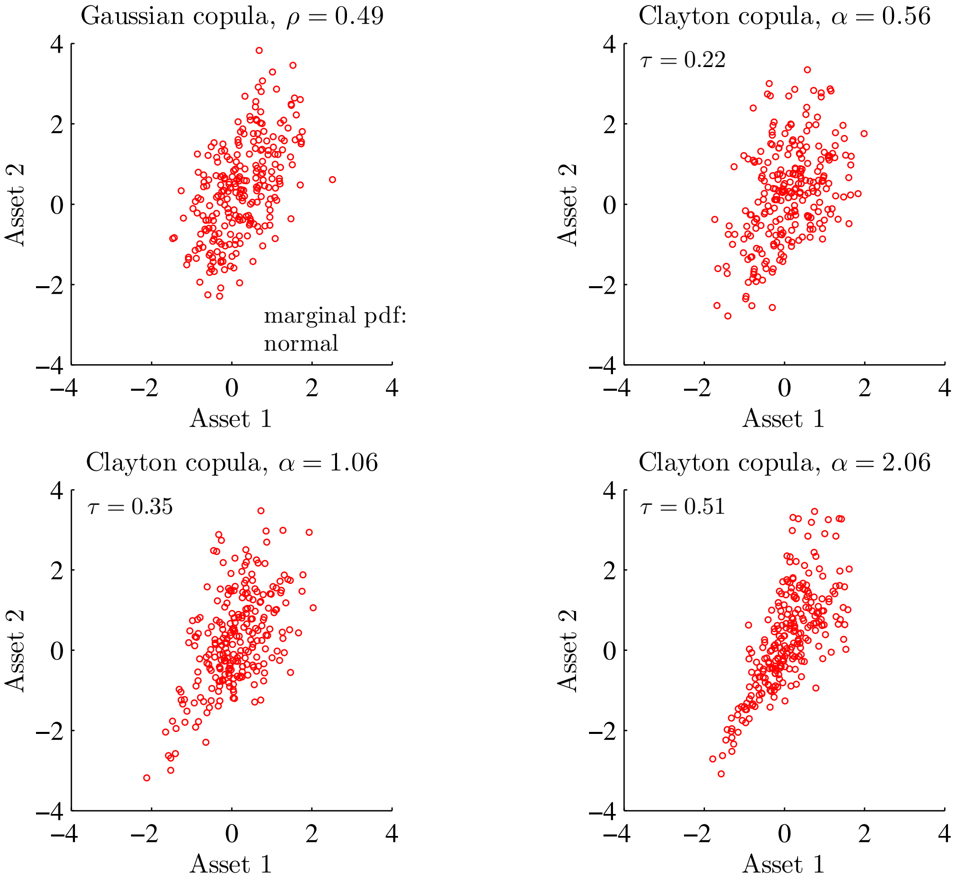


Figure 13.16: Example of scatter plots of two asset returns drawn from different copulas

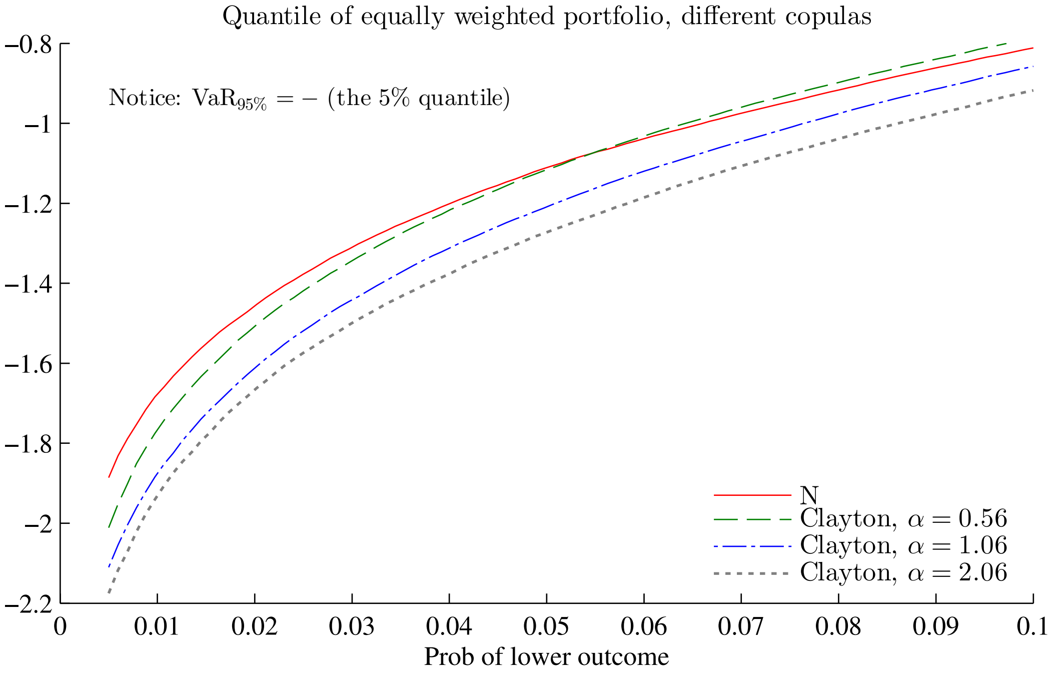


Figure 13.17: Quantiles of an equally weighted portfolio of two asset returns drawn from different copulas

14 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.

14.1 The Black-Scholes Model

14.1.1 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 $\max\left(0, S_{t+m}-K\right)$,where S_{t+m} is the asset price, and K the strike price. See Figure 14.1 for the timing convention.

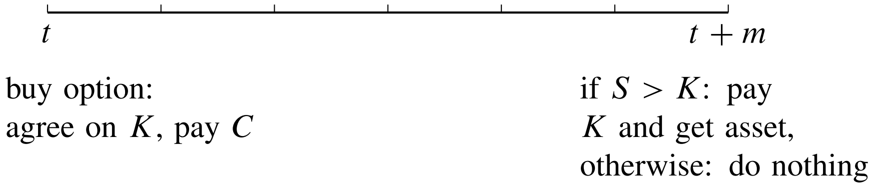


Figure 14.1: Timing convention of option contract

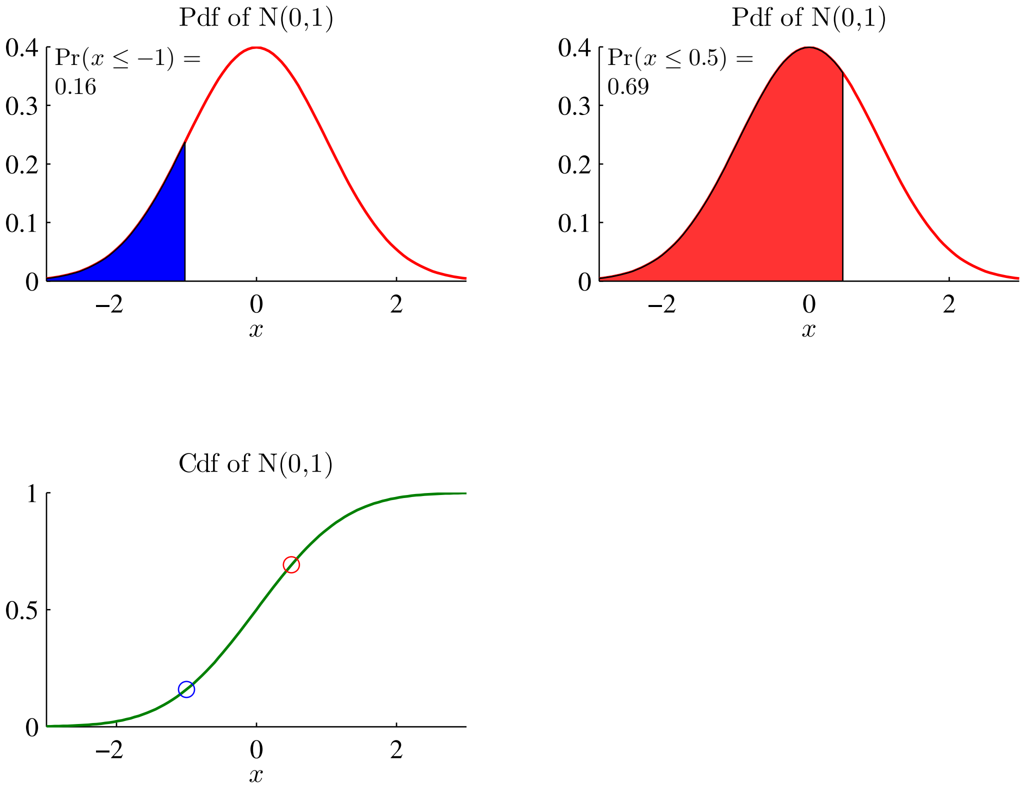


Figure 14.2: Pdf and cdf of N(0,1)

The Black-Scholes formula for a European call option price is

$$\begin{aligned} C_t &= S_t \Phi(d_1) - K e^{-rm} \Phi(d_1 - \sigma\sqrt{m}), \text{ where} \\ d_1 &= \frac{\ln(S_t/K) + \left(r + \sigma^2/2\right) m}{\sigma\sqrt{m}}. \end{aligned} \tag{14.1}$$

where $\Phi()$ 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 14.2. In this equation, S_0 is the price of the underlying asset in period t, and r is the continuously compounded interest rate (on an annualized basis).

Some basic properties of the model are illustrated in Figure 14.3. In particular, the call option price is increasing in the volöatility 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

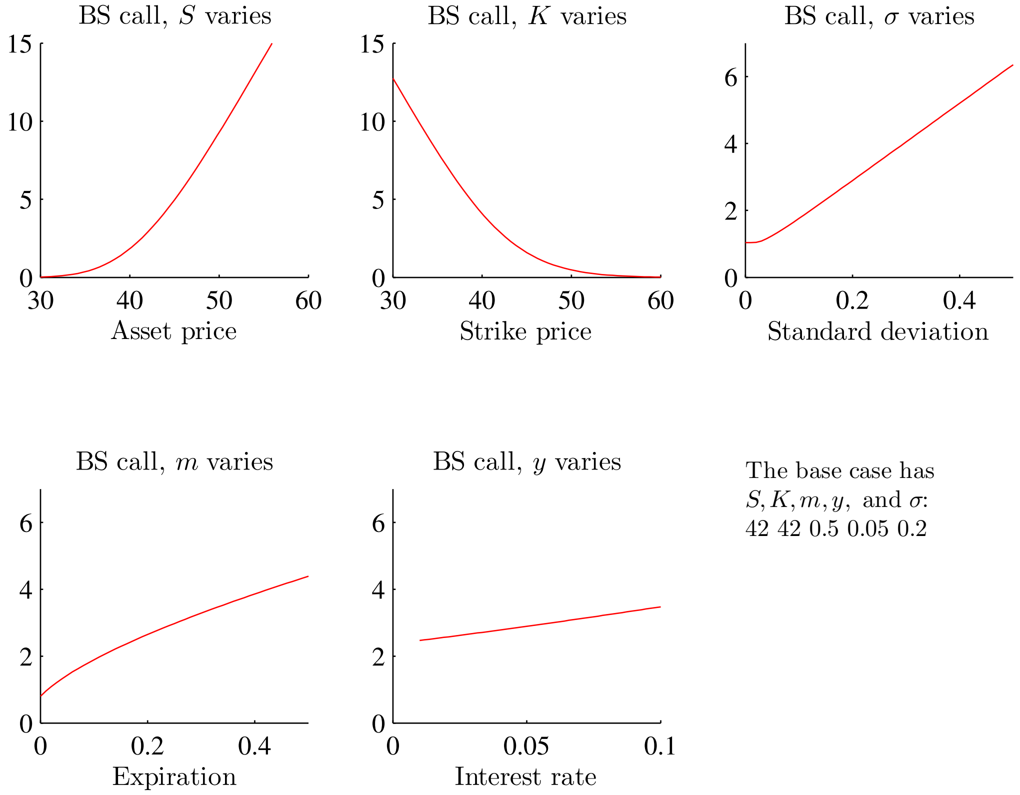


Figure 14.3: Call option price, Black-Scholes model

$t + m$ (conditional on the information in t) is normal with some mean α (not important for the option price) and the variance $m\sigma^2$

$$\ln S_{t+m} \sim N(\alpha, m\sigma^2). \tag{14.2}$$

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, a 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

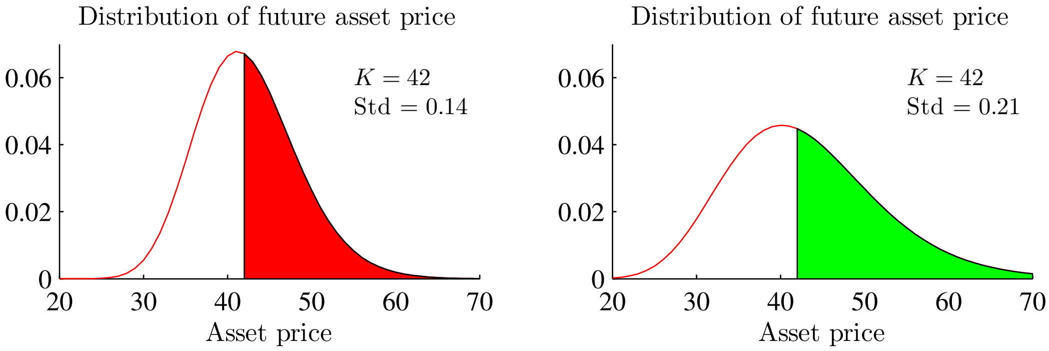


Figure 14.4: Distribution of future stock price

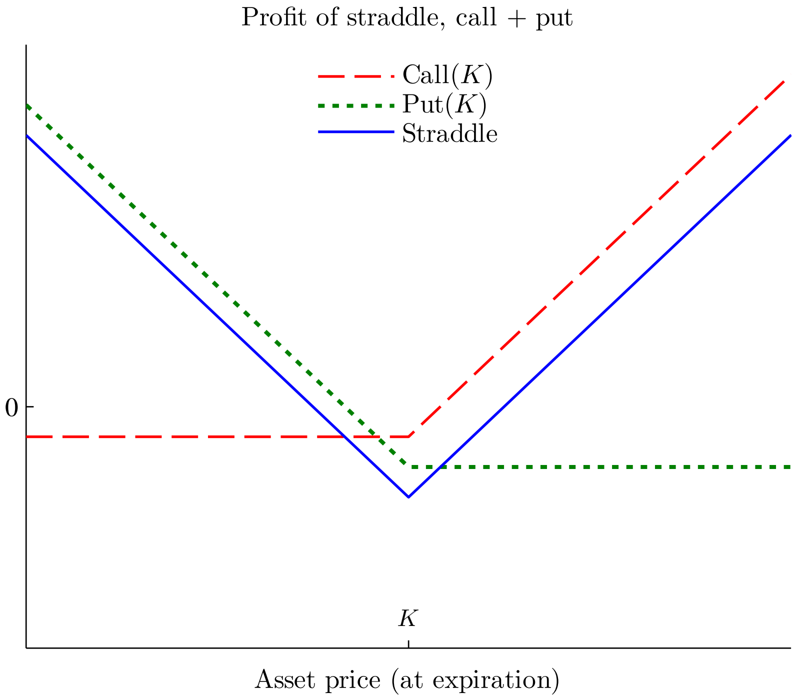


Figure 14.5: Profit of straddle portfolio

lot (in either direction) from the strike price, that is, when volatility is high. See Figures 14.4–14.5 for illustrations.

14.1.2 Implied Volatility

The pricing formula (14.1) contains only one unknown parameter: the standard deviation σ in the distribution of $\ln S_{t+m}$, see (14.2). 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. See Figure 14.3 for an illustration.

If the Black-Scholes formula is correct, that is, if the assumption in (14.2) 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. 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 relatively more probability mass in the tails ("fat tails") than a normal distribution has. See Figures 14.6–14.7 for illustrations. In addition, the implied volatilities seems to move considerably over time—see Figure 14.8 for a time series of implied volatilities

14.1.3 Brownian Motion without Mean Reversion: The Random Walk

The basic assumption behind the B-S formula (14.1) 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

$$d \ln S_t = \mu dt + \sigma dW_t, \tag{14.3}$$

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 (14.1) and (14.3) we notice that only the volatility (σ), not the drift (μ), show up in the option pricing formula. In essence, the drift is al-

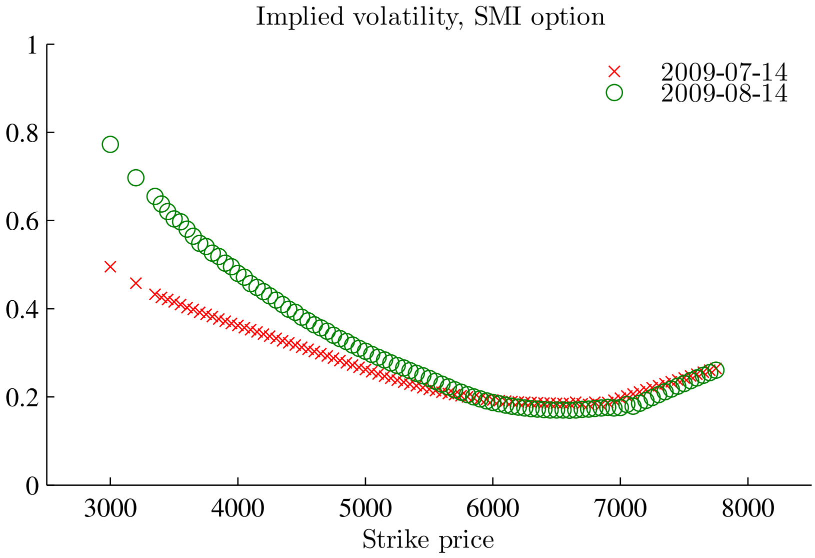


Figure 14.6: Implied volatilities of SMI options, selected dates

ready 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 14.1 *(Alternative stock price process*) If we instead of (14.3) 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 (14.3) 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.*

Remark 14.2 *((14.3) as a limit of a discrete time process*) (14.3) 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 (14.3) implies for discrete time intervals. It is straightforward to show that (14.3) implies that we have normally distributed changes (returns) and that the changes (returns) are uncorrelated (for non-overlapping data)

$$\ln(S_{t+h}/S_t) \sim N(\mu h, \sigma^2 h) \tag{14.4}$$

$$\text{Cov}[\ln(S_t/S_{t-h}), \ln(S_{t+h}/S_t)] = 0. \tag{14.5}$$

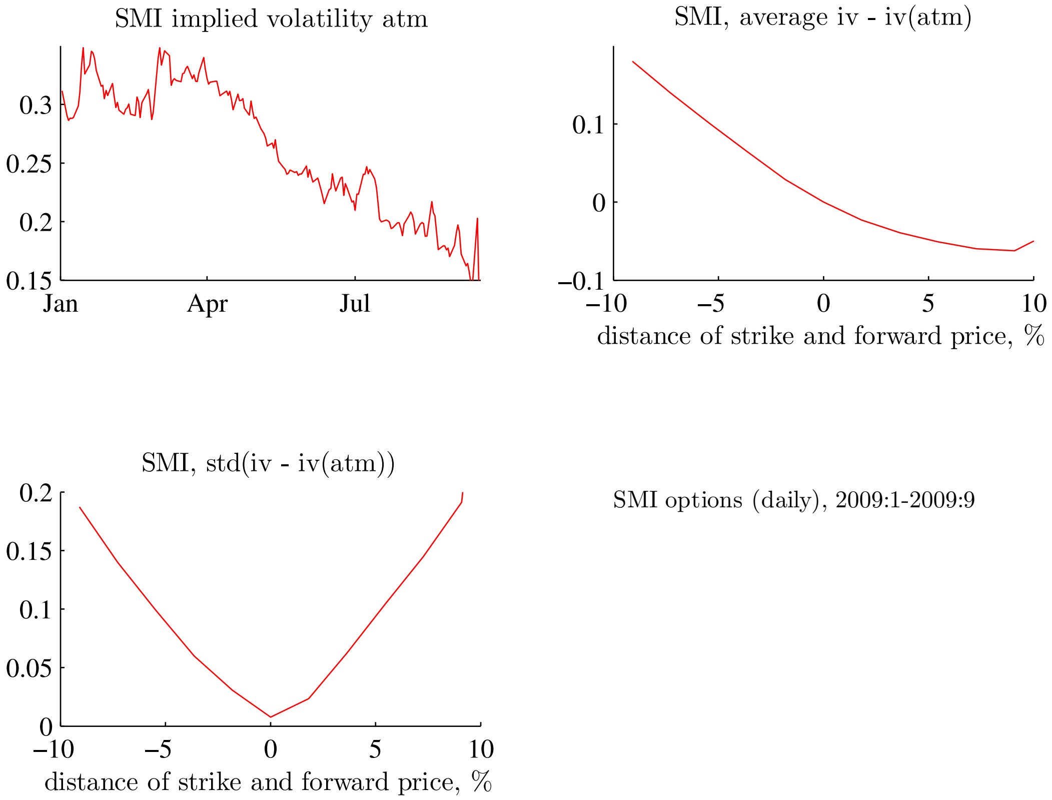


Figure 14.7: Implied volatilities

Notice that both the drift and the variance scale linearly with the horizon h. The reason is, or course, that the growth rates (even for the infinitesimal horizon) are iid.

