Lecture Notes in Empirical Finance (MSc, PhD)

Paul Söderlind[1]

19 April 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: EmpFinPhDAll.TeX.

Contents

1 Econometrics Cheat Sheet

Sections denoted by a star (*) is not required reading.

Reference: Cochrane (2005) 11 and 14; Singleton (2006) 2–4; DeMiguel, Garlappi, and Uppal (2009)

1.1 GMM

1.1.1 The Basic GMM

In general, the $q \times 1$ sample moment conditions in GMM are written

$$\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} g_t(\beta) = \mathbf{0}_{q\times 1}, \tag{1.1}$$

where $\bar{g}(\beta)$ is short hand notation for the sample average and where the value of the moment conditions clearly depend on the parameter vector. We let β_0 denote the true value of the $k \times 1$ parameter vector. The GMM estimator is

$$\hat{\beta}_{k\times 1} = \arg\min \bar{g}(\beta)' W \bar{g}(\beta), \tag{1.2}$$

where W is some symmetric positive definite $q \times q$ weighting matrix.

Example 1.1 *(Moment condition for a mean) To estimated the mean of x_t, use the following moment condition*

$$\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t - \mu = 0.$$

Example 1.2 *(Moments conditions for IV/2SLS/OLS) Consider the linear model $y_t = x_t'\beta_0 + u_t$, where x_t and β are $k \times 1$ vectors. Let z_t be a $q \times 1$ vector, with $q \geq k$. The sample moment conditions are*

$$\bar{g}\,(\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta) = \mathbf{0}_{q\times 1}$$

Let $q = k$ to get IV; let $z_t = x_t$ to get LS.

Example 1.3 *(Moments conditions for MLE) The maximum likelihood estimator maximizes the log likelihood function,* $\Sigma_{t=1}^{T} \ln L\left(w_t; \beta\right) / T$*, with the* K *first order conditions (one for each element in* β*)*

$$\bar{g}\left(\beta\right) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \ln L\left(w_t; \beta\right)}{\partial \beta} = \mathbf{0}_{K\times 1}$$

GMM estimators are typically asymptotically normally distributed, with a covariance matrix that depends on the covariance matrix of the moment conditions (evaluated at the true parameter values) and the possibly non-linear transformation of the moment conditions that defines the estimator. Let S_0 be the $(q \times q)$ covariance matrix of $\sqrt{T}\bar{g}(\beta_0)$ (evaluated at the true parameter values)

$$S_0 = \lim_{T\to\infty} \mathrm{Cov}\left[\sqrt{T}\bar{g}(\beta_0)\right] = \sum_{s=-\infty}^{\infty} \mathrm{Cov}\left[g_t(\beta_0), g_{t-s}(\beta_0)\right], \tag{1.3}$$

where $\mathrm{Cov}(x, y)$ is a matrix of covariances: element ij is $\mathrm{Cov}(x_i, y_j)$. value).

In addition, let D_0 be the $(q \times k)$ probability limit of the gradient (Jacobian) of the sample moment conditions with respect to the parameters (also evaluated at the true parameters)

$$D_0 = \mathrm{plim}\, \frac{\partial \bar{g}(\beta_0)}{\partial \beta'}. \tag{1.4}$$

Remark 1.4 *(Jacobian) The Jacobian is of the following format*

$$\frac{\partial \bar{g}(\beta_0)}{\partial \beta'} = \begin{bmatrix} \frac{\partial \bar{g}_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_1(\beta)}{\partial \beta_k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_q(\beta)}{\partial \beta_k} \end{bmatrix} \textit{(evaluated at } \beta_0\textit{).}$$

We then have that

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V) \text{ if } W = S_0^{-1}, \text{ where}$$
$$V = \left(D_0' S_0^{-1} D_0\right)^{-1}, \tag{1.5}$$

which assumes that we have used S_0^{-1} as the weighting matrix. This gives the most efficient GMM estimator—for a given set of moment conditions. The choice of the weighting

matrix is irrelevant if the model is exactly identified (as many moment conditions as parameters), so (1.5) can be applied to this case (even if we did not specify any weighting matrix at all). In practice, the gradient D_0 is approximated by using the point estimates and the available sample of data. The Newey-West estimator is commonly used to estimate the covariance matrix S_0. To implement $W = S_0^{-1}$, an iterative procedure is often used: start with $W = 1$, estimate the parameters, estimate $\hat{S}_0$, then (in a second step) use $W = \hat{S}_0^{-1}$ and reestimate. In most cases this iteration is stopped at this stage, but other researchers choose to continue iterating until the point estimates converge.

Example 1.5 *(Estimating a mean) For the moment condition in Example 1.1, assuming iid data gives*

$$S_0 = \text{Var}(x_t) = \sigma^2.$$

In addition,

$$D_0 = \frac{\partial \bar{g}(\mu_0)}{\partial \mu} = -1,$$

which in this case is just a constant (and does not need to be evaluated at true parameter). Combining gives

$$\sqrt{T}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, \sigma^2), \text{ so } \text{“}\hat{\mu} \sim N(\mu_0, \sigma^2/T).\text{”}$$

Remark 1.6 *(IV/2SLS/OLS) Let* $u_t = y_t - x_t'\beta$

$$S_0 = \text{Cov}\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} z_t u_t\right]$$
$$D_0 = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right) = -\Sigma_{zx}.$$

Under the Gauss-Markov assumptions S_0 *for OLS* ($z_t = x_t$) *can be simplified to*

$$S_0 = \sigma^2 \frac{1}{T}\sum_{t=1}^{T} x_t x_t' = \sigma^2 \Sigma_{xx},$$

so combining gives

$$V = \left[\Sigma_{xx}\left(\sigma^2 \Sigma_{xx}\right)^{-1}\Sigma_{xx}\right]^{-1} = \sigma^2 \Sigma_{xx}^{-1}.$$

To test if the moment conditions are satisfied, we notice that under the hull hypothesis (that the model is correctly specified)

$$\sqrt{T}\bar{g}\left(\beta_0\right) \xrightarrow{d} N\left(\mathbf{0}_{q\times 1}, S_0\right), \tag{1.6}$$

where q is the number of moment conditions. Since $\hat{\beta}$ chosen is such a way that k (number of parameters) linear combinations of the first order conditions always (in every sample) are zero, we get that there are effectively only $q-k$ non-degenerate random variables. We can therefore test the hypothesis that $\bar{g}\left(\beta_0\right)=0$ on the the "J test"

$$T\bar{g}(\hat{\beta})' S_0^{-1}\bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \text{ if } W = S_0^{-1}. \tag{1.7}$$

The left hand side equals T times of value of the loss function in (1.2) evaluated at the point estimates With no overidentifying restrictions (as many moment conditions as parameters) there are, of course, no restrictions to test. Indeed, the loss function value is then always zero at the point estimates.

1.1.2 GMM with a Suboptimal Weighting Matrix

It can be shown that if we use another weighting matrix than $W = S_0^{-1}$, then the variance-covariance matrix in (1.5) should be changed to

$$V_2 = \left(D_0' W D_0\right)^{-1} D_0' W S_0 W' D_0 \left(D_0' W D_0\right)^{-1}. \tag{1.8}$$

Similarly, the test of overidentifying restrictions becomes

$$T\bar{g}(\hat{\beta})' \Psi_2^{+} \bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \tag{1.9}$$

where Ψ_2^{+} is a generalized inverse of

$$\Psi_2 = \left[I_q - D_0\left(D_0' W D_0\right)^{-1} D_0' W\right] S_0 \left[I_q - D_0\left(D_0' W D_0\right)^{-1} D_0' W\right]'. \tag{1.10}$$

Remark 1.7 *(Quadratic form with degenerate covariance matrix) If the $n \times 1$ vector $X \sim N(0, \Sigma)$, where Σ has rank $r \leq n$ then $Y = X'\Sigma^{+}X \sim \chi^2_r$ where Σ^{+} is the pseudo inverse of Σ.*

Example 1.8 *(Pseudo inverse of a square matrix) For the matrix*

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}, \text{ we have } A^+ = \begin{bmatrix} 0.02 & 0.06 \\ 0.04 & 0.12 \end{bmatrix}.$$

1.1.3 GMM without a Loss Function

Suppose we sidestep the whole optimization issue and instead specify k linear combinations (as many as there are parameters) of the q moment conditions directly

$$\mathbf{0}_{k\times 1} = \underbrace{A}_{k\times q}\underbrace{\bar{g}(\hat{\beta})}_{q\times 1}, \tag{1.11}$$

where the matrix A is chosen by the researcher.

It is straightforward to show that the variance-covariance matrix in (1.5) should be changed to

$$V_3 = (A_0 D_0)^{-1} A_0 S_0 A_0' [(A_0 D_0)^{-1}]', \tag{1.12}$$

where A_0 is the probability limit of A (if it is random). Similarly, in the test of overidentifying restrictions (1.9), we should replace Ψ_2 by

$$\Psi_3 = [I_q - D_0 (A_0 D_0)^{-1} A_0] S_0 [I_q - D_0 (A_0 D_0)^{-1} A_0]'. \tag{1.13}$$

1.1.4 GMM Example 1: Estimate the Variance

Suppose x_t has a zero mean. To estimate the mean we specify the moment condition

$$g_t = x_t^2 - \sigma^2. \tag{1.14}$$

To derive the asymptotic distribution, we take look at the simple case when x_t is iid $N(0, \sigma^2)$ This gives $S_0 = \text{Var}(g_t)$, because of the iid assumption. We can simplify this further as

$$\begin{aligned} S_0 &= \text{E}(x_t^2 - \sigma^2)^2 \\ &= \text{E}(x_t^4 + \sigma^4 - 2x_t^2\sigma^2) = \text{E}\, x_t^4 - \sigma^4 \\ &= 2\sigma^4, \end{aligned} \tag{1.15}$$

where the second line is just algebra and the third line follows from the properties of normally distributed variables ($\mathrm{E}\, x_t^4 = 3\sigma^4$).

Note that the Jacobian is

$$D_0 = -1, \tag{1.16}$$

so the GMM formula says

$$\sqrt{T}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4). \tag{1.17}$$

1.1.5 GMM Example 2: The Means and Second Moments of Returns

Let R_t be a vector of net returns of N assets. We want to estimate the mean vector and the covariance matrix. The moment conditions for the mean vector are

$$\mathrm{E}\, R_t - \mu = \mathbf{0}_{N\times 1}, \tag{1.18}$$

and the moment conditions for the unique elements of the second moment matrix are

$$\mathrm{E}\,\mathrm{vech}(R_t R_t') - \mathrm{vech}(\Gamma) = \mathbf{0}_{N(N+1)/2\times 1}. \tag{1.19}$$

Remark 1.9 *(The vech operator) vech(A) where A is $m \times m$ gives an $m(m+1)/2 \times 1$ vector with the elements on and below the principal diagonal A stacked on top of each other (column wise). For instance, vech* $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}$.

Stack (1.18) and (1.19) and substitute the sample mean for the population expectation to get the GMM estimator

$$\frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} R_t \\ \mathrm{vech}(R_t R_t') \end{bmatrix} - \begin{bmatrix} \hat{\mu} \\ \mathrm{vech}(\hat{\Gamma}) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ \mathbf{0}_{N(N+1)/2\times 1} \end{bmatrix}. \tag{1.20}$$

In this case, $D_0 = -I$, so the covariance matrix of the parameter vector $(\hat{\mu}, \mathrm{vech}(\hat{\Gamma}))$ is just S_0 (defined in (1.3)), which is straightforward to estimate.

1.1.6 GMM Example 3: Non-Linear Least Squares

Consider the non-linear regression

$$y_t = F(x_t; \beta_0) + \varepsilon_t, \tag{1.21}$$

where $F(x_t; \beta_0)$ is a potentially non-linear equation of the regressors x_t, with a $k \times 1$ vector of parameters β_0. The non-linear least squares (NLS) approach is minimize the sum of squared residuals, that is, to solve

$$\hat{\beta} = \arg\min \textstyle\sum_{t=1}^{T} [y_t - F(x_t; \beta)]^2. \tag{1.22}$$

To express this as a GMM problem, use the first order conditions for (1.22) as moment conditions

$$\bar{g}(\beta) = -\frac{1}{T} \textstyle\sum_{t=1}^{T} \dfrac{\partial F(x_t; \beta)}{\partial \beta} [y_t - F(x_t; \beta)]. \tag{1.23}$$

The model is then exactly identified so the point estimates are found by setting all moment conditions to zero, $\bar{g}(\beta) = \mathbf{0}_{k\times 1}$. The distribution of the parameter estimates is thus as in (1.5). As usual, $S_0 = \text{Cov}[\sqrt{T}\bar{g}(\beta_0)]$, while the Jacobian is

$$\begin{aligned} D_0 &= \text{plim} \frac{\partial \bar{g}(\beta_0)}{\partial \beta'} \\ &= \text{plim} \frac{1}{T} \textstyle\sum_{t=1}^{T} \dfrac{\partial F(x_t; \beta)}{\partial \beta} \dfrac{\partial F(x_t; \beta)}{\partial \beta'} - \text{plim} \frac{1}{T} \textstyle\sum_{t=1}^{T} [y_t - F(x_t; \beta)] \dfrac{\partial^2 F(x_t; \beta)}{\partial \beta \partial \beta'}. \end{aligned} \tag{1.24}$$

Example 1.10 *(The derivatives with two parameters) With $\beta = [\beta_1, \beta_2]'$ we have*

$$\frac{\partial F(x_t; \beta)}{\partial \beta} = \begin{bmatrix} \partial F(x_t; \beta)/\partial \beta_1 \\ \partial F(x_t; \beta)/\partial \beta_2 \end{bmatrix}, \frac{\partial F(x_t; \beta)}{\partial \beta'} = \begin{bmatrix} \partial F(x_t; \beta)/\partial \beta_1 & \partial F(x_t; \beta)/\partial \beta_2 \end{bmatrix},$$

so the outer product of the gradient (first term) in (1.24) is a 2×2 matrix. Similarly, the matrix with the second derivatives (the Hessian) is also a 2×2 matrix

$$\frac{\partial^2 F(x_t; \beta)}{\partial \beta \partial \beta'} = \begin{bmatrix} \frac{\partial^2 F(x_t; \beta)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 F(x_t; \beta)}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 F(x_t; \beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 F(x_t; \beta)}{\partial \beta_2 \partial \beta_2} \end{bmatrix}.$$

1.2 MLE

1.2.1 The Basic MLE

Let L be the likelihood function of a sample, defined as the joint density of the sample

$$L = \text{pdf}(x_1, x_2, \ldots x_T; \theta) \tag{1.25}$$
$$= L_1 L_2 \ldots L_T, \tag{1.26}$$

where θ are the parameters of the density function. In the second line, we define the likelihood function as the product of the likelihood contributions of the different observations. For notational convenience, their dependence of the data and the parameters are suppressed.

The idea of MLE is to pick parameters to make the likelihood (or its log) value as large as possible

$$\hat{\theta} = \arg\max \ln L. \tag{1.27}$$

MLE is typically asymptotically normally distributed

$$\sqrt{N}(\hat{\theta} - \theta) \to^d N(0, V), \text{ where } V = I(\theta)^{-1} \text{ with} \tag{1.28}$$
$$I(\theta) = -\text{E}\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}/T \text{ or}$$
$$= -\text{E}\frac{\partial^2 \ln L_t}{\partial\theta\partial\theta'},$$

where $I(\theta)$ is the "information matrix." In the second line, the derivative is of the whole log likelihood function (1.25), while in the third line the derivative is of the likelihood contribution of observation t.

Alternatively, we can use the outer product of the gradients to calculate the information matrix as

$$J(\theta) = \text{E}\left[\frac{\partial \ln L_t}{\partial\theta}\frac{\partial \ln L_t}{\partial\theta'}\right]. \tag{1.29}$$

A key strength of MLE is that it is asymptotically efficient, that is, any linear combination of the parameters will have a smaller asymptotic variance than if we had used any other estimation method.