Remark 14.3 *(iid random variable in discrete time) Suppose x_t has the constant mean μ and a variance σ^2. Then $\mathrm{E}(x_t+x_{t-1}) = 2\mu$ and $\mathrm{Var}(x_t+x_{t-1}) = 2\sigma^2+2\,\mathrm{Cov}(x_t, x_{t-1})$. If x_t is iid, then the covariance is zero, so $\mathrm{Var}(x_t + x_{t-1}) = 2\sigma^2$. In this case, both mean and variance scale linearly with the horizon.*

14.1.4 Brownian Motion with Mean Reversion*

The mean reverting Ornstein-Uhlenbeck process is

$$d\ln S_t = \lambda(\mu - \ln S_t)dt + \sigma dW_t, \text{ with } \lambda > 0. \tag{14.6}$$

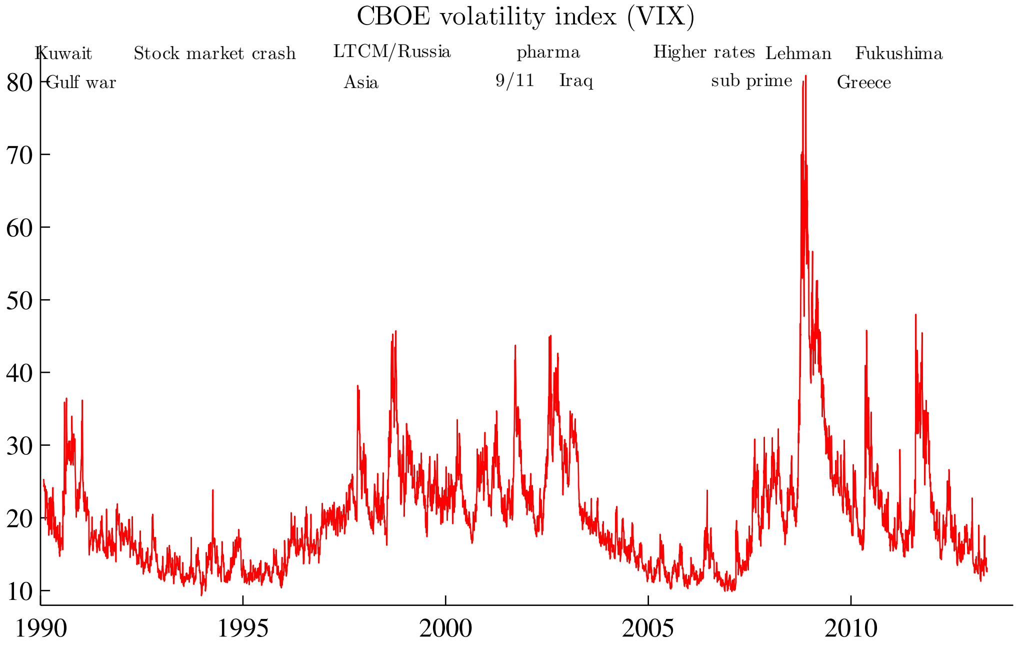


Figure 14.8: CBOE VIX, summary measure of implied volatities (30 days) on US stock markets

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 (14.6) on real life data, we (once again) have to understand what the model implies for discretely sampled data. It can be shown that it implies a discrete time AR(1)

$$\ln S_t = \alpha + \rho \ln S_{t-h} + \varepsilon_t, \text{ with} \tag{14.7}$$

$$\rho = e^{-\lambda h}, \ \alpha = \mu(1-\rho), \text{ and } \varepsilon_t \sim N\left[0, \sigma^2(1-\rho^2)/(2\lambda)\right]. \tag{14.8}$$

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 (14.8).

Example 14.4 *Suppose λ, μ and σ^2 are 2, 0, and 0.25 respectively—and the periods are years (so one unit of time corresponds to a year) . Equations (14.7)–(14.8) then gives the*

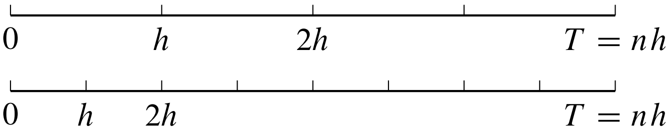


Figure 14.9: Two different samplings with same time span T

following AR(1) for weekly ($h = 1/52$) data

$$\ln S_t = 0.96 \ln S_{t-h} + \varepsilon_t \text{ with } \operatorname{Var}(\varepsilon_t) \approx 0.24.$$

14.2 Estimation of the Volatility of a Random Walk Process

This section discusses different ways of estimating the volatility. We will assume that we have data for observations in $\tau = 1, 2, .., n$. This could be 5-minute intervals, days or 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 14.9 for an illustration.

14.2.1 Standard Approach

We first estimate the variance for the sampling frequency we have, and then convert to the annual frequency.

According to (14.4) 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. The classical estimator of the variance of an iid data series is

$$\hat{s}^2 = \sum\nolimits_{\tau=1}^{n} (y_\tau - \bar{y})^2 / n, \text{ where} \tag{14.9}$$

$$\bar{y} = \sum\nolimits_{\tau=1}^{n} y_\tau / n. \tag{14.10}$$

(This is also the maximum likelihood estimator.) To annualize these numbers, use

$$\hat{\sigma}^2 = \hat{s}^2/h, \text{ and } \hat{\mu} = \bar{y}/h. \tag{14.11}$$

Example 14.5 *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 is can be quite important to subtract the mean drift, $\bar{y}$. Recall that for any random variable, we have

$$\sigma^2 = \mathrm{E}(x^2) - \mu^2, \tag{14.12}$$

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 14.6 *(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, (14.12) becomes*

$$0.16^2 = \mathrm{E}(x^2) - 0.08^2, \textit{ so } \mathrm{E}(x^2) \approx 0.18^2.$$

Assuming that the drift is zero gives an estimate of the variance equal to 0.18^2 *which is 25% too high.*

Remark 14.7 *(*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*

$$\frac{\textit{Second moment - variance}}{\textit{variance}} = \frac{m^2\mu^2}{m\sigma^2} = \frac{m\mu^2}{\sigma^2},$$

where we have used the fact that the second moment equals the variance plus the squared mean (cf (14.12)). 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 (14.9)–(14.10). 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 (14.9)–(14.11) are

$$\mathrm{Var}(\hat{\mu}) = \sigma^2/T \text{ and } \mathrm{Var}(\hat{\sigma}^2) = 2\sigma^4/n. \tag{14.13}$$

Therefore, to get a precise estimator of the mean drift, μ, a sample that stretches over a long period is crucial: 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 unimportant. For estimating the volatility (to use in the B-S model) we should therefore use high frequency data.

14.2.2 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 (14.9)–(14.10) 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

$$\hat{s}_{\tau}^{2} = \lambda \hat{s}_{\tau-1}^{2} + (1-\lambda)\left(y_{\tau-1} - \bar{y}\right)^{2}, \tag{14.14}$$

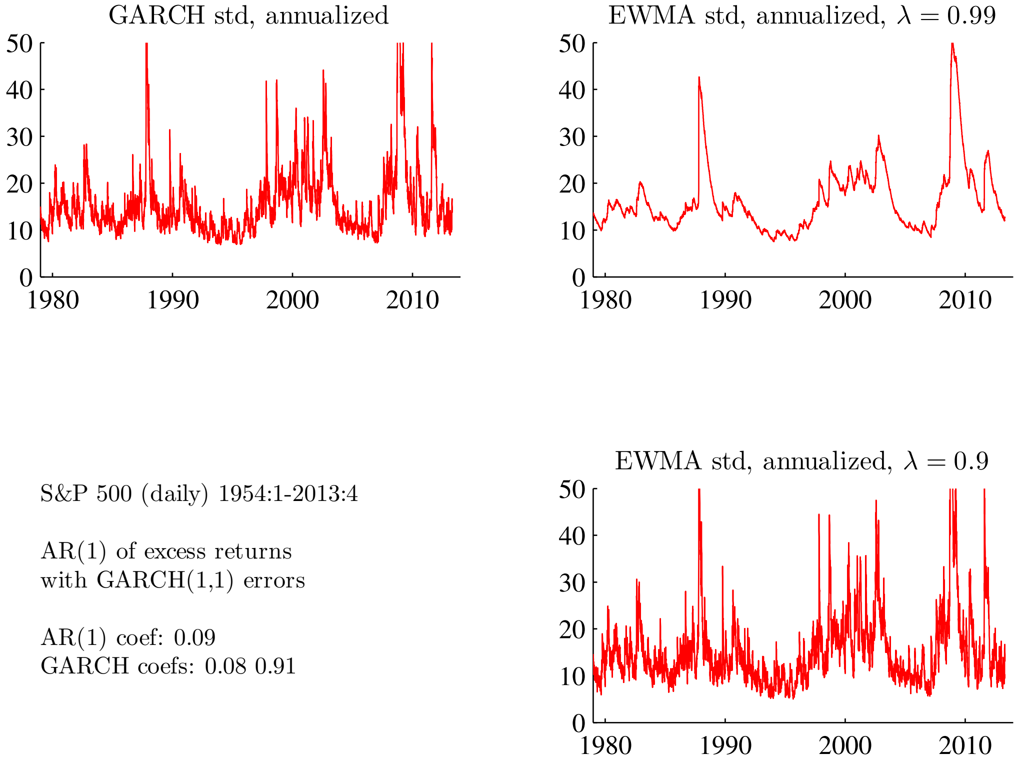
where τ is the current period and $\tau - 1$ the pervious period (say, today and yesterday). Clearly, a higher λ means that old data plays a larger role—and at the limit as λ goes towards one, we have the traditional estimator. See Figure 14.11 for a comparison using daily US equity returns. This method is commonly used by practitioners. For instance, the RISK Metrics is based on $\lambda = 0.94$ on daily data. Alternatively, λ can be chosen to minimize some criterion function.

Remark 14.8 *(EWMA with time-variation in the mean*) If we want also the mean to be time-varying, then we can use the estimator*

$$\begin{aligned}
\hat{s}_{\tau}^{2} &= (1-\lambda)\left[\left(y_{\tau-1} - \bar{y}_{\tau}\right)^{2} + \lambda\left(y_{\tau-2} - \bar{y}_{\tau}\right)^{2} + \lambda^{2}\left(y_{\tau-3} - \bar{y}_{\tau}\right)^{2} + \ldots\right] \\
\bar{y}_{\tau} &= \left[y_{\tau-1} + y_{\tau-2} + y_{\tau-3} + \ldots\right]/(\tau - 1).
\end{aligned}$$

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 does not allow for random volatility.

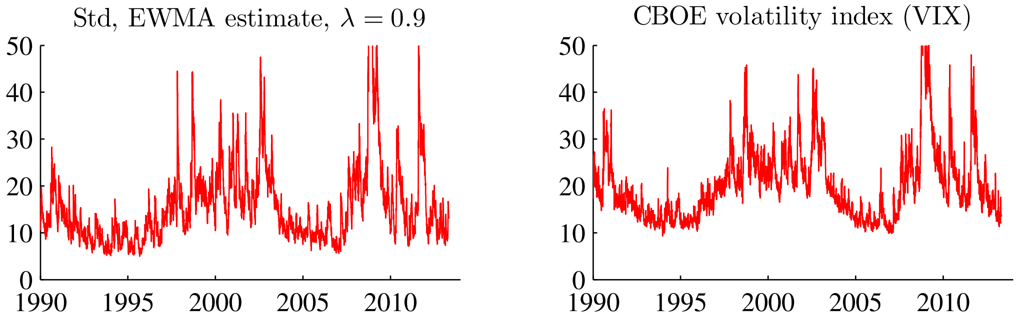


S&P 500 (daily) 1954:1-2013:4

AR(1) of excess returns
with GARCH(1,1) errors

AR(1) coef: 0.09
GARCH coefs: 0.08 0.91

Figure 14.10: Different estimates of US equity market volatility



S&P 500, daily data 1954:1-2013:4

Figure 14.11: Different estimates of US equity market volatility

14.2.3 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

$$s_\tau^2 = \alpha_0 + \alpha_1 u_{\tau-1}^2, \tag{14.15}$$

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 captures direct autoregression of the volatility

$$s_\tau^2 = \alpha_0 + \alpha_1 u_{\tau-1}^2 + \beta_1 s_{\tau-1}^2. \tag{14.16}$$

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 EMA in (14.14), 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 $\alpha_1 + \beta_1$.

14.2.4 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 14.12 for a comparison of the actual distribution of the log asset price (actually, cumulated returns, so assuming that the intial 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 futurelog asset price calculated as

$$\ln S_t + r_{t+h}, \text{ or} \tag{14.17}$$

$$\ln S_t + \textstyle\sum_{i=1}^{10} r_{t+ih}, \tag{14.18}$$

where each of the returns (r_{t+ih}) is drawn from an $N(0, s^2_{t+ih})$ distribution where the variance follows the GARCH(1,1) process like in (14.16).

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 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 14.9 *(Time-varying, but deterministic volatility*) A time-varying, but non-random volatility could be consistent with (14.2): 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.*

Bibliography

Bodie, Z., A. Kane, and A. J. Marcus, 2005, *Investments*, McGraw-Hill, Boston, 6th edn.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The econometrics of financial markets*, 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.

Gourieroux, C., and J. Jasiak, 2001, *Financial econometrics: problems, models, and methods*, Princeton University Press.

Heston, S. L., and S. Nandi, 2000, “A closed-form GARCH option valuation model,” *Review of Financial Studies*, 13, 585–625.

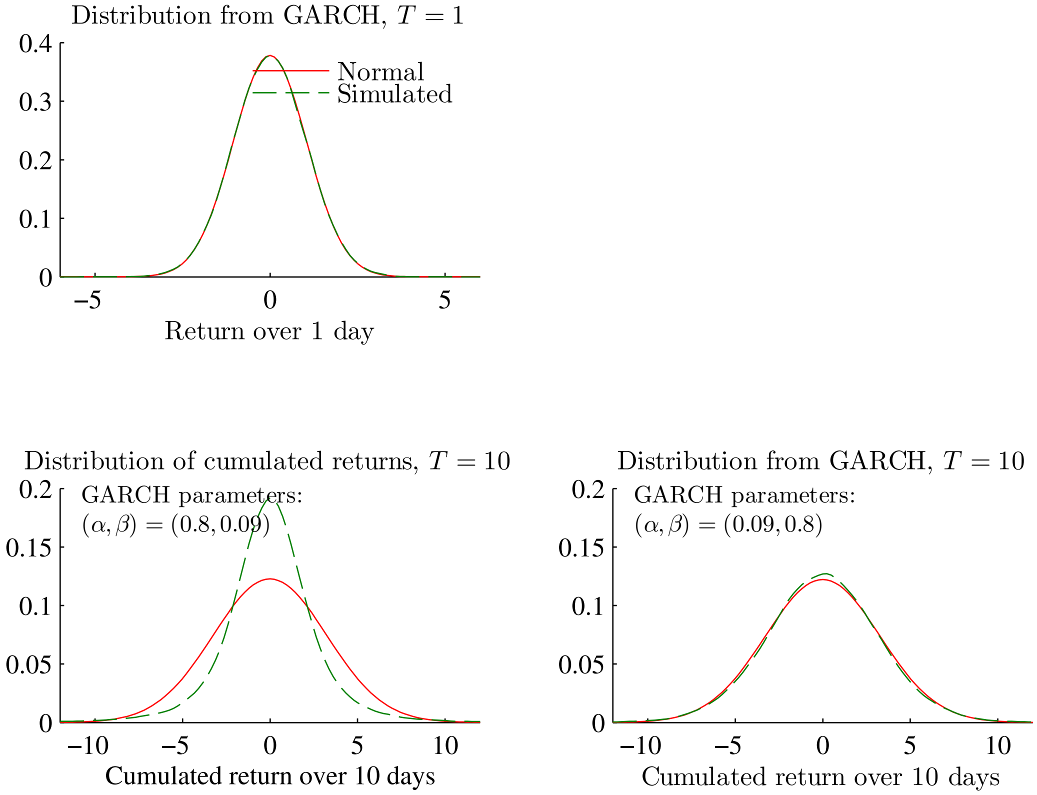


Figure 14.12: Comparison of normal and simulated distribution of *m*-period returns

Hull, J. C., 2006, *Options, futures, and other derivatives*, Prentice-Hall, Upper Saddle River, NJ, 6th edn.

Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.

15 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.