1.2.2 QMLE

A MLE based on the wrong likelihood function (distribution) may still be useful. Suppose we use the likelihood function L, so the estimator is defined by

$$\frac{\partial \ln L}{\partial \theta} = \mathbf{0}. \tag{1.30}$$

If this is the wrong likelihood function, but the expected value (under the true distribution) of $\partial \ln L/\partial\theta$ is indeed zero (at the true parameter values), then we can think of (1.30) as a set of GMM moment conditions—and the usual GMM results apply. The result is that this quasi-MLE (or pseudo-MLE) has the same sort of distribution as in (1.28), but with the variance-covariance matrix

$$V = I(\theta)^{-1} J(\theta) I(\theta)^{-1} \tag{1.31}$$

Example 1.11 *(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.*

1.2.3 MLE Example: Estimate the Variance

Suppose x_t is iid $N(0, \sigma^2)$. The pdf of x_t is

$$\text{pdf}(x_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{x_t^2}{\sigma^2}\right). \tag{1.32}$$

Since x_t and x_{t+1} are independent,

$$\begin{aligned} L &= \text{pdf}(x_1) \times \text{pdf}(x_2) \times ... \times \text{pdf}(x_T) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left(-\frac{1}{2}\frac{\sum_{t=1}^{T} x_t^2}{\sigma^2}\right), \text{ so} \end{aligned} \tag{1.33}$$

$$\ln L = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T} x_t^2. \tag{1.34}$$

The first order condition for optimum is

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{1}{2(\sigma^2)^2}\sum_{t=1}^{T} x_t^2 = 0 \text{ so}$$
$$\hat{\sigma}^2 = \sum_{t=1}^{T} x_t^2/T. \tag{1.35}$$

Differentiate the log likelihood once again to get

$$\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} = \frac{T}{2}\frac{1}{\sigma^4} - \frac{1}{(\sigma^2)^3}\sum_{t=1}^{T} x_t^2, \text{ so} \tag{1.36}$$
$$\mathrm{E}\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} = \frac{T}{2}\frac{1}{\sigma^4} - \frac{T}{(\sigma^2)^3}\sigma^2 = -\frac{T}{2\sigma^4} \tag{1.37}$$

The information matrix is therefore

$$I(\theta) = -\mathrm{E}\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2}/T = \frac{1}{2\sigma^4}, \tag{1.38}$$

so we have

$$\sqrt{T}(\hat{\sigma}^2 - \sigma^2) \to^d N(0, 2\sigma^4). \tag{1.39}$$

1.3 The Variance of a Sample Mean: The Newey-West Estimator

Many estimators (including GMM) are based on some sort of sample average. Unless we are sure that the series in the average is iid, we need an estimator of the variance (of the sample average) that takes serial correlation into account. The Newey-West estimator is probably the most popular.

Example 1.12 *(Variance of sample average) The variance of* $(x_1 + x_2)/2$ *is* $\mathrm{Var}(x_1)/4 + \mathrm{Var}(x_2)/4 + \mathrm{Cov}(x_1, x_2)/2$. *If* $\mathrm{Var}(x_i) = \sigma^2$ *for all* i, *then this is* $\sigma^2/2 + \mathrm{Cov}(x_1, x_2)/2$. *If there is no autocorrelation, then we have the traditional result,* $\mathrm{Var}(\bar{x}) = \sigma^2/T$.

Example 1.13 *(*x_t *is a scalar iid process.) When* x_t *is a scalar iid process, then*

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) &= \frac{1}{T^2}\sum_{t=1}^{T} \mathrm{Var}\,(x_t) \text{ (since independently distributed)} \\
&= \frac{1}{T^2} T\,\mathrm{Var}\,(x_t) \text{ (since identically distributed)} \\
&= \frac{1}{T}\mathrm{Var}\,(x_t)\,.
\end{aligned}$$

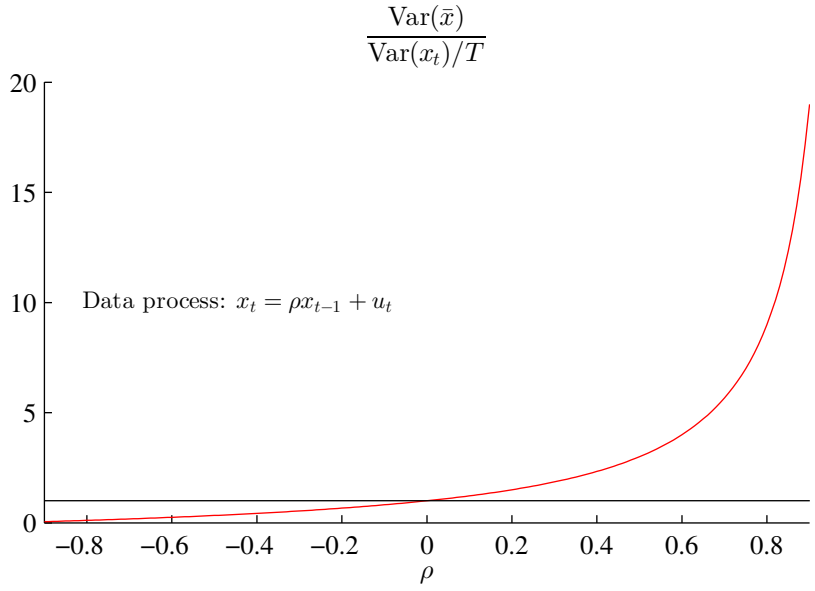


Figure 1.1: Variance of sample mean of an AR(1) series

This is the classical iid case. Clearly, $\lim_{T\Rightarrow\infty} \text{Var}(\bar{x}) = 0$. *By multiplying both sides by* T *we instead get* $\text{Var}(\sqrt{T}\bar{x}) = \text{Var}(x_t)$.

The Newey-West estimator of the variance-covariance matrix of the sample mean, $\bar{g}$, of $K \times 1$ vector g_t is

$$\widehat{\text{Cov}}\left(\sqrt{T}\bar{g}\right) = \sum_{s=-n}^{n}\left(1 - \frac{|s|}{n+1}\right)\widehat{\text{Cov}}\left(g_t, g_{t-s}\right) \tag{1.40}$$

$$= \widehat{\text{Cov}}\left(g_t, g_t\right) + \sum_{s=1}^{n}\left(1 - \frac{s}{n+1}\right)\left(\widehat{\text{Cov}}\left(g_t, g_{t-s}\right) + \widehat{\text{Cov}}\left(g_t, g_{t-s}\right)'\right) \tag{1.41}$$

where n is a finite "bandwidth" parameter.

Example 1.14 *(Newey-West estimator) With* $n = 1$ *in (1.40) the Newey-West estimator becomes*

$$\widehat{\text{Cov}}\left(\sqrt{T}\bar{g}\right) = \widehat{\text{Cov}}\left(g_t, g_t\right) + \frac{1}{2}\left(\widehat{\text{Cov}}\left(g_t, g_{t-1}\right) + \widehat{\text{Cov}}\left(g_t, g_{t-1}\right)'\right).$$

Example 1.15 *(Variance of sample mean of AR(1).) Let* $x_t = \rho x_t + u_t$*, where* $\text{Var}(u_t) = \sigma^2$*. Let* $R(s)$ *denote the* s*th autocovariance and notice that* $R(s) = \rho^{|s|}\sigma^2/\left(1-\rho^2\right)$*, so*

$$\text{Var}\left(\sqrt{T}\bar{x}\right) = \sum_{s=-\infty}^{\infty} R(s) = \frac{\sigma^2}{1-\rho^2}\sum_{s=-\infty}^{\infty}\rho^{|s|} = \frac{\sigma^2}{1-\rho^2}\frac{1+\rho}{1-\rho},$$

which is increasing in ρ *(provided* $|\rho| < 1$*, as required for stationarity). The variance of* $\sqrt{T}\bar{x}$ *is much larger for* ρ *close to one than for* ρ *close to zero: the high autocorrelation create long swings, so the mean cannot be estimated with good precision in a small sample. If we disregard all autocovariances, then we would conclude that the variance of* $\sqrt{T}\bar{x}$ *is* $\sigma^2/\left(1-\rho^2\right)$*, that is, the variance of* x_t*. This is much smaller (larger) than the true value when* $\rho > 0$ *(*$\rho < 0$*). For instance, with* $\rho = 0.9$*, it is 19 times too small. See Figure 1.1 for an illustration. Notice that* $T\,\text{Var}(\bar{x})\,/\,\text{Var}(x_t) = \text{Var}(\bar{x})\,/[\text{Var}(x_t)/T]$*, so the ratio shows the relation between the true variance of* $\bar{x}$ *and the classical estimator of it (based of the iid assumption).*

1.4 Testing (Linear) Joint Hypotheses

Consider an estimator $\hat{\beta}_{k\times 1}$ which satisfies

$$\sqrt{T}(\hat{\beta}-\beta_0) \xrightarrow{d} N\left(0, V_{k\times k}\right), \tag{1.42}$$

and suppose we want the asymptotic distribution of a linear transformation of β

$$\gamma_{q\times 1} = R\beta - a. \tag{1.43}$$

Under that null hypothesis (that $\gamma = \mathbf{0}$)

$$\begin{aligned}\sqrt{T}(R\beta-a) &\xrightarrow{d} N\left(0, \Lambda_{q\times q}\right), \text{ where}\\ \Lambda &= RVR'.\end{aligned} \tag{1.44}$$

Example 1.16 *(Testing 2 slope coefficients) Suppose we have estimated a model with three coefficients and the null hypothesis is*

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

We can write this as

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

The test of the joint hypothesis is based on

$$(R\beta - a)\Lambda^{-1}(R\beta - a)' \xrightarrow{d} \chi^2_q. \tag{1.45}$$

1.5 Testing (Nonlinear) Joint Hypotheses: The Delta Method

Consider an estimator $\hat{\beta}_{k\times 1}$ which satisfies

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left(0, V_{k\times k}\right), \tag{1.46}$$

and suppose we want the asymptotic distribution of a transformation of β

$$\gamma_{q\times 1} = f\left(\beta\right), \tag{1.47}$$

where $f\left(.\right)$ has continuous first derivatives. The result is

$$\sqrt{T}[f(\hat{\beta}) - f\left(\beta_0\right)] \xrightarrow{d} N\left(0, \Lambda_{q\times q}\right), \text{ where}$$

$$\Lambda = \frac{\partial f\left(\beta_0\right)}{\partial \beta'} V \frac{\partial f\left(\beta_0\right)'}{\partial \beta}, \text{ where } \frac{\partial f\left(\beta\right)}{\partial \beta'} = \begin{bmatrix} \frac{\partial f_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial f_1(\beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial f_q(\beta)}{\partial \beta_k} \end{bmatrix}_{q\times k} \tag{1.48}$$

The derivatives can sometimes be found analytically, otherwise numerical differentiation can be used. Now, a test can be done as in the same way as in (1.45).

Example 1.17 *(Quadratic function) Let $f(\beta) = \beta^2$ where β is a scalar. Then $\partial f\left(\beta\right)/\partial \beta = 2\beta$, so $\Lambda = 4\beta^2 V$, where $V = \text{Var}(\sqrt{T}\hat{\beta})$.*

Example 1.18 *(Testing a Sharpe ratio) Stack the mean ($\mu = \text{E}\, x_t$) and second moment ($\mu_2 = \text{E}\, x_t^2$) as $\beta = [\mu, \mu_2]'$. The Sharpe ratio is calculated as a function of β*

$$\frac{\text{E}(x)}{\sigma(x)} = f(\beta) = \frac{\mu}{(\mu_2 - \mu^2)^{1/2}}, \text{ so } \frac{\partial f(\beta)}{\partial \beta'} = \begin{bmatrix} \frac{\mu_2}{(\mu_2 - \mu^2)^{3/2}} & \frac{-\mu}{2(\mu_2 - \mu^2)^{3/2}} \end{bmatrix}.$$

If $\hat{\beta}$ is distributed as in (1.46), then (1.48) is straightforward to apply.

Example 1.19 *(Linear function) When $f(\beta) = R\beta - a$, then the Jacobian is $\frac{\partial f(\beta)}{\partial \beta'} = R$, so $\Lambda = RVR'$, just like in (1.44).*

Example 1.20 *(Testing a correlation of x_t and y_t, $\rho(x_t, y_t)$) For expositional simplicity, assume that both variables have zero means. The variances and the covariance are then be estimated by the moment conditions*

$$\sum_{t=1}^{T} m_t(\beta)/T = \mathbf{0}_{3\times 1} \text{ where } m_t = \begin{bmatrix} x_t^2 - \sigma_{xx} \\ y_t^2 - \sigma_{yy} \\ x_t y_t - \sigma_{xy} \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{bmatrix}.$$

The covariance matrix of these estimators is estimated as usual in GMM, making sure to account for autocorrelation of the data. The correlation is a simple function of these parameters

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

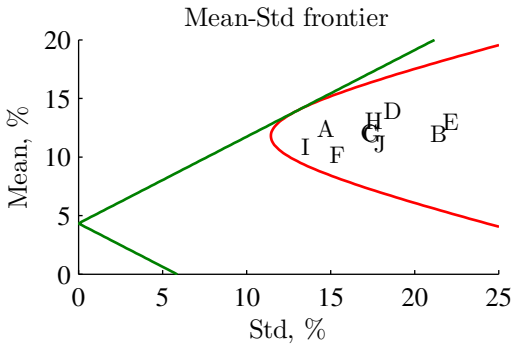
It is then straightforward to apply delta method (1.48).

Remark 1.21 *(Numerical derivatives) These derivatives can typically be very messy to calculate analytically, but numerical approximations often work fine. A very simple code can be structured as follows: let column j of $\partial f(\beta)/\partial \beta'$ be*

$$\begin{bmatrix} \frac{\partial f_1(\beta)}{\partial \beta_j} \\ \vdots \\ \frac{\partial f_q(\beta)}{\partial \beta_j} \end{bmatrix} = \frac{f(\tilde{\beta}) - f(\beta)}{\Delta}, \text{ where } \tilde{\beta} = \beta \text{ except that } \tilde{\beta}_j = \beta_j + \Delta.$$

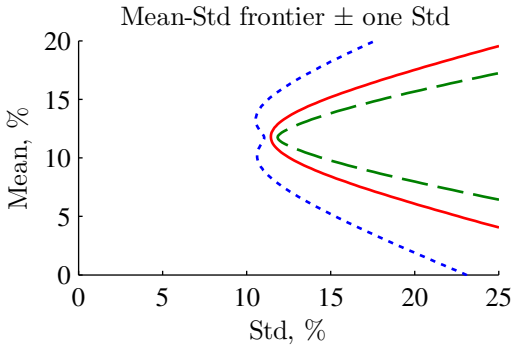
1.5.1 Delta Method Example 1: Confidence Bands around a Mean-Variance Frontier

A point on the mean-variance frontier at a given expected return is a non-linear function of the means and the second moment matrix estimated by 1.20. It is therefore straightforward to apply the delta method to calculate a confidence band around the estimate.



US industry portfolios, 1947:1-2011:12

	Mean	Std
A	12.43	14.18
B	11.95	20.94
C	12.05	16.79
D	13.95	18.16
E	13.02	21.66
F	10.17	14.95
G	11.99	16.91
H	13.06	17.04
I	10.87	13.26
J	11.10	17.58



SR(tangency), SR(EW) and
t-stat of difference
0.74 0.55 1.97

Figure 1.2: Mean-Variance frontier of US industry portfolios from Fama-French. Monthly returns are used in the calculations, but $100\sqrt{12\text{Variance}}$ is plotted against $100 * 12*\text{mean}$.

Figure 1.2 shows some empirical results. The uncertainty is lowest for the minimum variance portfolio (in a normal distribution, the uncertainty about an estimated variance is increasing in the true variance, $\text{Var}(\sqrt{T}\hat{\sigma}^2) = 2\sigma^4$).

Remark 1.22 *(MatLab coding) First, code a function $f(\beta; \mu_p)$ where $\beta = [\mu, \text{vech}(\Gamma)]$ that calculates the minimum standard deviation at a given expected return, μ_p. For this, you may find the duplication matrix (see remark) useful. Second, evaluate it, as well as the Jacobian, at the point estimates. Third, combine with the variance-covariance matrix of $[\hat{\mu}, \text{vech}(\hat{\Gamma})]$ to calculate the variance of the output (the minimum standard deviation). Repeat this for other values of the expected returns, μ_p.*

Remark 1.23 *(Duplication matrix) The duplication matrix D_m is defined such that for*

any symmetric $m \times m$ matrix A we have $\text{vec}(A) = D_m vech(A)$. *For instance,*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{21} \\ a_{22} \end{bmatrix} \text{ or } D_2 vech(A) = \text{vec}(A).$$

The duplication matrix is therefore useful for "inverting" the vech operator—the transformation from $\text{vec}(A)$ *is trivial.*

Remark 1.24 *(MatLab coding) The command reshape(x,m,n) creates an $m \times n$ matrix by putting the first m elements of x in column 1, the next m elements in column 2, etc.*

1.5.2 Delta Method Example 2: Testing the $1/N$ vs the Tangency Portfolio

Reference: DeMiguel, Garlappi, and Uppal (2009)

It has been argued that the (naive) $1/N$ diversification gives a portfolio performance which is not worse than an "optimal" portfolio. One way of testing this is to compare the the Sharpe ratios of the tangency and equally weighted portfolios. Both are functions of the first and second moments of the basic assets, so a delta method approach similar to the one for the MV frontier (see above) can be applied. Notice that this approach should incorporate the way (and hence the associated uncertainty) the first and second moments affect the portfolio weights of the tangency portfolio.

Figure 1.2 shows some empirical results.

Bibliography

Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.

DeMiguel, V., L. Garlappi, and R. Uppal, 2009, "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?," *Review of Financial Studies*, 22, 1915–1953.

Singleton, K. J., 2006, *Empirical dynamic asset pricing*, Princeton University Press.

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.