15.1 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 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 discrete variable. For instance, it could be a merger or not or if the monetary policy surprise was positive (lower interest 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).

The abnormal return of asset i in period t is

$$u_{it} = R_{it} - R_{it}^{normal}, \tag{15.1}$$

where R_{it} is the actual return and the last term is the normal return (which may differ across assets and time). The definition of the normal return is discussed in detail in Section 15.2. These returns could be nominal returns, but more likely (at least for slightly longer horizons) real returns or excess returns.

Suppose we have a sample of n such events ("assets"). To keep the notation (reason-

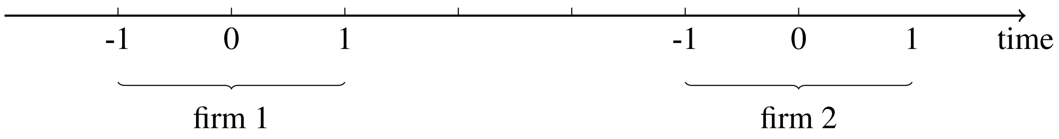


Figure 15.1: Event days and windows

ably) 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 15.1 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

$$\bar{u}_s = \textstyle\sum_{i=1}^{n} u_{is}/n. \tag{15.2}$$

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 (15.1) over some period around the event. It is often calculated from the beginning of the event window. For instance, if the event window starts at $-w$, then the q-period (day?) car for firm i is

$$\text{car}_{iq} = u_{i,-w} + u_{i,-w+1} + \ldots + u_{i,-w+q-1}. \tag{15.3}$$

The cross sectional average of the q-period car is

$$\overline{\text{car}}_q = \textstyle\sum_{i=1}^{n} \text{car}_{iq}/n. \tag{15.4}$$

See Figure 15.2 for an empirical example.

Example 15.1 *(Abnormal returns for ± day around event, two firms) Suppose there are two firms and the event window contains* ± 1 *day around the event day, and that the*

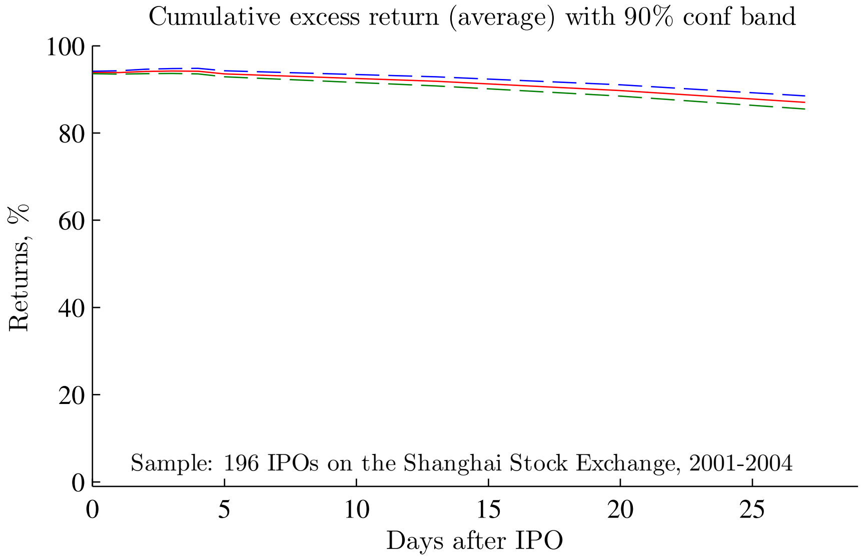


Figure 15.2: Event study of IPOs in Shanghai 2001–2004. (Data from Nou Lai.)

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

15.2 Models of Normal Returns

This section summarizes the most common ways of calculating the normal return in (15.1). The parameters in these models are typically estimated on a recent sample, the "estimation window," that ends before the event window. See Figure 15.3 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 in 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 around some mean μ_i

$$\begin{aligned} R_{it} &= \mu_i + \varepsilon_{it} \text{ with} \\ \mathrm{E}(\varepsilon_{it}) &= \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{i,t-s}) = 0. \end{aligned} \tag{15.5}$$

This mean is estimated by the sample average (during the estimation window). The normal return in (15.1) is then the estimated mean. $\hat{\mu}_i$ so the abnormal return (in the estimation window) becomes $\hat{\varepsilon}_{it}$. During the event window, we calculate the abnormal return as

$$u_{it} = R_{it} - \hat{\mu}_i. \tag{15.6}$$

The standard error of this is estimated by the standard error of $\hat{\varepsilon}_{it}$ (in the estimation window).

The *market model* is a linear regression of the return of asset i on the market return

$$\begin{aligned} R_{it} &= \alpha_i + \beta_i R_{mt} + \varepsilon_{it} \text{ with} \\ \mathrm{E}(\varepsilon_{it}) &= \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{i,t-s}) = \mathrm{Cov}(\varepsilon_{it}, R_{mt}) = 0. \end{aligned} \tag{15.7}$$

Notice that we typically do not impose the CAPM restrictions on the intercept in (15.7). The normal return in (15.1) is then calculated by combining the regression coefficients with the actual market return as $\hat{\alpha}_i + \hat{\beta}_i R_{mt}$, so the the abnormal return in the estimation window is $\hat{\varepsilon}_{it}$. For the event window we calculate the abnormal return as

$$u_{it} = R_{it} - \hat{\alpha}_i - \hat{\beta}_i R_{mt}. \tag{15.8}$$

The standard error of this is estimated by the standard error of $\hat{\varepsilon}_{it}$ (in the estimation window).

When we restrict $\alpha_i = 0$ and $\beta_i = 1$, then this approach is called the *market-adjusted-*

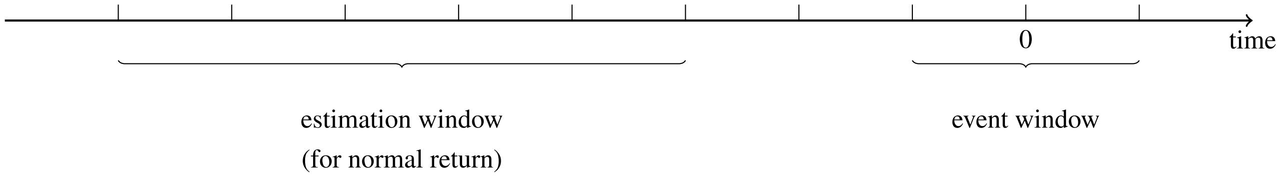


Figure 15.3: Event and estimation windows

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

$$u_{it} = R_{it} - R_{mt} \tag{15.9}$$

and the standard error of it is estimated by $\text{Std}(R_{it} - R_{mt})$ 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 (15.7) 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.

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 actual return for other small manufacturing firms (without events). In this case, the abnormal return becomes the difference between the actual return and the return on the matching portfolio. This type of *matching portfolio* is becoming increasingly popular. For the event window we calculate the abnormal return as

$$u_{it} = R_{it} - R_{pt}, \tag{15.10}$$

where R_{pt} is the return of the matching portfolio. The standard error of it is estimated by $\text{Std}(R_{it} - R_{pt})$ in the estimation window.

All the methods discussed here try to take into account the risk premium on the asset. It is captured by the mean in the constant mean mode, the beta in the market model, and by the way the matching portfolio is constructed. However, sometimes there is no data in the estimation window. The typical approach is then to use the actual market return as the normal return—that is, to use (15.7) but assuming that $\alpha_i = 0$ and $\beta_i = 1$. Clearly, this does not account for the risk premium on asset i, and is therefore a fairly rough guide.

Apart from accounting for the risk premium, does the choice of the model of the normal return matter a lot? 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 for the market model which the test more precise (see Section 15.3 for a discussion of how the variance of the abnormal return affects the variance of the test statistic).

To illustrate the importance of the model for normal returns, consider the market model (15.7). Under the null hypothesis that the event has no effect on the return, the abnormal return would be just the residual in the regression (15.7). It has the variance (assuming we know the model parameters)

$$\operatorname{Var}(u_{it}) = \operatorname{Var}(\varepsilon_{it}) = (1 - R^2)\operatorname{Var}(R_{it}), \tag{15.11}$$

where R^2 is the coefficient of determination of the regression (15.7).

Proof. (of (15.11)) From (15.7) we have (dropping the time subscripts)

$$\operatorname{Var}(R_i) = \beta_i^2 \operatorname{Var}(R_m) + \operatorname{Var}(\varepsilon_i).$$

We therefore get

$$\begin{aligned}
\operatorname{Var}(\varepsilon_i) &= \operatorname{Var}(R_i) - \beta_i^2 \operatorname{Var}(R_m)\\
&= \operatorname{Var}(R_i) - \operatorname{Cov}(R_i, R_m)^2 / \operatorname{Var}(R_m)\\
&= \operatorname{Var}(R_i) - \operatorname{Corr}(R_i, R_m)^2 \operatorname{Var}(R_i)\\
&= (1 - R^2)\operatorname{Var}(R_i).
\end{aligned}$$

The second equality follows from the fact that $\beta_i = \operatorname{Cov}(R_i, R_m)/\operatorname{Var}(R_m)$, the third equality from multiplying and dividing the last term by $\operatorname{Var}(R_i)$ and using the definition of the correlation, and the fourth equality from the fact that the coefficient of determination in a simple regression equals the squared correlation of the dependent variable and the regressor. ■

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 15.3). 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. More realistically, R^2 might be 0.43 (or less), so the market model gives a 25% decrease in the standard deviation, which is not a whole lot. 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.

15.3 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 are uncertain. Second, even if we knew the normal return for sure, the actual returns are random variables—and they will always deviate from their population mean 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. It can also be argued that the model for the normal return (for instance, a market model) should capture all common movements by the regressors — leaving the abnormal returns (the residuals) uncorrelated across firms. In contrast, if the events happen at the same time, the cross-correlation must be handled somehow. This is, for instance, the case if the events are macroeconomic announcements or monetary policy moves. An easy way to handle such synchronized (clustered) events is to form portfolios of those assets that share the event time—and then only use portfolios with non-overlapping events in the cross-sectional study. 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 (15.2), is then

$$\text{Var}(\bar{u}_s) = \left(\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2\right)/n^2 = \textstyle\sum_{i=1}^{n} \sigma_i^2/n^2, \tag{15.12}$$

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 since

$$\bar{u}_s / \,\text{Std}(\bar{u}_s) \to^d N(0,1). \tag{15.13}$$

The *cumulative abnormal return* over q period, $\text{car}_{i,q}$, can also be tested with a t-test. Since the returns are assumed to have no autocorrelation the variance of the $\text{car}_{i,q}$

$$\text{Var}(\text{car}_{iq}) = q\sigma_i^2. \tag{15.14}$$

This variance is increasing in q since we are considering cumulative returns (not the time average of returns).

The *cross-sectional average car$_{i,q}$* is then (similarly to (15.12))

$$\text{Var}(\overline{\text{car}}_q) = \left(q\sigma_1^2 + q\sigma_2^2 + ... + q\sigma_n^2\right)/n^2 = q\textstyle\sum_{i=1}^{n} \sigma_i^2/n^2, \tag{15.15}$$

if the abnormal returns are uncorrelated across time and assets.

Figures 4.2a–b in Campbell, Lo, and MacKinlay (1997) provide a nice example of an event study (based on the effect of earnings announcements).

Example 15.2 *(Variances of abnormal returns) If the standard deviations of the daily abnormal returns of the two firms in Example 15.1 are $\sigma_1 = 0.1$ and 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	$\left(0.1^2 + 0.2^2\right)/4$
0	0.1^2	0.2^2	$\left(0.1^2 + 0.2^2\right)/4$
1	0.1^2	0.2^2	$\left(0.1^2 + 0.2^2\right)/4$

Similarly, the variances for the cumulative abnormal returns are

Time	*Firm 1*	*Firm 2*	*Cross-sectional Average*
-1	0.1^2	0.2^2	$\left(0.1^2+0.2^2\right)/4$
0	2×0.1^2	2×0.2^2	$2\times\left(0.1^2+0.2^2\right)/4$
1	3×0.1^2	3×0.2^2	$3\times\left(0.1^2+0.2^2\right)/4$

Example 15.3 *(Tests of abnormal returns) By dividing the numbers in Example 15.1 by the square root of the numbers in Example 15.2 (that is, the standard deviations) we get the test statistics for the abnormal returns*

Time	*Firm 1*	*Firm 2*	*Cross-sectional Average*
-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

Time	*Firm 1*	*Firm 2*	*Cross-sectional Average*
-1	2	-0.5	0.4
0	8.5	6.7	9.8
1	7.5	6.4	9.0

15.4 Quantitative Events

Some events are not easily classified as discrete 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)

$$\text{car}_{iq} = a + bx_i + \zeta_i. \tag{15.16}$$

The slope coefficient is then a measure of how much the cumulative abnormal return reacts to a change of one unit of x_i.

Bibliography

Bodie, Z., A. Kane, and A. J. Marcus, 2005, *Investments*, McGraw-Hill, Boston, 6th edn.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The econometrics of financial markets*, 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.

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.

16 Kernel Density Estimation and Regression

16.1 Non-Parametric Regression

Reference: Campbell, Lo, and MacKinlay (1997) 12.3; Härdle (1990); Pagan and Ullah (1999); Mittelhammer, Judge, and Miller (2000) 21

16.1.1 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

$$y_t = b(x_t) + \varepsilon_t,\ \varepsilon_t \text{ is iid and } \mathrm{E}\,\varepsilon_t = \mathrm{Cov}\,[b(x_t), \varepsilon_t] = 0, \tag{16.1}$$

where $b()$ is an unknown, possibly non-linear, function.

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_s is far from x_t.

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

$$\begin{aligned}\hat{b}(x) &= \frac{w_1(x)y_1 + w_2(x)y_2 + \ldots + w_T(x)y_T}{w_1(x) + w_2(x) + \ldots + w_T(x)}\\ &= \frac{\sum_{t=1}^{T} w_t(x)y_t}{\sum_{t=1}^{T} w_t(x)}, \qquad (16.2)\end{aligned}$$

where $w_t(x)/\Sigma_{t=1}^{T} w_t(x)$ is the weight given to observation t. The function $w_t(x)$ is positive and (weakly) increasing in the distance between x_t and x. Note that denominator makes the weights sum to unity. The basic assumption behind (16.2) is that the $b(x)$ function is smooth so local (around x) averaging makes sense.

As an example of a $w(.)$ 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, $\mathrm{E}[y_t - \hat{b}(x)]^2$, and the expected squared acceleration, $\mathrm{E}[d^2\hat{b}(x)/dx^2]^2$. 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 $w(.)$.

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

$$w_t(x) = \frac{1}{h}\delta\left(\left|\frac{x_t - x}{h}\right| \leq 1/2\right), \text{where } \delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases} \qquad (16.3)$$

This weighting function puts the weight $1/h$ on all data point in the interval $x \pm h/2$ and zero on all other data points.

However, we can gain efficiency and get a smoother (across x values) estimate by using another density function that 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. 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. 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 of in a bell-shaped way.

See Figure 16.1 for an illustration.

A low value of h means that the weights taper off fast—the weight function is then a normal pdf with a low variance. With this particular kernel, we get the following weights t a point x

$$w_t(x) = \frac{\exp\left[-\left(\frac{x_t - x}{h}\right)^2/2\right]}{h\sqrt{2\pi}}. \tag{16.4}$$

When $h \to 0$, then $\hat{b}(x)$ evaluated at $x = x_t$ becomes just y_t, so no averaging is done. In contrast, as $h \to \infty$, $\hat{b}(x)$ becomes the sample average of y_t, so we have global averaging. Clearly, some value of h in between is needed.

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 16.2 for an illustration. Special corrections might be needed if there are a lot of observations stacked close to the boundary of the support of x (see Härdle (1990) 4.4).

See Figures 16.3–16.4 for an example. Note that the volatility is defined as the square of the drift minus expected drift (from the same estimation method).

A rule of thumb value of h is

$$h = T^{-1/5}|\gamma|^{-2/5}\sigma_\varepsilon^{2/5}(x_{\max} - x_{\min})^{1/5} \times 0.6, \tag{16.5}$$

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, 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

$$\text{EPE}(h) = \sum\nolimits_{t=1}^{T}\left[y_t - \hat{b}_{-t}(x_t, h)\right]^2/T, \tag{16.6}$$

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 (16.6) 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 16.5 for an example.