B Matlab Code

B.1 Autocovariance

Remark B.1 *(MatLab coding) Suppose we have an $T \times K$ matrix g with g_t' in row t. We want to calculate $\widehat{\text{Cov}}(g_t, g_{t-s}) = \Sigma_{t=s+1}^{T}(g_t - \bar{g})(g_{t-s} - \bar{g})'/T$ as in*

```
g_gbar = g - repmat(mean(g),T,1);      %has zero means
Cov_s = g_gbar(s+1:T,:)'*g_gbar(1:T-s,:)/T;
```

B.2 Numerical Derivatives

A simple forward approximation:

```
fb = f(b);
df_db = zeros(q,k);
for j = 1:k;        %loop over columns (parameters)
```

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).

```
  bj          = b;
  bj(j)       = b(j)+Delta;
  df_db(:,j) = (f(bj)- fb)/Delta;
end;
```

2 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 2.1.

2.1 Monte Carlo Simulations

2.1.1 Monte Carlo Simulations in the Simplest Case

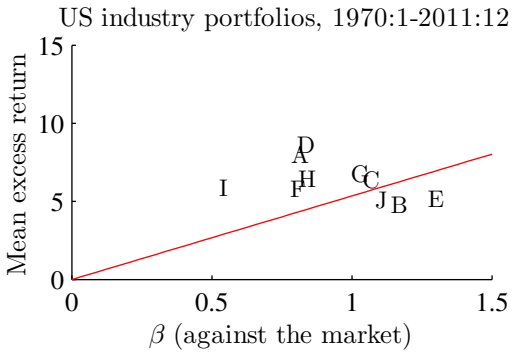
Monte Carlo simulations is essentially a way to generate many artificial (small) samples from a parameterized model and then estimating the statistic on each of those samples. The distribution of the statistic is then used as the small sample distribution of the estimator.

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

$$y_t = x_t'\beta + u_t, \tag{2.1}$$

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,



	alpha	t LS	t NW	t boot
all	NaN	NaN	NaN	NaN
A (NoDur)	3.79	2.76	2.75	2.74
B (Durbl)	-1.33	-0.64	-0.65	-0.64
C (Manuf)	0.84	0.85	0.84	0.84
D (Enrgy)	4.30	1.90	1.91	1.94
E (HiTec)	-1.64	-0.88	-0.88	-0.87
F (Telcm)	1.65	0.94	0.94	0.95
G (Shops)	1.46	0.95	0.96	0.95
H (Hlth)	2.10	1.17	1.19	1.18
I (Utils)	3.03	1.68	1.65	1.63
J (Other)	-0.70	-0.63	-0.62	-0.62

NW uses 1 lag
The bootstrap samples pairs of (y_t, x_t)
3000 simulations

Figure 2.1: CAPM, US industry portfolios, different t-stats

$g(\hat{\beta})$. 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

$$\tilde{y}_t = \tilde{x}_t'\beta + \tilde{u}_t, \tag{2.2}$$

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 2.2–2.3 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 2.1 *(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* $\text{Cov}(\varepsilon_t) = \text{E}\, Pu_t u_t' P' = PIP' = \Sigma$.

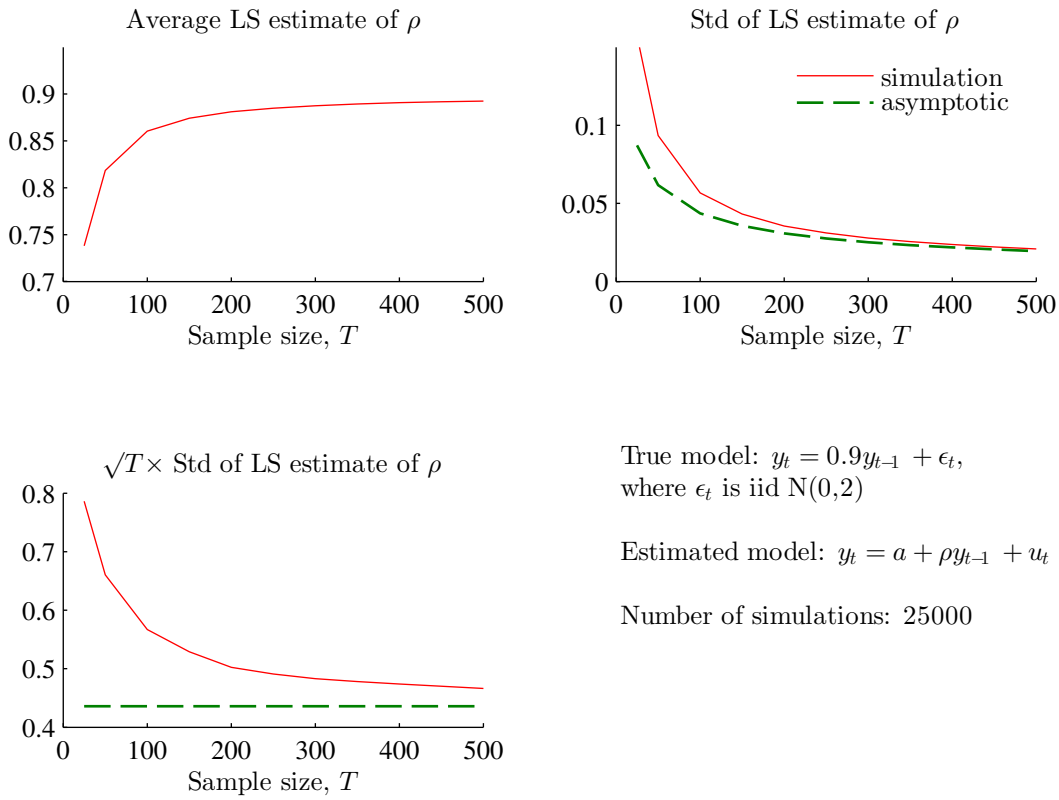


Figure 2.2: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

2.1.2 Monte Carlo Simulations when x_t Includes Lags of y_t

If x_t contains lags of y_t, then we must set up the simulations so that feature is preserved in every artificial sample that we create. For instance, suppose x_t includes y_{t-1} and another vector z_t of variables which are independent of $u_{t\pm s}$ for all s. We can then generate an artificial sample as follows. First, create a sample $\tilde{z}_t$ for $t = 1, \ldots, T$ by some time series model (for instance, a VAR) or by taking the observed sample itself. Second, observation t of $(\tilde{x}_t, \tilde{y}_t)$ is generated as

$$\tilde{x}_t = \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{z}_t \end{bmatrix} \text{ and } \tilde{y}_t = \tilde{x}_t'\beta + \tilde{u}_t \text{ for } t = 1, \ldots, T \tag{2.3}$$

We clearly need the initial value $\tilde{y}_0$ to start up the artificial sample—and then the rest of the sample ($t = 1, 2, ...$) is calculated recursively.

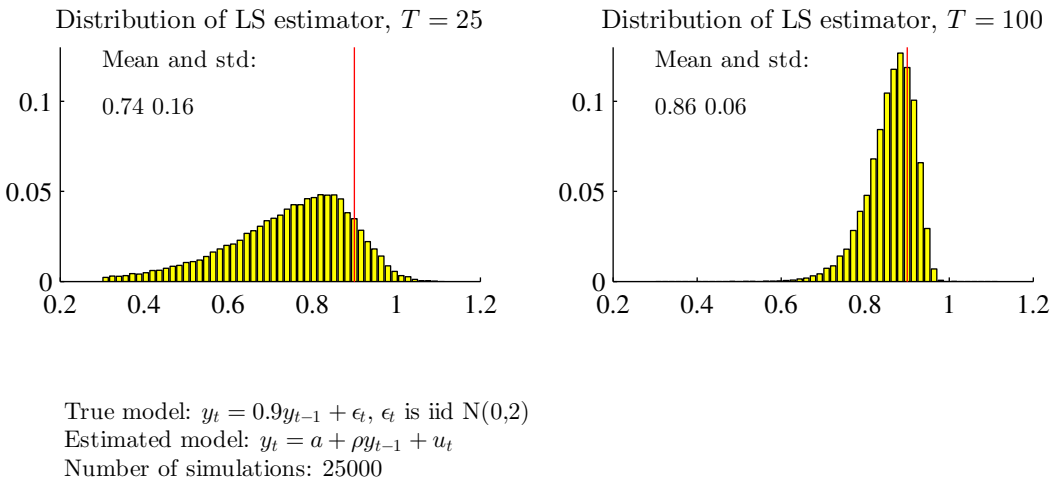


Figure 2.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

For instance, for a *VAR(2) model* (where there is no z_t)

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t, \tag{2.4}$$

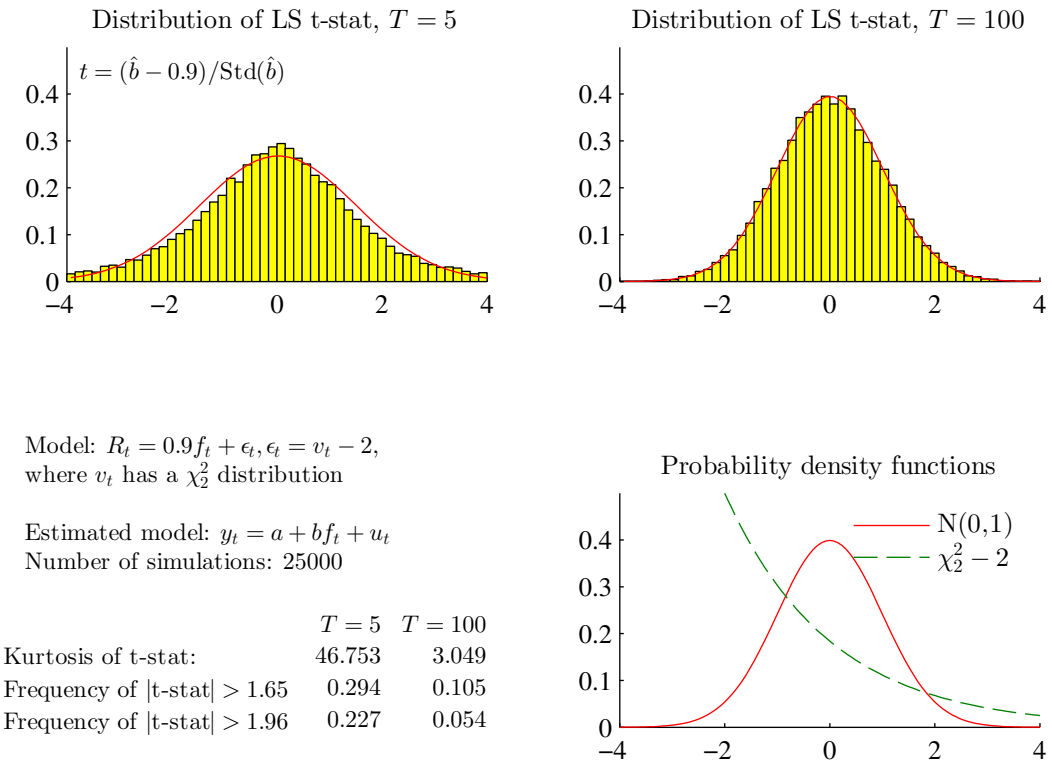
the procedure is straightforward. First, estimate the model on data and record the estimates $(A_1, A_2, \text{Var}(u_t))$. Second, draw a new time series of residuals, $\tilde{u}_t$ for $t = 1, \ldots, T$ and construct an artificial sample recursively (first $t = 1$, then $t = 2$ and so forth) as

$$\tilde{y}_t = A_1 \tilde{y}_{t-1} + A_2 \tilde{y}_{t-2} + \tilde{u}_t. \tag{2.5}$$

(This requires some starting values for y_{-1} and y_0.) Third, re-estimate the model on the the artificial sample, $\tilde{y}_t$ for $t = 1, \ldots, T$.

2.1.3 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 2.4* for an example.



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

Figure 2.4: Results from a Monte Carlo experiment with thick-tailed errors.

Remark 2.2 *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 2.3 *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.*

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))

$$\tilde{u}_t = a_1\tilde{u}_{t-1} + a_2\tilde{u}_{t-2} + \tilde{\varepsilon}_t. \tag{2.6}$$

Alternatively, *heteroskedastic errors* can be generated by, for instance, a GARCH(1,1) model

$$u_t \sim N(0, \sigma_t^2), \text{ where } \sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{2.7}$$

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

$$\varepsilon_t^2 = c' w_t + \eta_t, \tag{2.8}$$

where w_t include the squares and cross product of all the regressors.

2.2 Bootstrapping

2.2.1 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 (2.1), 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

$$\tilde{y}_t = x_t' \beta + \tilde{u}_t, \tag{2.9}$$

   where $\tilde{u}_t$ is drawn (with replacement) from the fitted residual and where β is the point estimate.

Example 2.4 *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 (2.9) works also when y_t is a vector of dependent variables—and will then help retain the cross-sectional correlation of the residuals.

2.2.2 Bootstrapping when x_t Includes Lags of y_t

When x_t contains lagged values of y_t, then we have to modify the approach in (2.9) since $\tilde{u}_t$ can become correlated with x_t. For instance, if x_t includes y_{t-1} and we happen to sample $\tilde{u}_t = u_{t-1}$, then we get a non-zero correlation. The easiest way to handle this is as in the Monte Carlo simulations in (2.3), but where $\tilde{u}_t$ are drawn (with replacement) from the sample of fitted residuals. The same carries over to the *VAR model* in (2.4)–(2.5).

2.2.3 Bootstrapping when Errors Are Heteroskedastic

Suppose now that the errors are heteroskedastic, but serially uncorrelated. If the heteroskedasticity is unrelated to the regressors, then we can still use (2.9).

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 2.5 *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.

2.2.4 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 2.5–2.6 for an illustration.

2.2.5 Other Approaches

There are many other ways to do bootstrapping. For instance, we could sample the regressors and residuals independently of each other and construct an artificial sample of the dependent variable $\tilde{y}_t = \tilde{x}_t'\hat{\beta} + \tilde{u}_t$. This clearly makes sense if the residuals and regressors are independent of each other and errors are iid. In that case, the advantage of this approach is that we do not keep the regressors fixed.

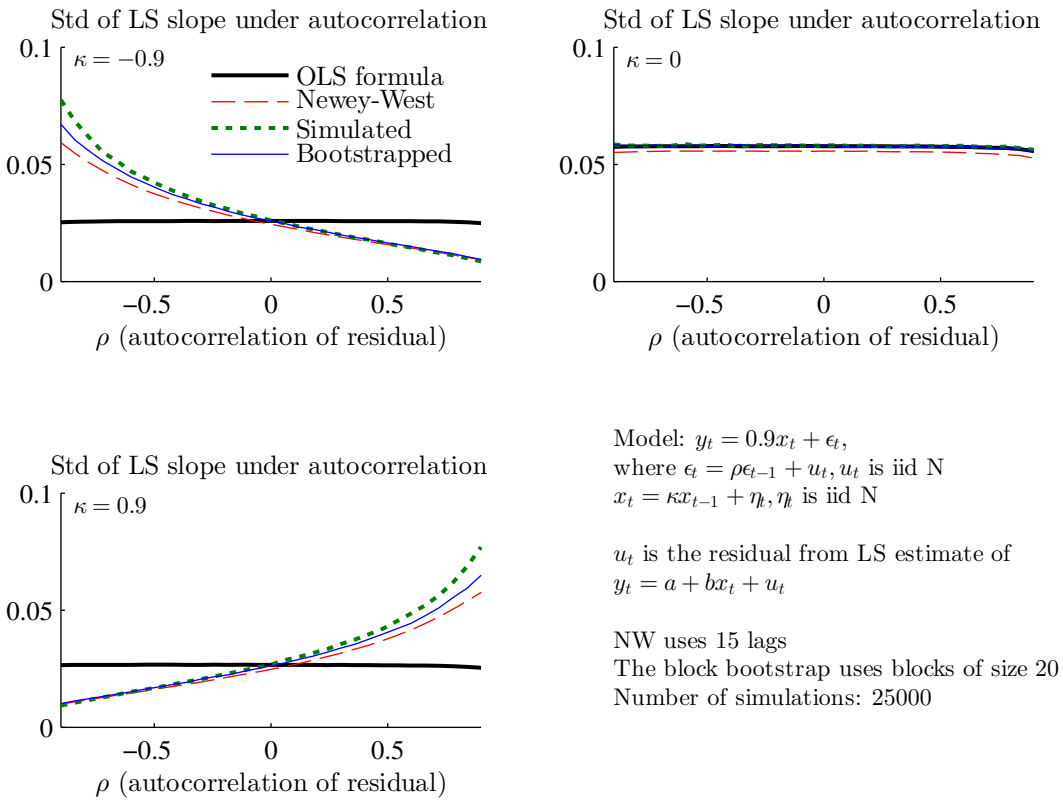


Figure 2.5: Standard error of OLS estimator, autocorrelated errors

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.

Efron, B., and R. J. Tibshirani, 1993, *An introduction to the bootstrap*, Chapman and Hall, New York.