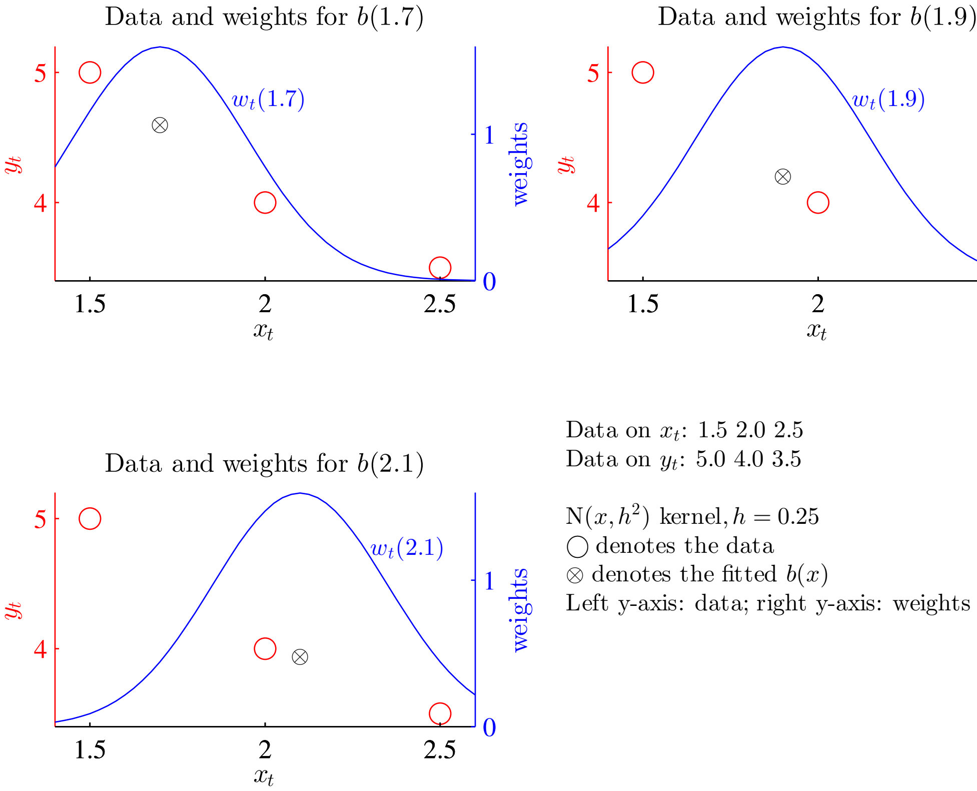


Figure 16.1: Example of kernel regression with three data points

Remark 16.1 *(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 $t = 3, 4, ..., T$. Calculate the EPE as in (16.6).
Step 4: redo Steps 2–3, 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

$$\sqrt{Th}\left[\hat{b}(x) - \mathrm{E}\,\hat{b}(x)\right] \to^{d} N\left[0, \frac{1}{2\sqrt{\pi}} \frac{\sigma^2(x)}{f(x)}\right], \tag{16.7}$$

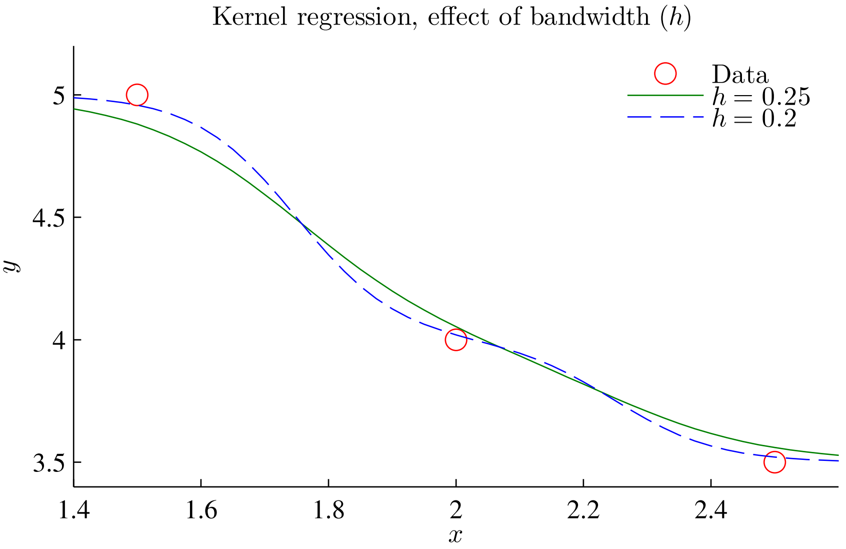


Figure 16.2: Example of kernel regression with three data points

where $\sigma^2(x)$ is the variance of the residuals in (16.1) and $f(x)$ the marginal density of x. Clearly, this means that we have (with sloppy notation)

$$\hat{b}(x) \text{ ``} \rightarrow^d \text{''} N\left[\mathrm{E}\,\hat{b}(x), \frac{1}{2\sqrt{\pi}}\frac{\sigma^2(x)}{f(x)}\frac{1}{Th}\right], \tag{16.8}$$

To estimate the density function needed in (16.7), we can use a kernel density estimator of the pdf at some point x

$$\hat{f}(x) = \frac{1}{Th_x}\sum\nolimits_{t=1}^{T} K\left(\frac{x_t - x}{h_x}\right), \text{ where} \tag{16.9}$$
$$K(u) = \frac{\exp\left(-u^2/2\right)}{\sqrt{2\pi}}.$$

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. (Clearly, using $K\left[(x_t - x)/h_x\right]/h_x$ is the same as using pdf of $N(x, h_x^2)$.) Notice that h_x need not be the same as the bandwidth (h) used in the kernel regression.

See *Figure 16.6* for an example where the width of the confidence band varies across x values—mostly because the sample contains few observations close to some x values.

(However, the assumption of independent observations can be questioned in this case.)

To estimate the function $\sigma^2(x)$ in (16.7), we use a non-parametric regression of the squared fitted residuals on x_t

$$\hat{\varepsilon}_t^2 = \sigma^2(x_t), \text{ where } \hat{\varepsilon}_t = y_t - \hat{b}(x_t), \tag{16.10}$$

where $\hat{b}(x_t)$ are the fitted values from the non-parametric regression (16.1). Notice that this approach allows the variance to depend on the x value.

Example 16.2 *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 (16.4) is*

$$\begin{aligned}\sum\nolimits_{t=1}^{T} w_t(x) y_t &= \left(e^{-(1.5-1.9)^2/2} \times 5 + e^{-(2-1.9)^2/2} \times 4 + e^{-(2.5-1.9)^2/2} \times 3.5\right)/\sqrt{2\pi}\\ &\approx (0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5)/\sqrt{2\pi}\\ &= 11.52/\sqrt{2\pi}.\end{aligned}$$

The denominator is

$$\begin{aligned}\sum\nolimits_{t=1}^{T} w_t(x) &= \left(e^{-(1.5-1.9)^2/2} + e^{-(2-1.9)^2/2} + e^{-(2.5-1.9)^2/2}\right)/\sqrt{2\pi}\\ &\approx 2.75/\sqrt{2\pi}.\end{aligned}$$

The estimate at $x = 1.9$ *is therefore*

$$\hat{b}(1.9) \approx 11.52/2.75 \approx 4.19.$$

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.

16.1.2 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \ \varepsilon_t \text{ is iid and } \mathrm{E}\,\varepsilon_t = 0. \tag{16.11}$$

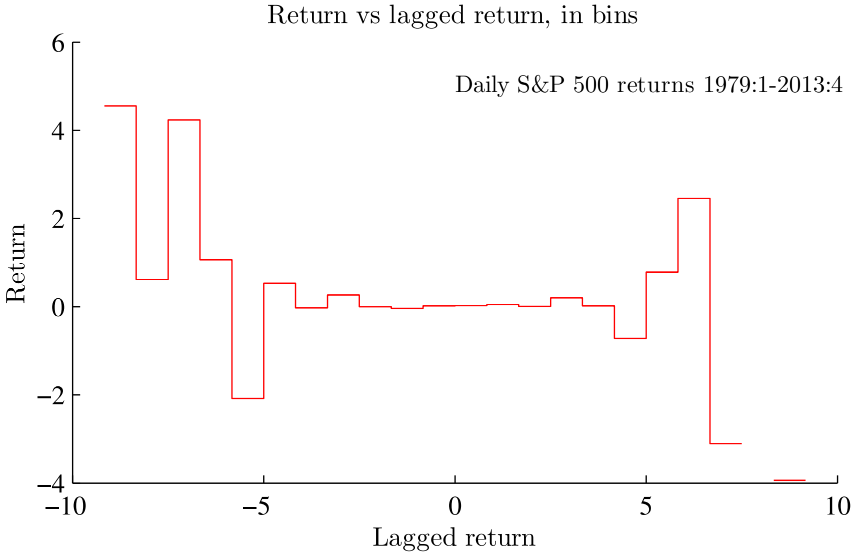


Figure 16.3: Crude non-parametric regression

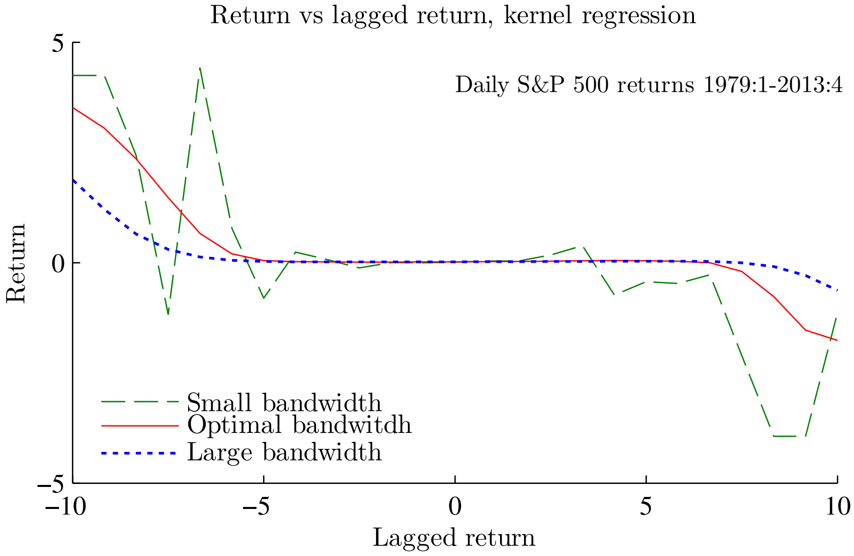


Figure 16.4: Non-parametric regression, importance of bandwidth

This makes the estimation problem much harder since there are typically few observations in every bivariate bin (rectangle) of x and z. For instance, 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

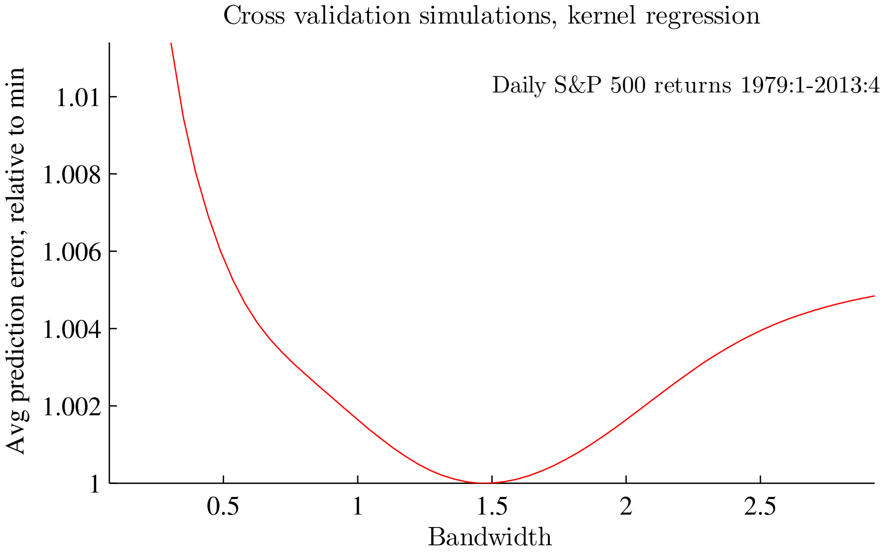


Figure 16.5: Cross-validation

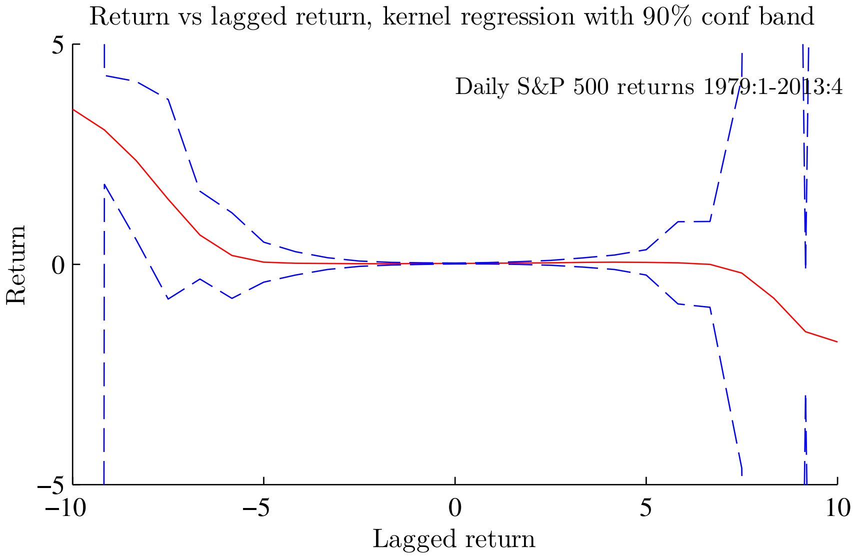


Figure 16.6: Non-parametric regression with confidence bands

of observations in every bin.

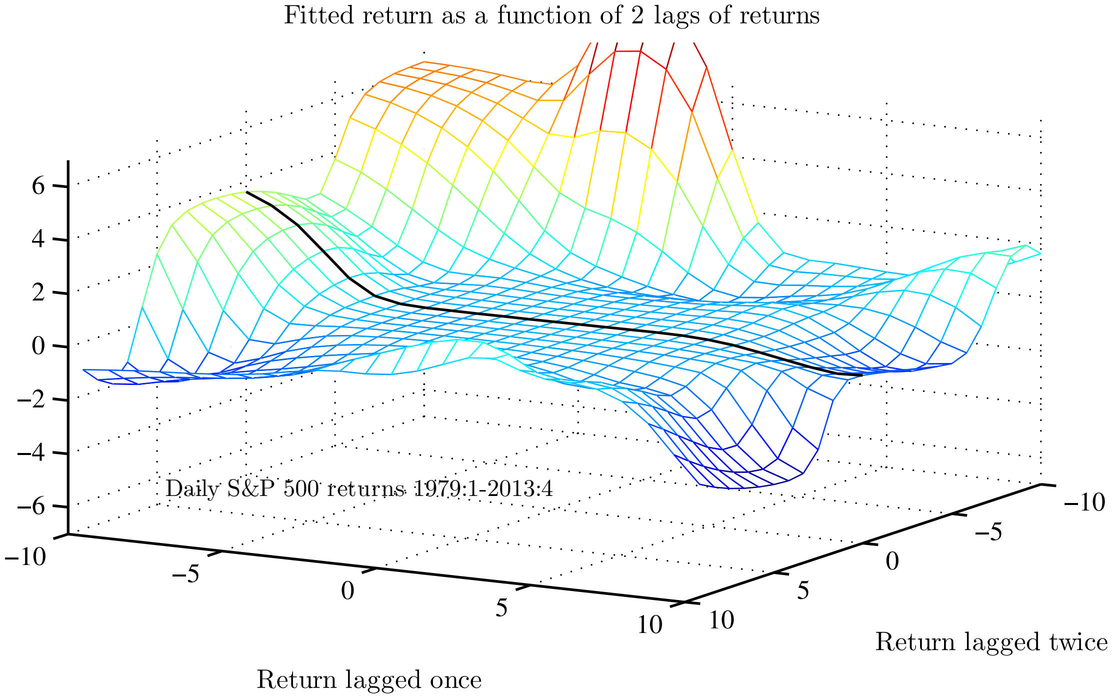


Figure 16.7: Non-parametric regression with two regressors

In any case, the most common way to implement the kernel regressor is to let

$$\hat{b}(x,z) = \frac{\sum_{t=1}^{T} w_t(x) w_t(z) y_t}{\sum_{t=1}^{T} w_t(x) w_t(z)}, \tag{16.12}$$

where $w_t(x)$ and $w_t(z)$ are two kernels like in (16.4) 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 $w_t(x) w_t(z)$, which is high if both x_t and z_t are close to x and z respectively.

See Figure 16.7 for an example.

16.2 Examples of Non-Parametric Estimation

16.2.1 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 model assumes that the

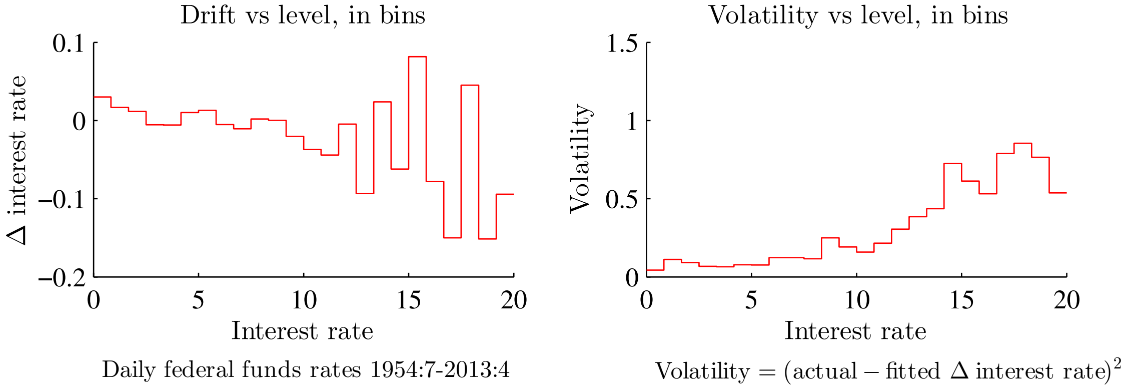


Figure 16.8: Crude non-parametric estimation

(de-meaned) short interest rate, r_t, is a mean-reverting AR(1) process

$$r_t = \rho r_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim N(0, \sigma^2), \text{ so} \tag{16.13}$$

$$r_t - r_{t-1} = (\rho - 1) r_{t-1} + \varepsilon_t, \tag{16.14}$$

and that all term premia are constant. This means that the drift is decreasing in the interest rate, but that the volatility is constant. For instance, if $\rho = 0.95$ (a very peristent interest rate), then (16.14) is

$$r_t - r_{t-1} = -0.05 r_{t-1} + \varepsilon_t, \tag{16.15}$$

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 (16.13).) It can then be shown that all interest rates (for different maturities) are linear functions of short interest rates.

To capture more movements in the yield curve, models with richer dynamics are used. For instance, Cox, Ingersoll, and Ross (1985) construct a model which implies that the short interest rate follows an AR(1) as in (16.13) 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)). Figures 16.8–16.11 give an example. Note that the volatility is defined as the square of the drift minus expected drift (from the same estimation method).