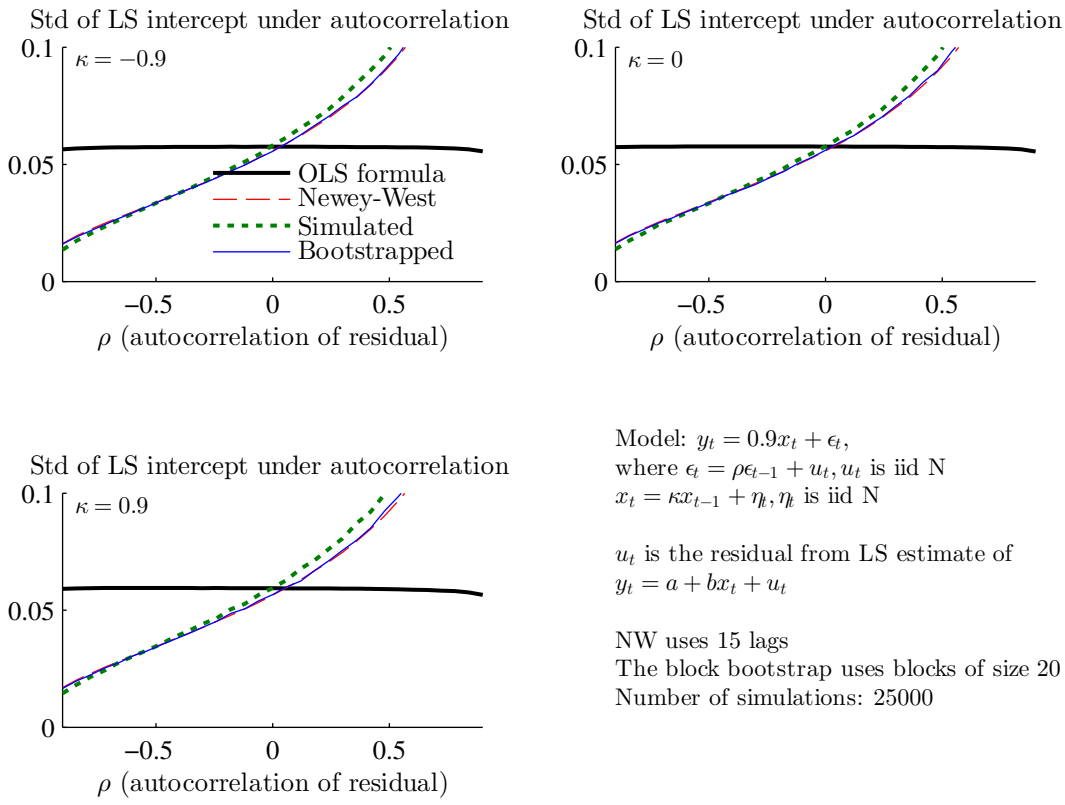


Figure 2.6: Standard error of OLS estimator, autocorrelated errors

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.

3 Return Distributions

Sections denoted by a star (*) is not required reading.

3.1 Estimating and Testing Distributions

Reference: Harvey (1989) 260, Davidson and MacKinnon (1993) 267, Silverman (1986); Mittelhammer (1996), DeGroot (1986)

3.1.1 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,

$$u_i = F_i(x_i) = \Pr(X_i \leq x_i). \tag{3.1}$$

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

$$F_i(x_i) = \int_{s=-\infty}^{x_i} f_i(s)ds. \tag{3.2}$$

(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 3.1 *(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, also called the "quantile function." See Figure 3.1 for an illustration.*

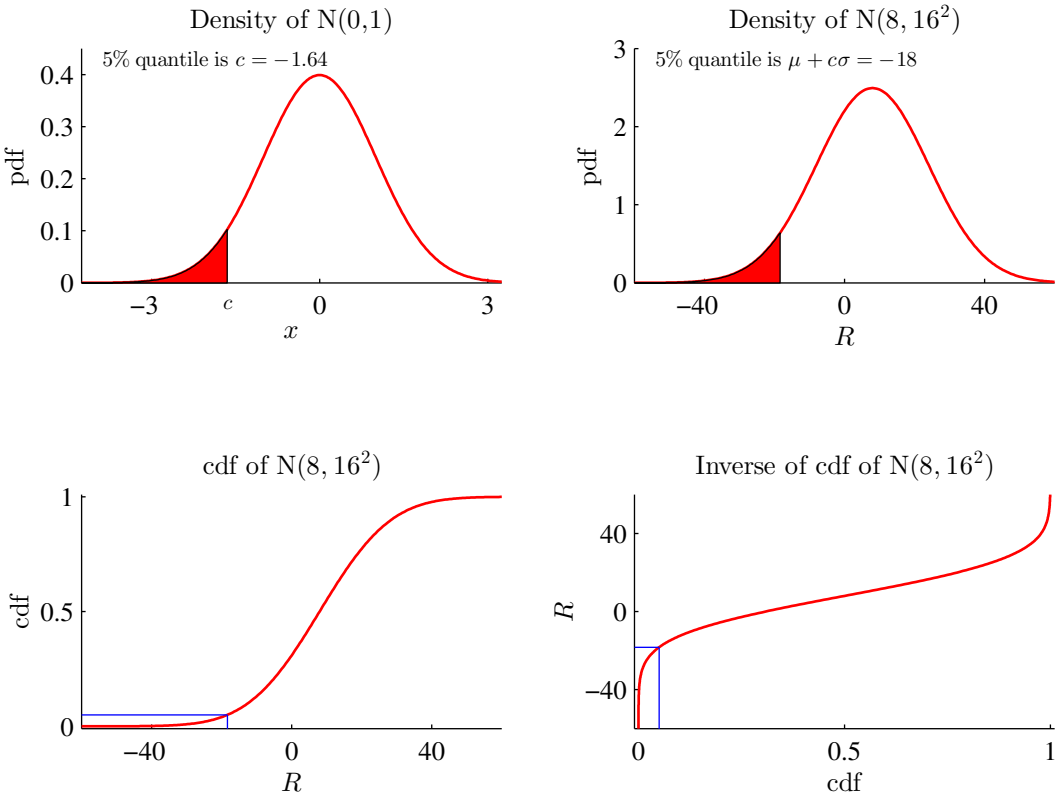


Figure 3.1: Finding quantiles of a N(μ,σ^2) distribution

3.1.2 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 3.2–3.4 for illustrations.

Remark 3.2 *(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.