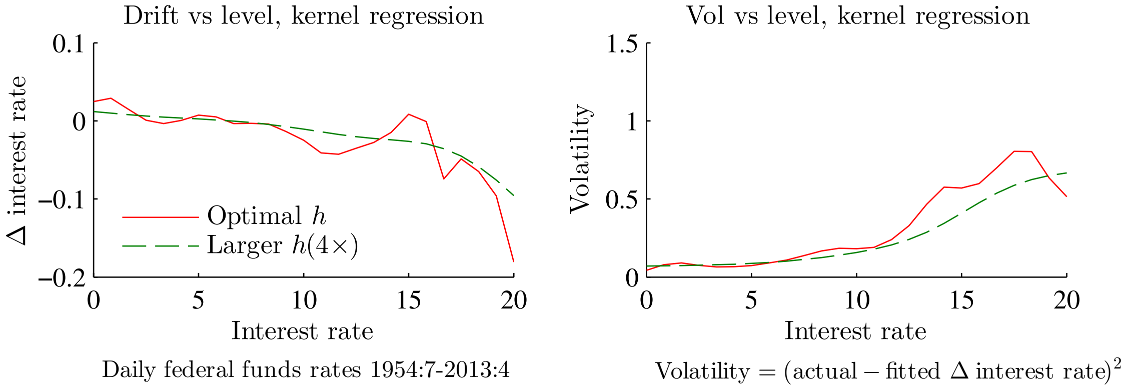


Figure 16.9: Kernel regression, importance of bandwidth

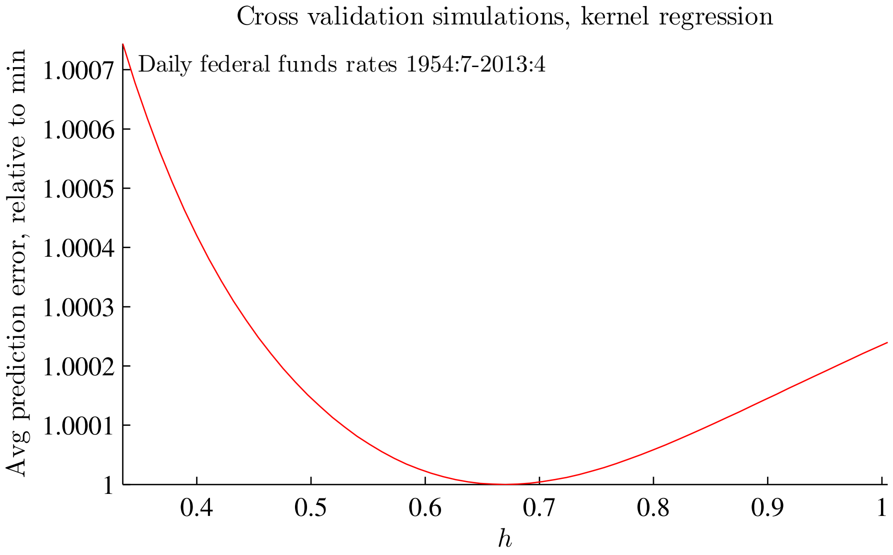
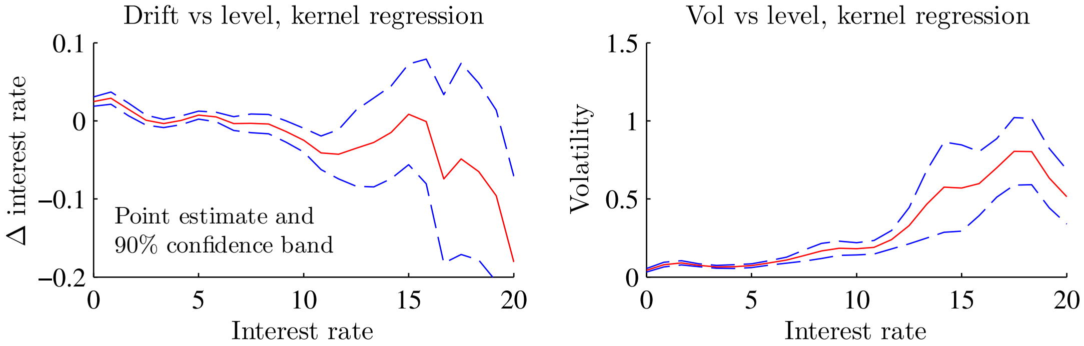


Figure 16.10: Cross-validation

16.2.2 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 than what is compatible with the normal distribution ("fat tails").


Drift vs level, kernel regression

Vol vs level, kernel regression

Volatility = (actual − fitted ∆ interest rate)^2

Daily federal funds rates 1954:7-2013:4
The bandwith is from cross-validation

Figure 16.11: Kernel regression, confidence band

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 asset, strike price, time to expiry, interest rate, and dividends). For instance, Ait-Sahalia and Lo (1998) applies 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. Whereas the distributions for short horizons are not too different from normal distributions, this is not true for longer horizons.

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.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The Econometrics of Financial Markets*, Princeton University Press, Princeton, New Jersey.

Cox, J. C., J. E. Ingersoll, and S. A. Ross, 1985, "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–407.

Härdle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric Foundations*, Cambridge University Press, Cambridge.

Pagan, A., and A. Ullah, 1999, *Nonparametric Econometrics*, Cambridge University Press.

17 Simulating the Finite Sample Properties*

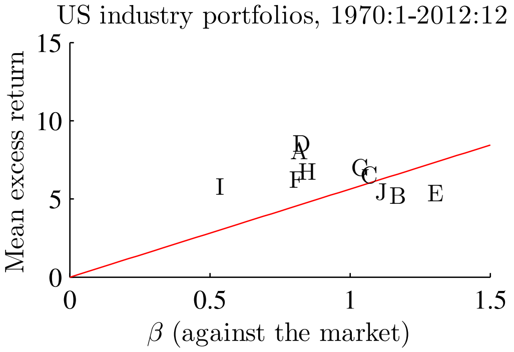
Reference: Greene (2000) 5.3 and Horowitz (2001)

Additional references: Cochrane (2001) 15.2; Davidson and MacKinnon (1993) 21; Davison and Hinkley (1997); Efron and Tibshirani (1993) (bootstrapping, chap 9 in particular); and Berkowitz and Kilian (2000) (bootstrapping in time series models)

We know the small sample properties of regression coefficients in linear models with fixed regressors and iid normal error terms. Monte Carlo simulations and bootstrapping are two common techniques used to understand the small sample properties when these conditions are not satisfied.

How they 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 is not needed, see Figure 17.1.



	alpha	t LS	t NW	t boot
all	NaN	NaN	NaN	NaN
A (NoDur)	3.62	2.72	2.71	2.70
B (Durbl)	-1.21	-0.58	-0.59	-0.59
C (Manuf)	0.70	0.72	0.71	0.69
D (Enrgy)	4.06	1.80	1.81	1.81
E (HiTec)	-1.82	-1.00	-1.00	-0.98
F (Telcm)	1.82	1.07	1.06	1.05
G (Shops)	1.37	0.94	0.94	0.94
H (Hlth)	2.13	1.22	1.24	1.24
I (Utils)	2.87	1.61	1.58	1.56
J (Other)	-0.65	-0.61	-0.60	-0.61

NW uses 1 lag
The bootstrap samples pairs of (y_t, x_t)
3000 simulations

Figure 17.1: CAPM, US industry portfolios, different t-stats

17.1 Monte Carlo Simulations

17.1.1 Monte Carlo Simulations in the Simplest Case

Monte Carlo simulations is essentially a way to generate many artificial (small) samples from a parameterized model and then estimating the statistic on each of those samples. The distribution of the statistic is then used as the small sample distribution of the estimator.

The following is an example of how Monte Carlo simulations could be done in the special case of a linear model with a scalar dependent variable

$$y_t = x_t'\beta + u_t, \tag{17.1}$$

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 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 asymptotic 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 $t = 1, \ldots, T$. Draw random numbers $\tilde{u}_t$ for $t = 1, \ldots, T$ and use those together with the artificial

sample of $\tilde{x}_t$ to calculate an artificial sample $\tilde{y}_t$ for $t = 1, \ldots, T$ from

$$\tilde{y}_t = \tilde{x}_t'\beta + \tilde{u}_t, \tag{17.2}$$

by using the prespecified values of the coefficients β.

2. Calculate an estimate $\hat{\beta}$ and record it along with the value of $g(\hat{\beta})$ 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 $\hat{\beta}$, $g(\hat{\beta})$, 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 $\hat{\beta}$, $g(\hat{\beta})$, 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?

See Figures 17.2–17.3 for an example.

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 17.1 *(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 to calculate the lower triangular P such that $\Sigma = PP'$ (note that Gauss and MatLab returns P' instead of P). Draw u_t from an $N(0, I)$ distribution (randn in MatLab, rndn in Gauss), and define $\varepsilon_t = \mu + Pu_t$. Note that* $\text{Cov}(\varepsilon_t) = \text{E}\, Pu_tu_t'P' = PIP' = \Sigma$.

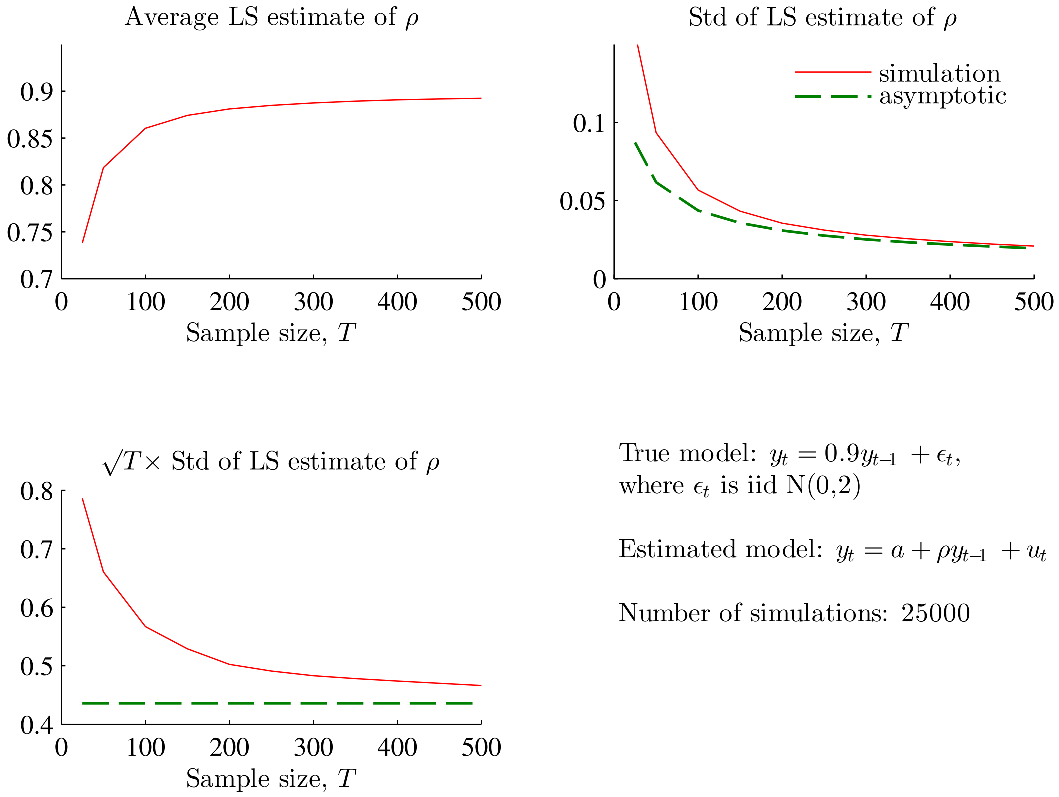


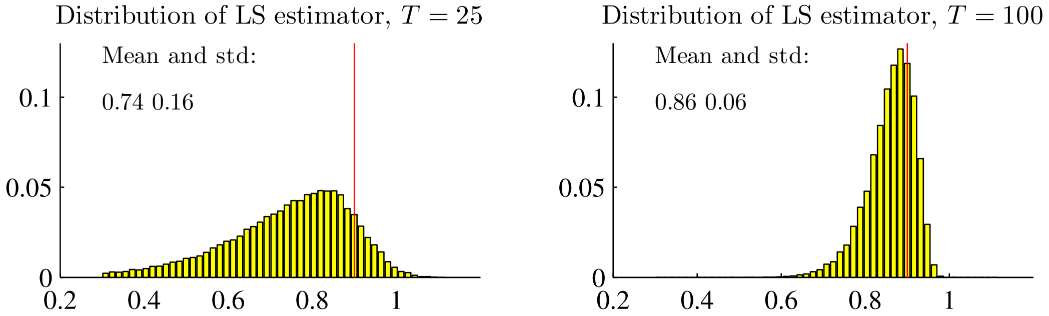
Figure 17.2: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

17.1.2 Monte Carlo Simulations with more Complicated Errors*

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 17.4* for an example.

Remark 17.2 *Let $X \sim U(0,1)$ and consider the transformation $Y = F^{-1}(X)$, where $F^{-1}()$ is the inverse of a strictly increasing cumulative distribution function F, then Y has the cdf F.*

Example 17.3 *The exponential cdf is $x = 1-\exp(-\theta y)$ with inverse $y = -\ln(1-x)/\theta$. Draw x from $U(0.1)$ and transform to y to get an exponentially distributed variable.*



True model: $y_t = 0.9y_{t-1} + \epsilon_t$, ϵ_t is iid N(0,2)
Estimated model: $y_t = a + \rho y_{t-1} + u_t$
Number of simulations: 25000

Figure 17.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

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 that process from the fitted errors and then generate artificial samples of errors (here by an AR(2))

$$\tilde{u}_t = a_1\tilde{u}_{t-1} + a_2\tilde{u}_{t-2} + \tilde{\varepsilon}_t. \tag{17.3}$$

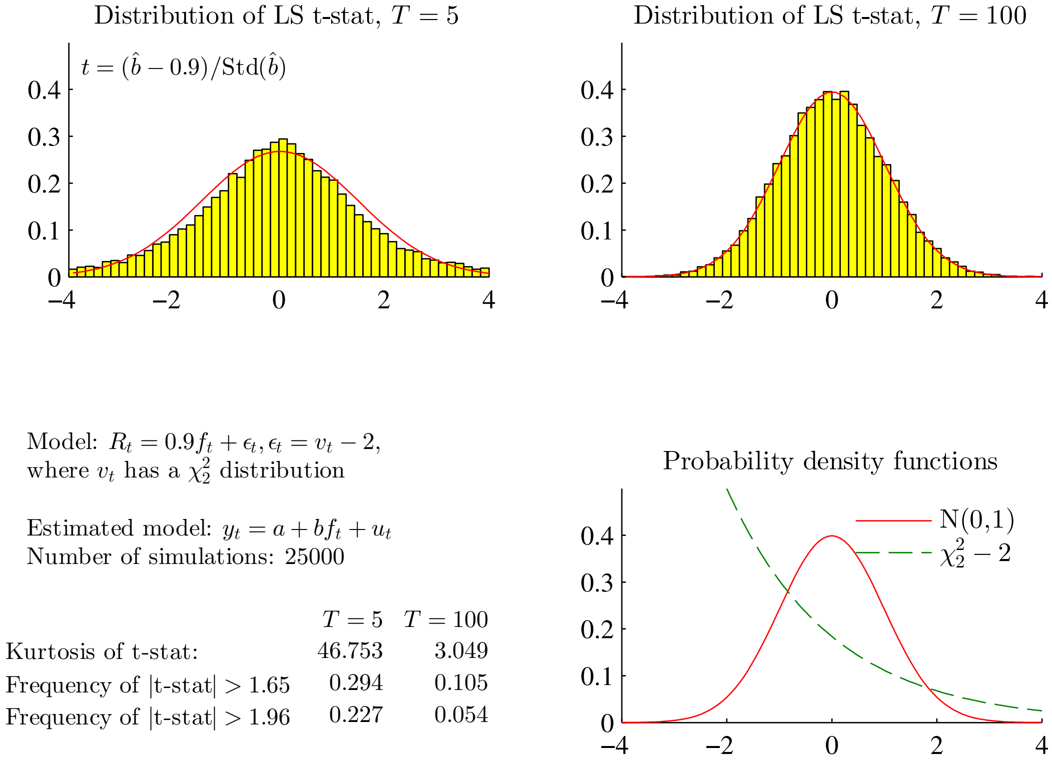
Alternatively, *heteroskedastic errors* can be generated by, for instance, a GARCH(1,1) model

$$u_t \sim N(0, \sigma_t^2), \text{ where } \sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{17.4}$$

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 fitted values from

$$\varepsilon_t^2 = c'w_t + \eta_t, \tag{17.5}$$

where w_t include the squares and cross product of all the regressors.



	$T = 5$	$T = 100$
Kurtosis of t-stat:	46.753	3.049
Frequency of $\lvert$t-stat$\rvert > 1.65$	0.294	0.105
Frequency of $\lvert$t-stat$\rvert > 1.96$	0.227	0.054

Figure 17.4: Results from a Monte Carlo experiment with thick-tailed errors.

17.2 Bootstrapping

17.2.1 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 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 (17.1), 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 $t = 1, \ldots, T$ by

$$\tilde{y}_t = x_t'\beta + \tilde{u}_t, \tag{17.6}$$

   where $\tilde{u}_t$ is drawn (with replacement) from the fitted residual and where β is the point estimate.

Example 17.4 *With $T = 3$, the artificial sample could be*

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x_1'\beta_0 + u_2, x_1) \\ (x_2'\beta_0 + u_1, x_2) \\ (x_3'\beta_0 + u_2, x_3) \end{bmatrix}.$$

The approach in (17.6) works also when y_t is a vector of dependent variables—and will then help retain the cross-sectional correlation of the residuals.

17.2.2 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 (17.6).

On 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 $(\tilde{y}_t, \tilde{x}_t) = (x_s'\beta_0 + u_s, x_s)$ 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 17.5 *With $T = 3$, the artificial sample could be*

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x_2'\beta_0 + u_2, x_2) \\ (x_3'\beta_0 + u_3, x_3) \\ (x_3'\beta_0 + u_3, x_3) \end{bmatrix}$$

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.

17.2.3 Autocorrelated Errors*

It is quite hard to handle the case when the errors are serially dependent, since we must the sample in such a way that we do not destroy the autocorrelation structure of the data. A common approach is to fit a 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 data. For instance, suppose the sample has 10 observations, and we decide to create blocks of 3 observations. The first block is $(\hat{u}_1, \hat{u}_2, \hat{u}_3)$, the second block is $(\hat{u}_2, \hat{u}_3, \hat{u}_4)$, and so forth until the last block, $(\hat{u}_8, \hat{u}_9, \hat{u}_{10})$. If we need a sample of length 3τ, say, then we simply draw τ of those block randomly (with replacement) and stack them to form a longer series. To handle end point effects (so that all data points have the same probability to be drawn), we also create blocks by "wrapping" the data around a circle. In practice, this means that we add a the following blocks: $(\hat{u}_{10}, \hat{u}_1, \hat{u}_2)$ and $(\hat{u}_9, \hat{u}_{10}, \hat{u}_1)$. The length of the blocks should clearly depend on the degree of autocorrelation, but $T^{1/3}$ is sometimes recommended as a rough guide. An alternative approach is to have non-overlapping blocks. See Berkowitz and Kilian (2000) for some other approaches.

See Figures 17.5–17.6 for an illustration.

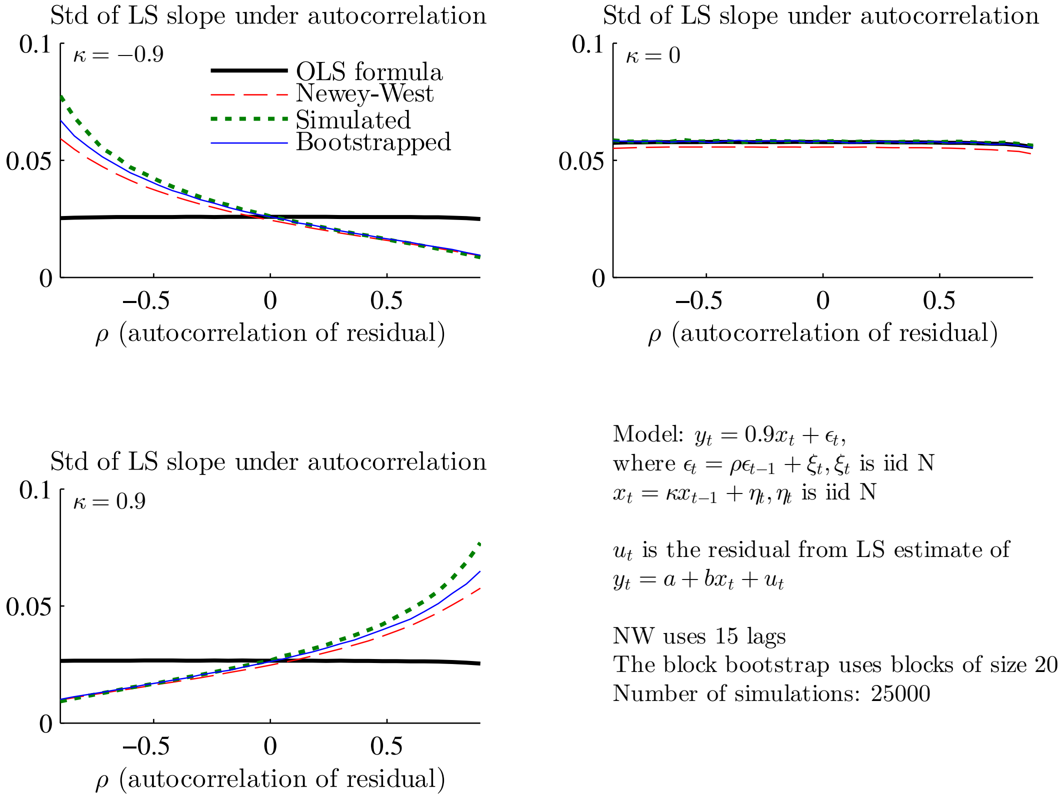
Bibliography

Berkowitz, J., and L. Kilian, 2000, "Recent developments in bootstrapping time series," *Econometric-Reviews*, 19, 1–48.

Cochrane, J. H., 2001, *Asset pricing*, Princeton University Press, Princeton, New Jersey.

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.



Model: $y_t = 0.9x_t + \epsilon_t$,
where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t, \xi_t$ is iid N
$x_t = \kappa x_{t-1} + \eta_t, \eta_t$ is iid N

u_t is the residual from LS estimate of
$y_t = a + bx_t + u_t$

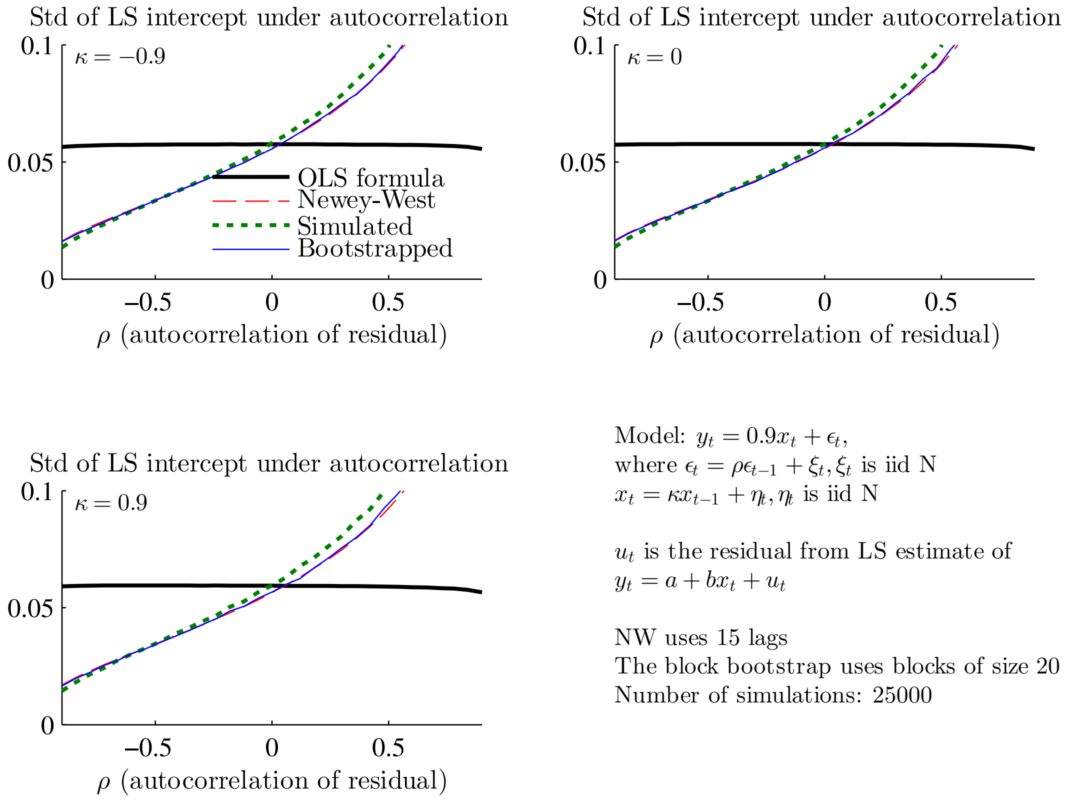
NW uses 15 lags
The block bootstrap uses blocks of size 20
Number of simulations: 25000

Figure 17.5: Standard error of OLS estimator, autocorrelated errors

Efron, B., and R. J. Tibshirani, 1993, *An introduction to the bootstrap*, Chapman and Hall, New York.

Greene, W. H., 2000, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Horowitz, J. L., 2001, "The Bootstrap," in J.J. Heckman, and E. Leamer (ed.), *Handbook of Econometrics* . , vol. 5, Elsevier.


Std of LS intercept under autocorrelation

Std of LS intercept under autocorrelation

Std of LS intercept under autocorrelation

Model: $y_t = 0.9x_t + \epsilon_t$,
where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t, \xi_t$ is iid N
$x_t = \kappa x_{t-1} + \eta_t, \eta_t$ is iid N

u_t is the residual from LS estimate of
$y_t = a + bx_t + u_t$

NW uses 15 lags
The block bootstrap uses blocks of size 20
Number of simulations: 25000

Figure 17.6: Standard error of OLS estimator, autocorrelated errors

18 Panel Data*

References: Verbeek (2012) 10 and Baltagi (2008)

18.1 Introduction to Panel Data

A panel data set has data on a cross-section ($i = 1, 2, \ldots, N$, individuals or firms) over many time periods ($t = 1, 2, \ldots, T$). The aim is to estimate a linear relation between the dependent variable and the regressors

$$y_{it} = x_{it}'\beta + \varepsilon_{it}. \tag{18.1}$$

For instance, data on the dependent variable might have this structure

$$\begin{bmatrix} & \underline{i=1} & \underline{i=2} & \cdots & \underline{i=N} \\ t=1: & y_{11} & y_{21} & & y_{N1} \\ t=2: & y_{12} & y_{22} & & y_{N2} \\ \vdots & & & & \\ t=T: & y_{1T} & y_{2T} & & y_{NT} \end{bmatrix} \tag{18.2}$$

The structure for each of the regressors is similar.

The most basic estimation approach is to just run LS (on all observations "stacked"). This is not that bad (although GLS might be more efficient), especially since there is typically lots of data points. However, we may want to allow for individual (i) effects.

18.2 Fixed Effects Model

In the fixed effects model, we allow for different individual intercepts

$$y_{it} = \mu_i + x_{it}'\beta + \varepsilon_{it}, \varepsilon_{it} \text{ is iid} N(0, \sigma_\varepsilon^2), \tag{18.3}$$

and maintain the basic assumption that ε_{jt} is uncorrelated with x_{it} (across all i and j).

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. (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 (that gives the same point estimates as OLS with dummies) consider the following approach. First, take average across time (for a given i) of y_{it} and x_{it} in (18.3. That is, think (but d not run any estimation...) of forming the cross-sectional regression

$$\bar{y}_i = \mu_i + \bar{x}_i'\beta + \bar{\varepsilon}_i, \text{ where} \tag{18.4}$$

$$\bar{y}_i = \frac{1}{T}\sum\nolimits_{t=1}^{T} y_{it} \text{ and } \bar{x}_i = \frac{1}{T}\sum\nolimits_{t=1}^{T} x_{it}. \tag{18.5}$$

Second, subtract from (18.3) to get

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i). \tag{18.6}$$

At this stage, estimate β by running LS on all observations of (18.6) "stacked." 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 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

$$\mu_i = \bar{y}_i - \bar{x}_i'\hat{\beta}_{FE}. \tag{18.7}$$

Clearly, the within estimator wipes out all regressors that are constant across time for a given individual (say, gender and schooling) : they are effectly merged with the individual means (μ_i).

We can apply the usual tests (possibly, small-sample adjustment of standard errors). In particular, we can estimate the standard error of the residual as

$$\sigma_u^2 = \frac{1}{TN}\sum\nolimits_{t=1}^{T}\sum\nolimits_{i=1}^{N}\hat{u}_{it}^2, \text{ where} \tag{18.8}$$

$$\hat{u}_{it}^2 = y_{it} = \hat{\mu}_i - x_{it}'\hat{\beta}_{FE},$$

and the covariance matrix of the slope coefficients as

$$\text{Var}(\hat{\beta}_{FE}) = \sigma^2 S_{xx}^{-1} \text{ where } S_{xx} = \textstyle\sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'. \tag{18.9}$$

Notice that these results (on the variance of the slope coefficients) rely on the assumption that the residuals are uncorrelated across time and individuals.

Example 18.1 *$N = 2, T = 2$. 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 18.2 *(Lagged dependent variable as regressor.) If $y_{i,t-1}$ is among the regressors x_{it}, then the within estimator (18.6) is biased in small 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 (18.6) is then*

$$y_{it} - \textstyle\sum_{t=1}^{T} y_{it}/T = \left(y_{i,t-1} - \sum_{t=2}^{T} y_{i,t-1}/(T-1)\right)\beta + \left(\varepsilon_{it} - \sum_{t=1}^{T}\varepsilon_{it}/T\right).$$

The problem is that $y_{i,t-1}$ is correlated with $\sum_{t=1}^{T}\varepsilon_{it}/T$ since the latter contains $\varepsilon_{i,t-1}$ which affects $y_{i,t-1}$ directly. In addition, $\sum_{t=2}^{T} y_{i,t-1}/T$ contains $y_{i,t}$ which is correlated with ε_{it}. It can be shown that this bias can be substantial for panels with small T.

An another way of estimating the fixed effects model is to difference away the μ_i by taking *first-differences* (in time)

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + \underbrace{\varepsilon_{it} - \varepsilon_{i,t-1}}_{\text{residual } u_{it}}. \tag{18.10}$$

This can be estimated by OLS, but we could adjust the covariance matrix of the slope coefficients, since the residuals are now (negatively) autocorrelated ($\varepsilon_{i,t-1}$ shows up both

	LS	Fixed eff	Between	GLS
exper/100	7.84	4.11	10.64	4.57
	(8.25)	(6.21)	(4.05)	(7.12)
exper2/100	−0.20	−0.04	−0.32	−0.06
	(−5.04)	(−1.50)	(−2.83)	(−2.37)
tenure/100	1.21	1.39	1.25	1.38
	(2.47)	(4.25)	(0.90)	(4.32)
tenure2/100	−0.02	−0.09	−0.02	−0.07
	(−0.85)	(−4.36)	(−0.20)	(−3.77)
south	−0.20	−0.02	−0.20	−0.13
	(−13.51)	(−0.45)	(−6.67)	(−5.70)
union	0.11	0.06	0.12	0.07
	(6.72)	(4.47)	(3.09)	(5.57)

Table 18.1: 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.

in t and $t-1$, but with different signs). Becuase of the negative autocorrelation, unadjusted standard errors are likely to overstate the uncertainty—and carn therefore be used as a conservative approach. 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).

Example 18.3 *$N = 2, T = 2$. Stack the data for individual $i = 1$ first and those for individual $i = N$ second*

$$\text{Cov}\begin{pmatrix} u_{1,T-1} \\ u_{1T} \\ u_{N,T-1} \\ u_{NT} \end{pmatrix} = \text{Cov}\begin{pmatrix} \varepsilon_{1,T-1} - \varepsilon_{1,T-2} \\ \varepsilon_{1,T} - \varepsilon_{1,T-1} \\ \varepsilon_{N,T-1} - \varepsilon_{N,T-2} \\ \varepsilon_{N,T} - \varepsilon_{N,T-1} \end{pmatrix} = \begin{bmatrix} 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & 0 & 0 \\ -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 \\ 0 & 0 & -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 \end{bmatrix}.$$

Remark 18.4 *(Difference-in-difference estimator) Suppose there are only two periods ($T = 2$) and that one of the regerssors is a dummy that equals one for a some individuals who got a "treatment" (say, extra schooling) between the two periods and zero for the other individuals. Running the first-difference method (18.10) and studying the coefficient of that dummy variable is then the so called "difference-in-difference" method. It* *measures how much the dependent variable changed for those with treatment compared to the change for those without the treatment.*

Remark 18.5 *(Lagged dependent variable as regressor) If $y_{i,t-1}$ is among the regressors x_{it}, then the first-difference method (18.10) does not work (OLS is inconsistent). 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).*

18.3 Random Effects Model

The random effects model allows for *random* individual "intercepts" (μ_i)

$$y_{it} = \beta_0 + x_{it}'\beta + \mu_i + \varepsilon_{it}, \text{ where} \tag{18.11}$$

$$\varepsilon_{it} \text{ is iid} N(0, \sigma_\varepsilon^2) \text{ and } \mu_i \text{ is iid} N(0, \sigma_\mu^2). \tag{18.12}$$

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. For instance, suppose the panel is drawn for a large sample so as to be representative. This means, effectively, that the sample contains values of (y_{it}, x_{it}) that match those of the population. An example could be that one of the x_{it} variables measure age of individuals—and the sample is drawn so that it has the same agre distribution as the population. In this setting, a random μ_i makes sense as a proxy for whatever information we leave out. Clearly, if the 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 therefore write the regression as

$$y_{it} = \beta_0 + x_{it}'\beta + u_{it}, \text{where } u_{it} = \mu_i + \varepsilon_{it}, \tag{18.13}$$

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

$$\bar{y}_i = \beta_0 + \bar{x}_i'\beta + \underbrace{\mu_i + \bar{\varepsilon}_i}_{\text{residual}_i}, \tag{18.14}$$

is also consistent, but discards all time-series information. Third, LS on

$$y_{it} = \beta_0 + x_{it}'\beta + \underbrace{\mu_i + \varepsilon_{it}}_{\text{residual}_{it}} \tag{18.15}$$

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.

Example 18.6 *$N = 2, T = 2$. 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 18.7 *(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*

$$\begin{aligned} y_{it} - \vartheta\bar{y}_i &= \beta_0(1-\vartheta) + (x_{it} - \vartheta\bar{x}_i)'\beta + \upsilon_{it}, \text{ where} \\ \vartheta &= 1 - \sqrt{\sigma_u^2/(\sigma_u^2 + T\sigma_\mu^2)}. \end{aligned}$$

In this equation, σ_u^2 is the variance of the residuals in the "within regression" as estimated in (18.8) 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" (18.14).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 (18.6). This happens when $\sigma_\mu^2 >> \sigma_u^2$ 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.

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 (cf. the FE model) works, since it effectively eliminates the excluded variable from the system.

Bibliography

Baltagi, D. H., 2008, *Econometric Analysis of Panel Data*, Wiley, 4th edn.

Verbeek, M., 2012, *A guide to modern econometrics*, Wiley, 4th edn.

19 Binary Choice Models*

Reference: Verbeek (2012) 7

19.1 Binary Choice Model

19.1.1 Basic Model Setup

A binary variable

$$y_i = \begin{cases} 0 & \text{firm } i \text{ doesn't pay dividends} \\ 1 & \text{firm } i \text{ pays dividends} \end{cases} \tag{19.1}$$

We know a few things about firm i: x_i (industry, size, profitability...)

Model: the probability that firm i pays dividends is some function of x_i

$$\Pr(y_i = 1|x_i) = F(x_i'\beta) \tag{19.2}$$

Idea: if x_i contains profitability, then (presumably) most firms with high profits will have dividends. What you estimate is (effectively) how the typical pattern changes with profits.

What function $F()$ should we use in (19.2)? Mostly a matter of convenience. A *probit model* assumed that $F()$ is a standard normal cumulative distribution function, see Figure 19.1. Other choices of $F()$ give *logit model* ($F()$ is a logistic function) or *linear probability model* ($F(x_i'\beta) = x_i'\beta$). See Figure 19.2 for an illustration.

How to interpret the results? Mostly by looking at the marginal effects

$$\frac{\partial F(x_i'\beta)}{\partial x_i} \tag{19.3}$$

For instance, how does the probability of having dividends change when profits changes?

Example 19.1 *Constant plus two more regressors (w and z): $x_i'\beta = \beta_0 + \beta_1 w_i + \beta_2 z_i$, then*

$$\frac{\partial F(x_i'\beta)}{\partial w_i} = f(\beta_0 + \beta_1 w_i + \beta_2 z_i)\beta_1,$$

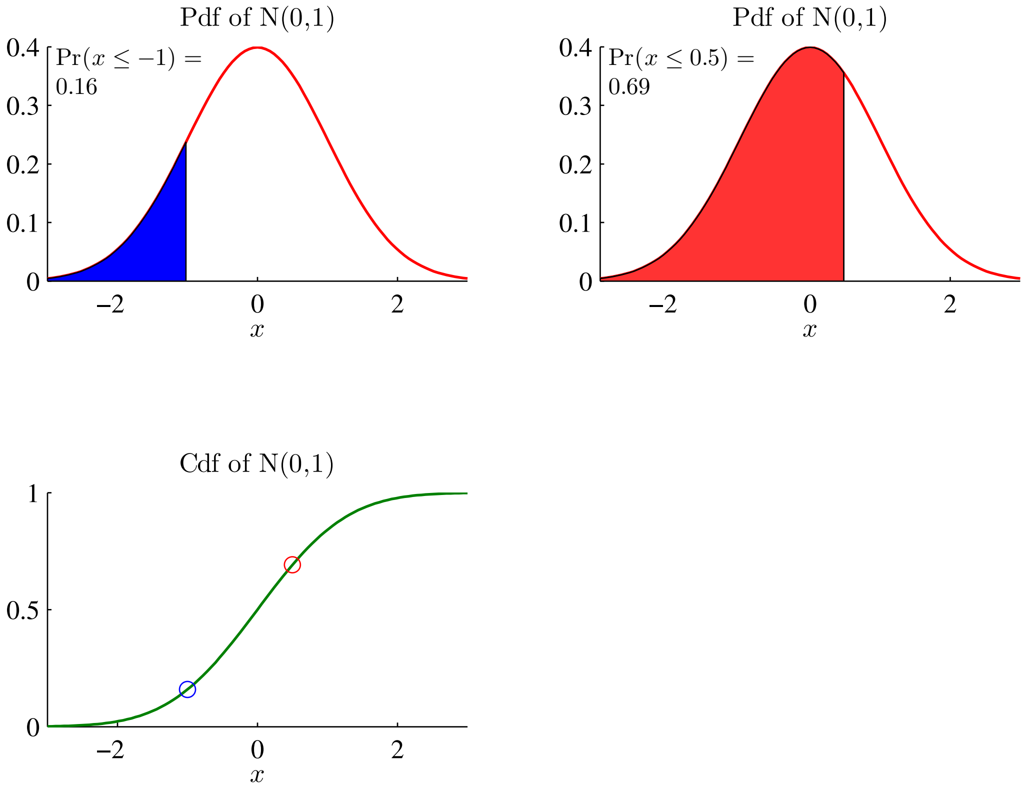


Figure 19.1: Pdf and cdf of N(0,1)

where $f()$ is the derivative of $F()$. Sign(derivative)=sign(β_1) Calculated at some typical value of $\beta_0 + \beta_1 w_i + \beta_2 z_i$.

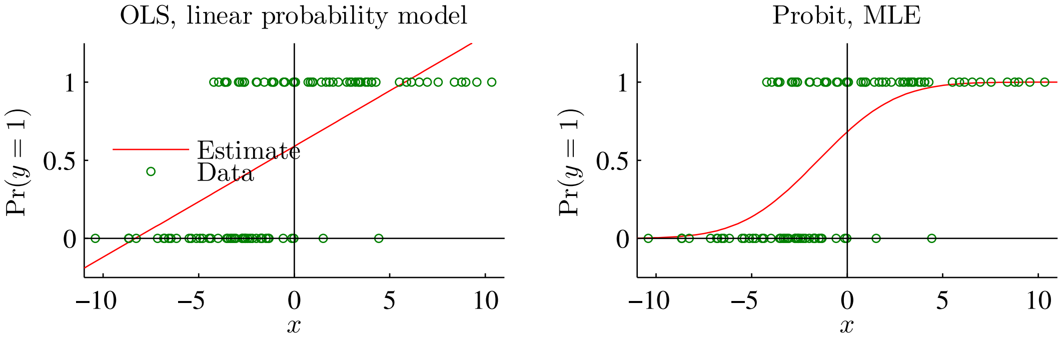
Example 19.2 *If a regressor is a dummy variable, then use a simple difference instead of attempting a derivative. For instance, if z_i is either 0 or 1, then we can use*

$$F(\beta_0 + \beta_1 w_i + \beta_2) - F(\beta_0 + \beta_1 w_i).$$

This is calculated at some typical value of $\beta_0 + \beta_1 w_i$.

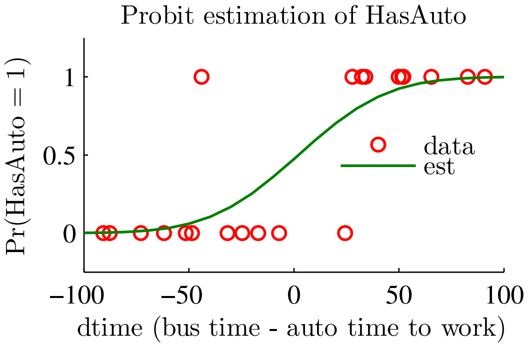
Notice: the ratio of two coefficients equals the ratio of their marginal effect on the probability

$$\beta_k/\beta_m = \frac{\partial F(x_i'\beta)}{\partial x_{k,i}} / \frac{\partial F(x_i'\beta)}{\partial x_{m,i}}$$



$y = 1$ if $y^* > 0$ and zero otherwise
$y^* = 2 + 1.3x + \epsilon, \epsilon \sim N(0, 4)$
$x \sim N(0, 5)$

Figure 19.2: Example of probit model



Probit estimation of HasAuto (0 or 1)

	coeff	t-stat
const	-0.06	-0.16
dtime	0.03	2.92

McFadden's R^2: 0.58

Figure 19.3: Example of probit model, Hill et al (2008), Table 16.1

19.1.2 Estimation

The model is typically estimated with MLE. To do that we need to construct the likelihood function.

Remark 19.3 *Bernoulli distribution.* $\Pr(y_i = 1) = p_i$, $\Pr(y_i = 0) = 1 - p_i$.

Assume independent observations (firm 1 and 2). Then, the probabilities (likelihoods) for the different outcomes are

$$\begin{aligned}
\Pr(y_1 = 1 \text{ and } y_2 = 1) &= p_1 p_2 \\
\Pr(y_1 = 1 \text{ and } y_2 = 0) &= p_1(1 - p_2) \\
\Pr(y_1 = 0 \text{ and } y_2 = 1) &= (1 - p_1)\, p_2 \\
\Pr(y_1 = 0 \text{ and } y_2 = 0) &= (1 - p_1)\,(1 - p_2)
\end{aligned} \tag{19.4}$$

This list will be long (and messy to program) when there are many observations (firms). We therefore use an alternative way of writing the same thing as in (19.4). First, notice that

$$p_1^{y_1}(1 - p_1)^{1-y_1} = \begin{cases} p_1 & \text{if } y_1 = 1 \\ 1 - p_1 & \text{if } y_1 = 0. \end{cases} \tag{19.5}$$

For the sample wit two data points, the probability (likelihood) can be written

$$L = p_1^{y_1}(1 - p_1)^{1-y_1} \times p_2^{y_2}(1 - p_2)^{1-y_2}. \tag{19.6}$$

Let $p_i = F(x_i'\beta)$ from (19.2) and use in (19.6) to get a likelihood function for two data points

$$L = F(x_1'\beta)^{y_1}\left[1 - F(x_1'\beta)\right]^{1-y_1} \times F(x_2'\beta)^{y_2}\left[1 - F(x_2'\beta)\right]^{1-y_2}.$$

or as log (after slight rearranging)

$$\begin{aligned}
\ln L = {} & y_1 \ln F(x_1'\beta) + y_2 \ln F(x_2'\beta) \\
& + (1 - y_1)\ln\left[1 - F(x_1'\beta)\right] + (1 - y_2)\ln\left[1 - F(x_2'\beta)\right].
\end{aligned} \tag{19.7}$$

For N data points, this generalizes to

$$\ln L = \textstyle\sum_{i=1}^{N} y_i \ln F(x_i'\beta) + (1 - y_i)\ln\left[1 - F(x_i'\beta)\right]. \tag{19.8}$$

We find the ML estimate by maximizing this log likelihood function with respect to the parameters β.

See Figure 19.3 for an empirical example.

19.1.3 Goodness of Fit

To measure of 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 common with a traditional R^2. It is

$$\text{McFadden's } R^2 = 1 - \frac{\text{log likelihood value (at max)}}{\text{log likelihood value (all coeffs=0, except constant)}}. \tag{19.9}$$

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 (19.9) is therefore between 0 (as bad as a model with only a constant) and 1 (perfect model).

Example 19.4 *If* $\ln L = \ln 0.9$ *(at max) and the model with only a constant has* $\ln L = \ln 0.5$

$$\textit{McFadden's } R^2 = 1 - \frac{\ln 0.9}{\ln 0.5} \approx 0.84$$

If instead, the model has $\ln L = \ln 0.8$ *(at max), then*

$$\textit{McFadden's } R^2 = 1 - \frac{\ln 0.8}{\ln 0.5} \approx 0.68$$

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

$$\hat{y}_i = \begin{cases} 1 & \text{if } F(x_i'\hat{\beta}) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{19.10}$$

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.

$$\begin{array}{cccc} & \hat{y}_i = 0 & \hat{y}_i = 1 & \text{Total} \\ y_i = 0: & n_{00} & n_{01} & N_0 \\ y_i = 1: & n_{10} & n_{11} & N_1 \\ \text{Total:} & \hat{N}_0 & \hat{N}_1 & N \end{array} \tag{19.11}$$

For instance, n_{01} is the number of data points for which $y_i = 0$ but $\hat{y}_i = 1$. Similarly, N_0 is the number of observations for which $y_i = 0$ (and it clearly equals $n_{00} + n_{01}$). Define an "R^2_{pred}" for the prediction as

$$\text{“}R^2_{pred}\text{”} = 1 - \frac{\text{number of incorrect predictions}}{\text{number of incorrect predictions, constant probabilities}}. \tag{19.12}$$

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

$$\begin{aligned}\hat{p} &= \text{Fraction of } (y_1 = 1)\\ &= N_1/N. \end{aligned} \tag{19.13}$$

If $\hat{p} \leq 0.5$, so the (naive) constant probability model predicts $y_i = 0$, then the number of incorrect predictions 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 (19.12) as

$$\text{“}R^2_{pred}\text{”} = \begin{cases} 1 - \frac{n_{10}+n_{01}}{N_1} & \text{if } \hat{p} \leq 0.5 \\ 1 - \frac{n_{10}+n_{01}}{N_0} & \text{if } \hat{p} > 0.5. \end{cases}$$

Example 19.5 *Let x_i be a scalar. Suppose we have the following data*

$$\begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 1 \end{bmatrix} \text{ and}$$
$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 1.5 & -1.2 & 0.5 & -0.7 \end{bmatrix}$$

See Figure 19.4

Suppose $\beta = 0$, then we get the following values

$$F(x_i'\beta) = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}$$

$$\begin{aligned} & y_i \log F(x_i'\beta) + (1 - y_i) \log \left[1 - F(x_i'\beta)\right] \\ & \approx \begin{bmatrix} -0.69 & -0.69 & -0.69 & -0.69 \end{bmatrix} \\ \log L & \approx -2.77 \end{aligned}$$

Now, suppose instead that $\beta = 1$

$$F(x_i'\beta) \approx \begin{bmatrix} 0.93 & 0.12 & 0.69 & 0.24 \end{bmatrix}$$

$$\begin{aligned} & y_i \log F(x_i'\beta) + (1 - y_i) \log \left[1 - F(x_i'\beta)\right] \\ & \approx \begin{bmatrix} -0.07 & -0.12 & -0.37 & -1.42 \end{bmatrix} \\ \log L & \approx -1.98, \end{aligned}$$

which is higher than at $\beta = 0$*. If* $\beta = 1$ *happened to maximize the likelihood function (it almost does...), then*

$$\textit{McFadden's } R^2 = 1 - \frac{-1.98}{-2.77} \approx 0.29$$

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*

$$\text{“}R^2_{pred}\text{”} = 1 - \frac{1}{1+0} = 0.$$

19.1.4 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 $(0, 1, 2, 3, ..)$.

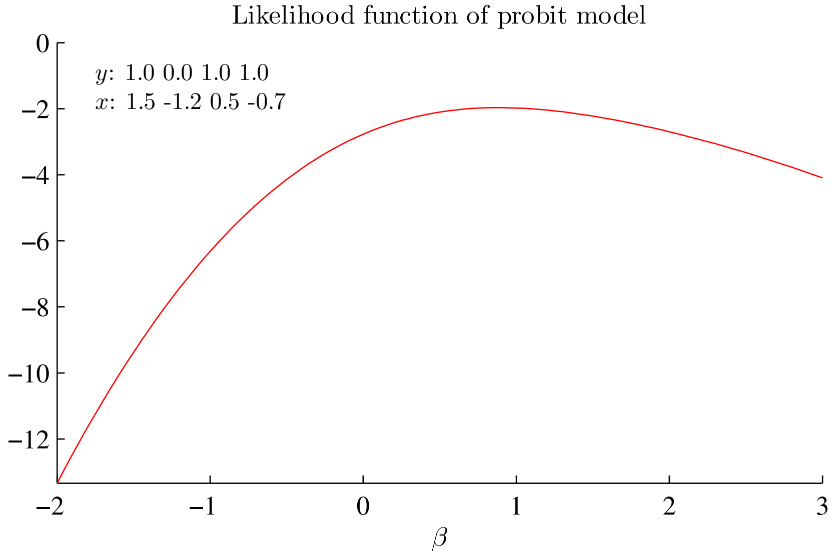


Figure 19.4: Example of ML estimation of probit model

19.2 Truncated Regression Model

19.2.1 Basic Model Setup

Suppose the correct model is linear

$$y_i^* = x_i'\beta + \varepsilon_i, \varepsilon_i \sim \text{iid}N(0, \sigma^2), \tag{19.14}$$

but that data (also regressors) are completely missing if $y_i^* \leq 0$

$$\begin{bmatrix} y_i = y_i^* & \text{if } y_i^* > 0 \\ (y_i, x_i) \text{ not observed} & \text{otherwise.} \end{bmatrix} \tag{19.15}$$

The problem with this is that the sample is no longer random. For instance, if y_i^* is dividends, x_i is profits—and it so happens that firms with low dividends are not in the sample. See Figure 19.7 for an illustration.

In fact, running OLS of

$$y_i = x_i'\beta + \varepsilon_i \tag{19.16}$$

on the available data will give biased (and inconsistent) estimates.

The reason is that we only use those data points where y_i is unusually high (for a given value of $x_i'\beta$). To be precise, the expected value of y_i, conditional on $x_i'\beta$ and that we observe the data, is

$$\begin{aligned} \mathrm{E}\,(y_i|y_i > 0, x_i) &= x_i'\beta + \mathrm{E}\left(\varepsilon_i|y_i^* > 0\right) && (19.17)\\ &= x_i'\beta + \mathrm{E}\left(\varepsilon_i|\varepsilon_i > -x_i'\beta\right). && (19.18)\end{aligned}$$

The second line follows from the fact that $y_i^* > 0$ happens when $x_i'\beta + \varepsilon_i > 0$ (see (19.15)) which happens when $\varepsilon_i > -x_i'\beta$. The key result is that last term is positive (recall $\mathrm{E}\,\varepsilon_i = \mathrm{E}\,(\varepsilon_i|\varepsilon_i > -\infty) = 0$), which make the OLS estimates inconsistent. The result in (19.18) means that our data has a higher mean that the corresponding $x_i'\beta$ would motivate. Since OLS creates the estimates to make sure that the residual has a zero mean, so OLS will tilt the coefficient estimates away from β. The basic reason is that $\mathrm{E}\left(\varepsilon_i|\varepsilon_i > -x_i'\beta\right)$ varies if x_i, so it acts like an extra error term that is correlated with the regressor—which is a classical reason for why OLS is inconsistent. The following examples illustrate how.

Remark 19.6 *(Truncated normal distribution) Let $\varepsilon \sim N(\mu, \sigma^2)$, then*

$$\mathrm{E}(\varepsilon|\varepsilon > a) = \mu + \sigma\frac{\phi(a_0)}{1 - \Phi(a_0)} \text{ and } a_0 = (a - \mu)/\sigma$$

See Figure 19.5.

Example 19.7 *As a trivial example, suppose the model is $y_i^* = 0 + \varepsilon_i$ with $\varepsilon_i \sim iidN(0, 1)$. Then*

$$\begin{aligned} \mathrm{E}\,(y_i|y_i > 0, x_i) &= 0 + \mathrm{E}\,(\varepsilon_i|\varepsilon_i > 0)\\ &= 0 + \frac{\phi(0)}{1 - \Phi(0)} = \sqrt{2/\pi} \approx 0.80,\end{aligned}$$

which is far from the true mean (0). OLS will therefore estimate an intercept of around 0.8 instead of 0.

Example 19.8 *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*

$$\begin{aligned} \mathrm{E}\,(y_i|y_i > 0, x_i) &= 2x_i + \mathrm{E}\,(\varepsilon_i|\varepsilon_i > -2x_i)\\ &= 2x_i + \frac{\phi(-2x_i)}{1 - \Phi(-2x_i)}\end{aligned}$$

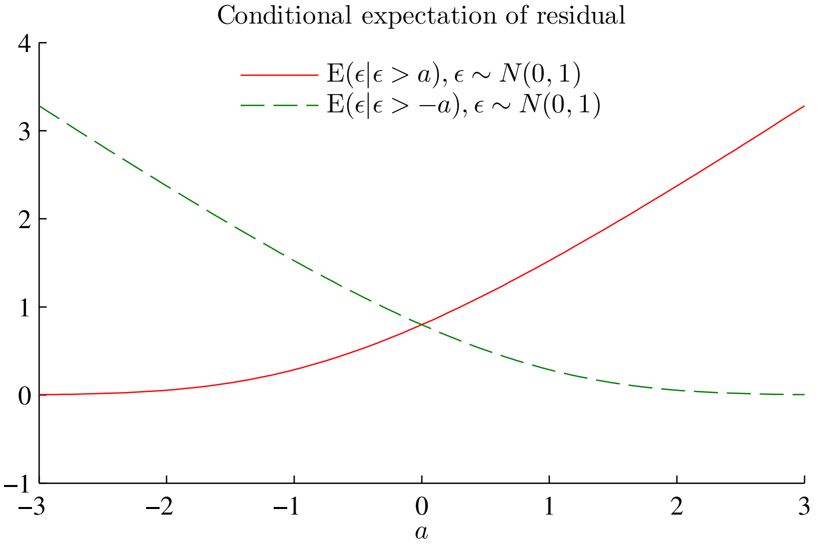


Figure 19.5: Expectations of a truncated variable

For some selected values of x_i we have

$$\begin{aligned}
\mathrm{E}(y_i|y_i > 0, x_i) &= \\
&= 2x_i + \mathrm{E}(\varepsilon_i|\varepsilon_i > -2x_i) \\
&= \begin{cases} 2\times(-1) + \mathrm{E}(\varepsilon_i|\varepsilon_i > 2) & x = -1 \\ 2\times 0 + \mathrm{E}(\varepsilon_i|\varepsilon_i > 0) & x_i = 0 \\ 2\times 1 + \mathrm{E}(\varepsilon_i|\varepsilon_i > -2) & x_i = 1 \end{cases} \\
&= \begin{cases} 2\times(-1) + 2.37 = 0.37 & x = -1 \\ 2\times 0 + 0.8 = 0.80 & x_i = 0 \\ 2\times 1 + 0.06 = 2.06 & x_i = 1 \end{cases}
\end{aligned}$$

so the slope is lower than 2: OLS will therefor fit a slope coefficient that is lower than 2. See Figure 19.6. The basic point is that $\mathrm{E}(\varepsilon_i|\varepsilon_i > -2x_i)$ *is much higher for low than 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.)*

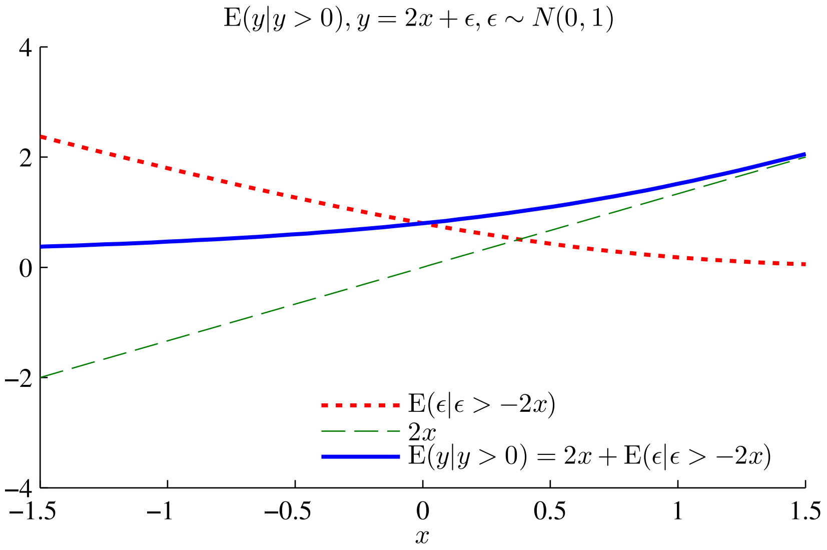


Figure 19.6: Expectations of a truncated variable

19.2.2 Estimation

Remark 19.9 *(Distribution of truncated a random variable) Let the density function of a random variable X be* pdf(x)*. The density function, conditional on $a < X \leq b$ is* pdf$(x|a < X \leq b) = \text{pdf}(x)/\Pr(a < X \leq b)$*. Clearly, it could be that $a = -\infty$ and/or $b = \infty$..*

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 $\varepsilon_i > -x_i'\beta$)

$$\text{pdf}(\varepsilon_i|\varepsilon_i > -x_i'\beta) = \frac{\text{pdf}(\varepsilon_i)}{\Pr(\varepsilon_i > -x_i'\beta)}. \tag{19.19}$$

If $\varepsilon_i \sim N(0, \sigma^2)$, the denominator is

$$\begin{aligned}\Pr(\varepsilon_i > -x_i'\beta) &= \Pr\left(\varepsilon_i/\sigma > -x_i'\beta/\sigma\right) &(19.20)\\ &= 1 - \Phi\left(-x_i'\beta/\sigma\right) &(19.21)\\ &= \Phi\left(x_i'\beta/\sigma\right). &(19.22)\end{aligned}$$

The second line follows from $N(0, 1)$ being symmetric around 0, so $\Phi(z) = 1 - \Phi(-z)$.

Combine (19.22) with a $N(0, \sigma^2)$ distribution for ε_i, replace ε_i by $y_i - x_i'\beta$ and take logs to get

$$\ln L = \sum_i \ln L_i \text{, where} \tag{19.23}$$

$$L_i = \text{pdf}(\varepsilon_i | \varepsilon_i > -x_i'\beta) \times \Pr(y_i > 0) \tag{19.24}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\left(y_i - x_i'\beta\right)^2}{\sigma^2}\right) / \Phi\left(x_i'\beta/\sigma\right). \tag{19.25}$$

We maximize this likelihood function with respect to β and σ^2 (numerical optimization). Notice: $\Phi\left(x_i'\beta/\sigma\right)$ is the new part compared with OLS. See Figure 19.7 for an illustration.

19.3 Censored Regression Model (Tobit Model)

The censored regression model is similar to truncated model, but we are 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

$$y_i^* = x_i'\beta + \varepsilon_i, \varepsilon_i \sim \text{iid}N(0, \sigma^2) \tag{19.26}$$

$$\text{Data: } y_i = \begin{cases} (y_i^*, x_i) & \text{if } y_i^* > 0 \\ (0, x_i) & \text{otherwise.} \end{cases}$$

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.

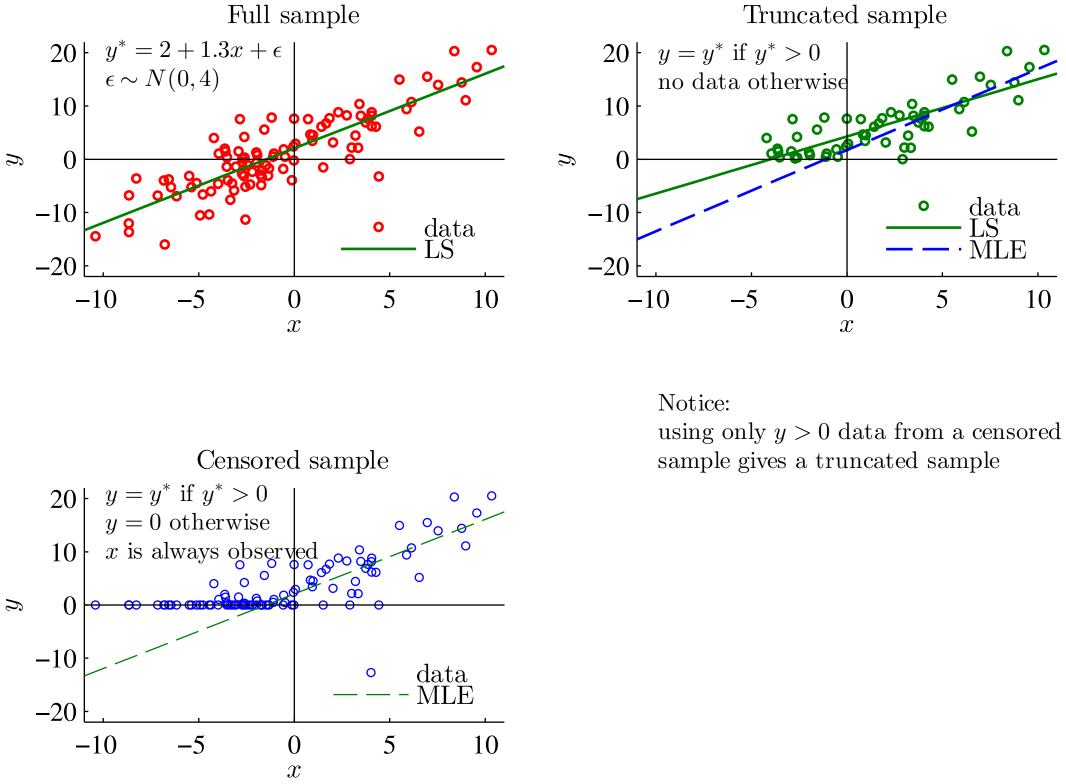
If we estimate $y_i = x_i'\beta + \varepsilon_i$ (with LS), using all data with $y_i > 0$, then we are in same situation as in truncated model: LS is not consistent. See Figure 19.7.

Example: y_i^* is dividends, x_i is profits—firms with low dividends are assigned a common value (normalized to $y_i = 0$) in the survey.

19.3.1 Estimation of Censored Regression Model

Remark 19.10 *(Likelihood function with different states) The likelihood contribution of observation i is* pdf(y_i) *which can also be written $pdf(y_i|state\ K) \times$* Pr(*state K*). *See Remark 19.9*

There are two states: $y_i^* \leq 0$ and $y_i^* > 0$.



Notice:
using only $y > 0$ data from a censored
sample gives a truncated sample

Figure 19.7: Estimation on full, truncated and censored sample

State $y_i^* \leq 0$ (that is, no data on y_i^* but on x_i) happens when $y_i^* = x_i'\beta + \varepsilon_i \leq 0$, that is, when $\varepsilon_i \leq -x_i'\beta$. The probability of this is

$$\begin{aligned}\Pr(\varepsilon_i \leq -x_i'\beta) &= \Pr(\varepsilon_i/\sigma \leq -x_i'\beta/\sigma)\\ &= \Phi(-x_i'\beta/\sigma). \qquad (19.27)\end{aligned}$$

(By symmetry of the normal distribution, this also equals $1 - \Phi(x_i'\beta/\sigma)$.) The conditional density function in this state has the constant value of one, so the likelihood contribution (see Remark 19.10) is

$$\begin{aligned}L_i(\text{if } y_i^* \leq 0) &= \text{pdf}(y_i|y_i^* \leq 0) \times \Pr(y_i^* \leq 0)\\ &= 1 \times \Phi(-x_i'\beta/\sigma). \qquad (19.28)\end{aligned}$$

State $y_i^* > 0$ happens in the same way as in the truncated model (19.19), but the dif-

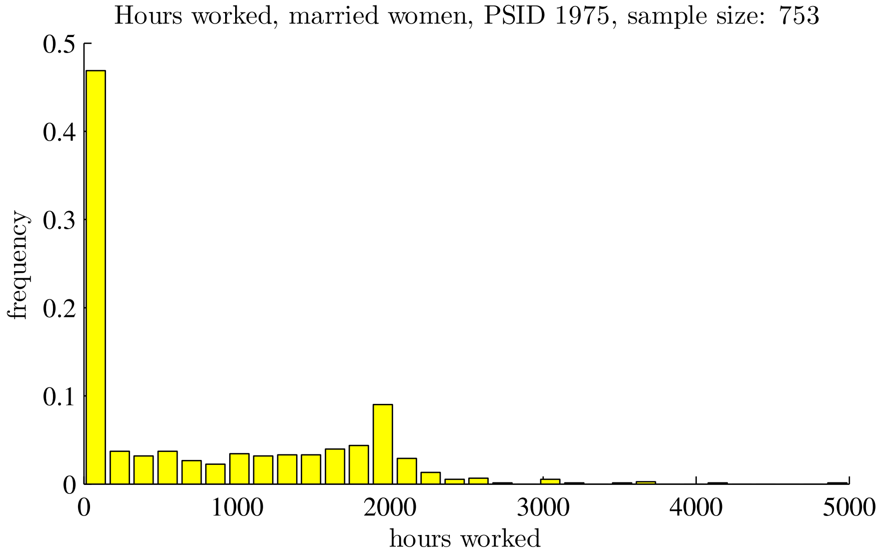


Figure 19.8: Example of probit model, Hill et al (2008), Table 16.1

ference here is that the contribution to the likelihood function (again, see Remark 19.10) is

$$\begin{aligned} L_i(\text{if } y_i^* > 0) &= \text{pdf}(\varepsilon_i | \varepsilon_i > -x_i'\beta) \times \Pr(\varepsilon_i > -x_i'\beta) \\ &= \text{pdf}(\varepsilon_i). \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{\left(y_i - x_i'\beta\right)_i^2}{\sigma^2} \right). \end{aligned} \tag{19.29}$$

The likelihood function is defined by (19.27) and (19.29). 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 (19.28)—since we know all x_i values.

19.3.2 Interpretation of the Tobit Model

We could be interested in several things. First, how is probability of $y_i = 0$ affected by a change in regressor k? The derivative provides an answer

$$\frac{\partial \Pr(y_i = 0)}{\partial x_{ik}} = -\phi(x_i'\beta/\sigma)\beta_k/\sigma. \tag{19.30}$$

	OLS	MLE
const	1335.3	1349.9
	(5.7)	3.5
educ	27.1	73.3
	(2.2)	3.5
exper	48.0	80.5
	(13.2)	12.2
age	−31.3	−60.8
	(−7.9)	−8.4
kids16	−447.9	−918.9
	(−7.7)	−8.1
sigma	753.0	1133.7
		26.8
Nobs	753.0	

Table 19.1: Tobit estimation of hours worked. Example of a tobit model, Hill et al (2008), Table 16.8. Numbers in parentheses are t-stats.

This derivative has a absolute value when $x_i'\beta \approx 0$, since a small change in x_k can then tip the balance towards $y_i = 0$. In contrast, when $x_i'\beta$ is very small or very large, then a small change in x_k does not matter much (as we are already safely in $y_i = 0$ or $y_i = 1$ territory). Second, wow is the expected value of y_i affected by a change in regressor k? Once again,we can calculate a derivative

$$\frac{\partial \mathrm{E}\, y_i}{\partial x_{ik}} = \Phi\left(x_i'\beta/\sigma\right)\beta_k. \tag{19.31}$$

Notice that this derivative depends on $x_i'\beta$. For low values of $x_i'\beta$, the derivative is close to zero (since $\Phi\left(x_i'\beta/\sigma\right) \approx 0$). In contrast, for high values of $x_i'\beta$, the derivative is close to β_k.

19.4 Heckit: Sample Selection Model

Recall that *in a Tobit model*, $x_i'\beta + \varepsilon_i$ decide *both* the probability of observing y_i^* and its value. “Heckit” models relax that.

A *sample selection model* is a two equation model

$$w_i^* = x_{1i}'\beta_1 + \varepsilon_{1i} \tag{19.32}$$

$$h_i^* = x_{2i}'\beta_2 + \varepsilon_{2i}. \tag{19.33}$$

For instance, w_i^* could be individual productivity and h_i^* could be labour supply, and x_{1i} and x_{2i} could contain information about education, age, etc. In this special case where $\text{Corr}(h_i^*, w_i^*) = 1$, then we are back in standard Tobit model.

It is typical to assume that the residuals in the two equations could be correlated

$$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix}\right). \tag{19.34}$$

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 but is likely to affect both productivity and the labour supply choice.

The data on w_i^* only observed for people who work (their hourly wage), and h_i^* is only observed as 0/1 (doesn't work/works)

$$\text{Data: } \begin{cases} w_i = w_i^*, h_i = 1 & \text{if } h_i^* > 0 \\ w_i \text{ not observed}, h_i = 0 & \text{otherwise} \end{cases} \tag{19.35}$$

To understand the properties of this model, notice that the expected value of w_i, conditional on $h_i = 1$, is

$$\begin{aligned} \text{E}(w_i|h_i = 1) &= x_{1i}'\beta_1 + \underbrace{\text{E}(\varepsilon_{1i}|h_i = 1)}_{\text{E}\left(\varepsilon_{1i}|\varepsilon_{2i} > -x_{2i}'\beta_2\right)} \\ &= x_{1i}'\beta_1 + \text{E}\left(\varepsilon_{1i}|\varepsilon_{2i} > -x_{2i}'\beta_2\right) \\ &= x_{1i}'\beta_1 + \sigma_{12}\lambda_i, \text{ where } \lambda_i = \frac{\phi(x_{2i}'\beta_2)}{\Phi(x_{2i}'\beta_2)}, \end{aligned} \tag{19.36}$$

where $\phi()$ and $\Phi()$ are the standard normal pdf and cdf (λ_i is called the inverse Mill's ratio or Heckman's lambda). Showing this is straightforward, but a bit tedious. The point of (19.36) that the correlation of the residuals in the two equations (19.32)–(19.33) is crucial. In fact, when $\sigma_{12} = 0$, then we can estimate (19.32) with OLS. Otherwise, it is biased (and inconsistent).

Another way to see this: for the observable data (when $h_i = 1$)

$$w_i = x'_{1i}\beta_1 + \varepsilon_{1i}$$

and the issue is: $\mathrm{E}(x_{1i}\varepsilon_{1i}) = 0$ for this data? To keep it simple, suppose x_{2i} includes just a constant: w_i observed only when $\varepsilon_{2i} > 0$. If $\mathrm{Corr}(\varepsilon_{1i}, \varepsilon_{2i}) > 0$, 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 $\sigma_{12} \neq 0$? Must think about the economics. 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 $\mathrm{Cov}(x_{1i}, \lambda_i) = 0$ (although $\sigma_{12} \neq 0$)? Well, then OLS on

$$w_i = x'_{1i}\beta + \varepsilon_{1i}$$

is consistent (recall the case of uncorrelated regressors: can then estimate one slope coefficient at a time). The conclusion is that the bias of OLS comes from $\mathrm{Cov}(x_{1i}, x_{2i}) \neq 0$ since then $\mathrm{Cov}(x_{1i}, \lambda_i) \neq 0$: extreme case $x_{1i} = x_{2i}$.

19.4.1 Estimation

Use MLE or Heckman's 2-step approach, which is as follows:

1. Estimate (19.33) with Probit method (recall $h_i = 0$ or 1). We are then estimating β_2 in $\Pr(h_i = 1) = F(x'_{2i}\beta_2)$. Extract $x'_{2i}\hat{\beta}_2$ and create $\hat{\lambda}_i$ as in (19.36).

2. Estimate (β_1 and σ_{12}) with LS

$$w_i = x'_{1i}\beta_1 + \sigma_{12}\hat{\lambda}_i + \eta_i \tag{19.37}$$

   on the data where w_i is observed (and not artificial set to zero or some other value).

Properties: consistent, may need to adjust standard errors (unless you test under the null hypothesis that $\sigma_{12} = 0$).

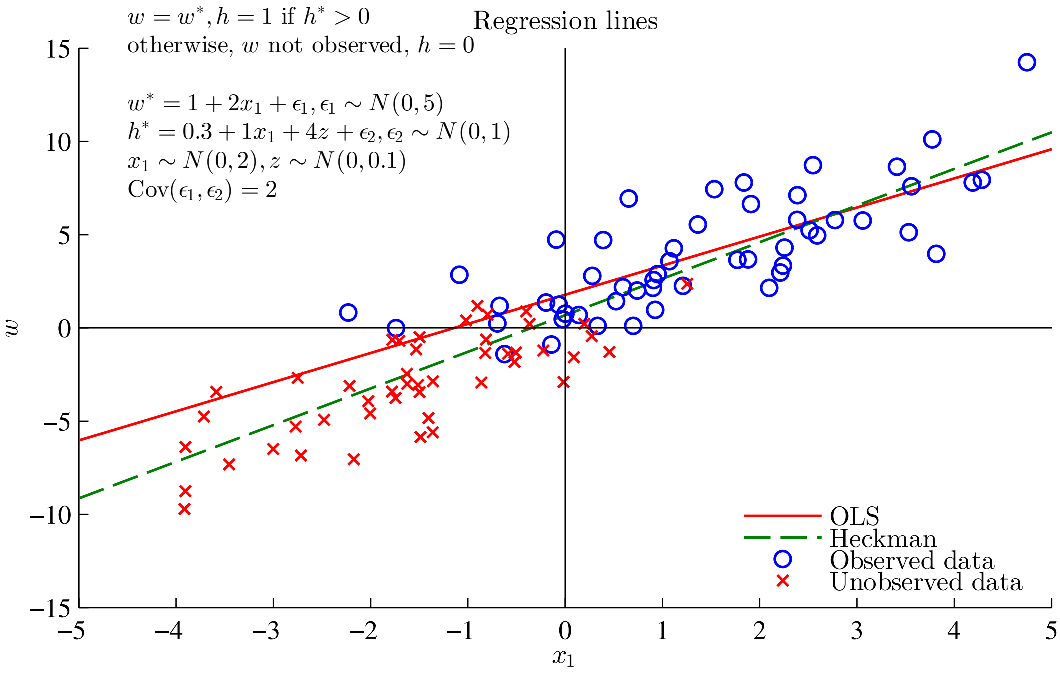


Figure 19.9: Sample selection model

	LS	Heckman
const	−0.40	1.21
	(−2.10)	(2.11)
educ	0.11	0.05
	7.73	2.23
exper	0.02	0.02
	3.90	4.18
lambda		−1.56
		−2.98
R2	0.15	
Nobs	753.00	

Table 19.2: OLS and Heckman estimation of log wages, married women, PSID 1975. Example of a Heckman model, Hill et al (2008), Table 16.8. Numbers in parentheses are t-stats.

Bibliography

Verbeek, M., 2012, *A guide to modern econometrics*, Wiley, 4th edn.

	Tobit
const	1.05
	(2.20)
age	−0.01
	(−2.98)
educ	0.05
	(3.54)
kids	−0.19
	(−2.45)
mtr	−0.96
	(−2.33)

Table 19.3: Tobit estimation of labour market participation, hours>0. 1st step of Heckman estimation. Example of a Heckman model, Hill et al (2008), Table 16.8. Numbers in parentheses are t-stats.