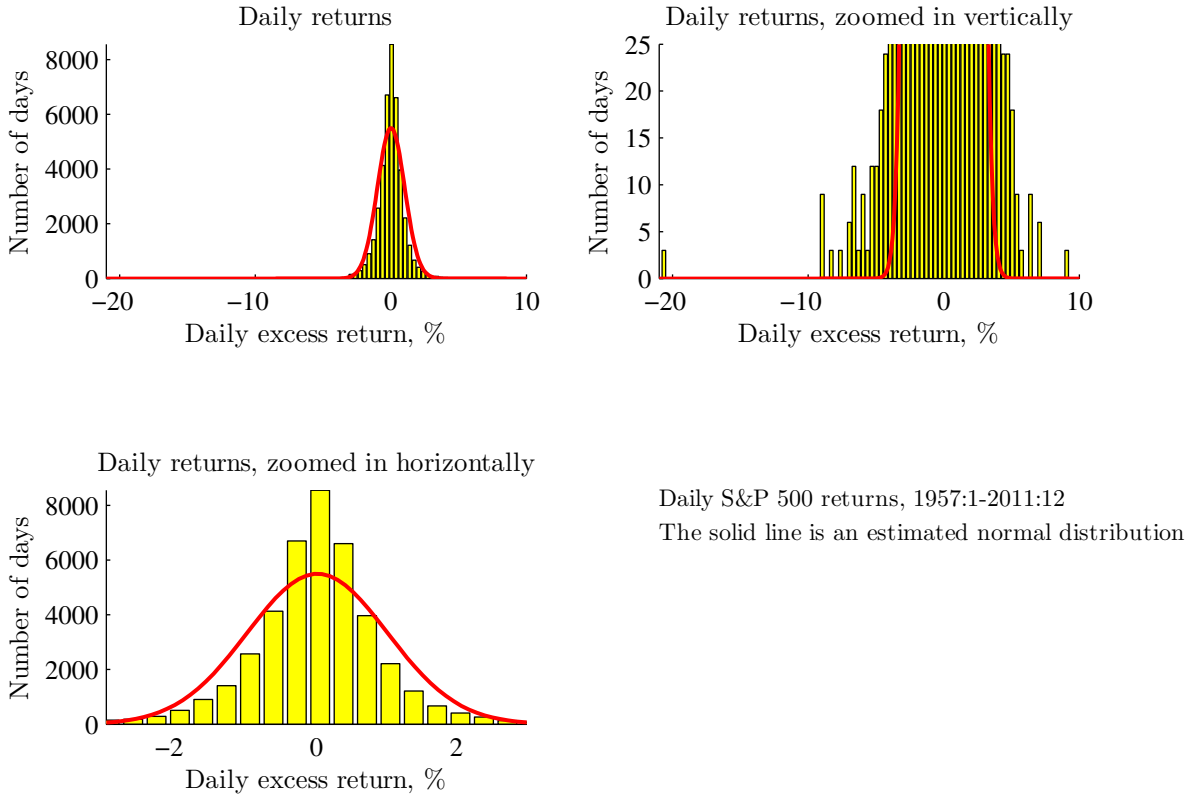


Figure 3.2: Distribution of daily S&P returns

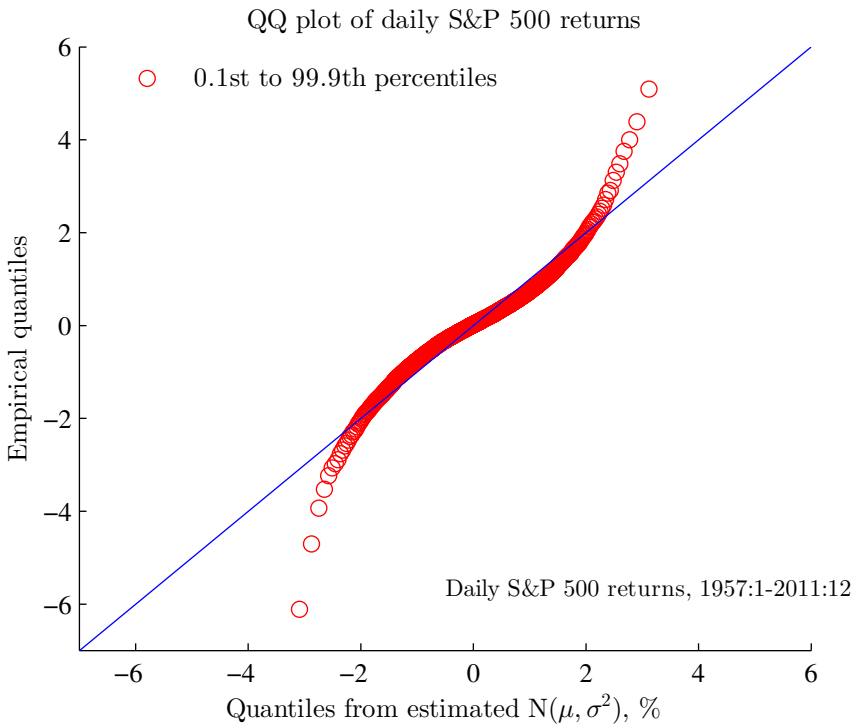


Figure 3.3: Quantiles of daily S&P returns

3.1.3 Parametric Tests of Normal Distribution

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

$$\begin{array}{lcll} & & \text{Test statistic} & \text{Distribution} \\ \text{skewness} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^3 & N\left(0,6/T\right) \\ \text{kurtosis} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^4 & N\left(3,24/T\right) \\ \text{Bera-Jarque} & = & \frac{T}{6}\text{skewness}^2+\frac{T}{24}\left(\text{kurtosis}-3\right)^2 & \chi_2^2. \end{array} \tag{3.3}$$

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (3.3) 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

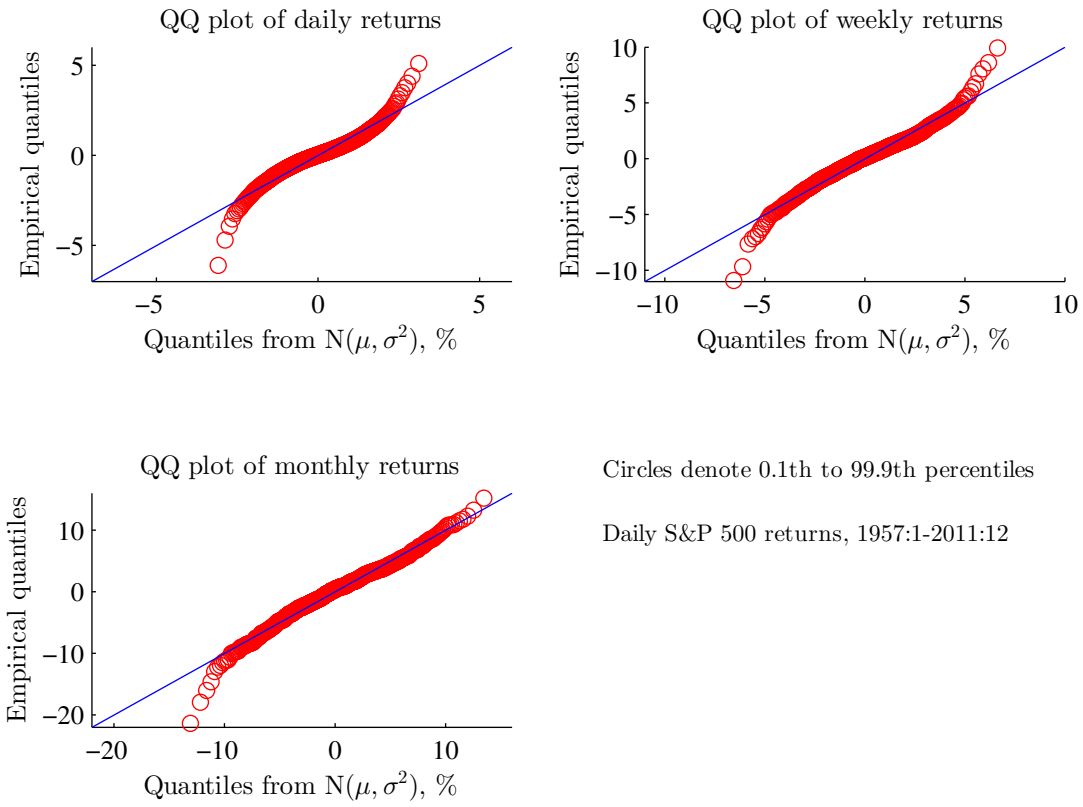


Figure 3.4: Distribution of S&P returns (different horizons)

sum of the square of two uncorrelated $N(0,1)$ variables, which has a χ^2_2 distribution.

The Bera-Jarque test can also be implemented as a test of overidentifying restrictions in GMM. The moment conditions

$$g(\mu,\sigma^2) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} x_t-\mu \\ (x_t-\mu)^2-\sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4-3\sigma^4 \end{bmatrix}, \tag{3.4}$$

should all be zero if x_t is $N(\mu,\sigma^2)$. We can estimate the two parameters, μ and σ^2, by using the first two moment conditions only, and then test if all four moment conditions are satisfied. It can be shown that this is the same as the Bera-Jarque test if x_t is indeed iid $N(\mu,\sigma^2)$.

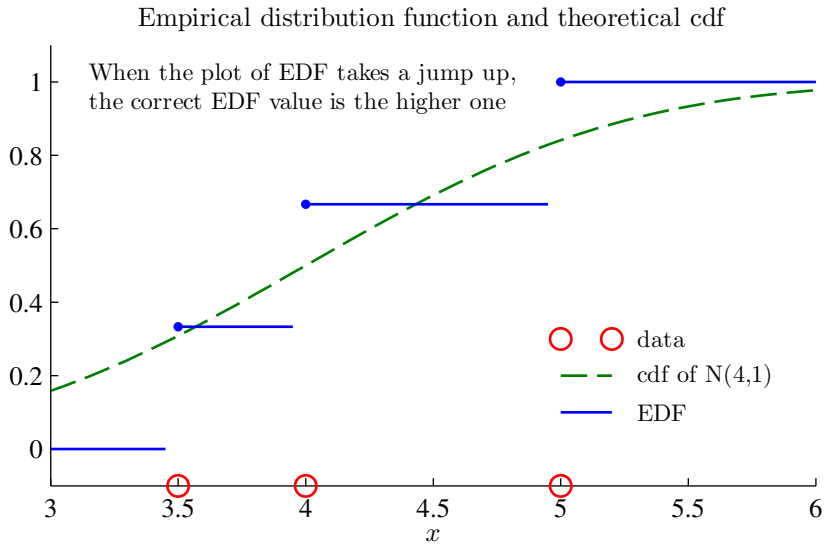


Figure 3.5: Example of empirical distribution function

3.1.4 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,

$$\text{EDF}(x) = \frac{1}{T}\sum_{t=1}^{T}\delta(x_t \leq x), \text{ where} \tag{3.5}$$

$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

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 3.5* for an illustration.

Example 3.3 *(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 3.5.*

Define the absolute value of the maximum distance

$$D_T = \max_{x_t} |\text{EDF}(x_t) - F(x_t)|. \tag{3.6}$$

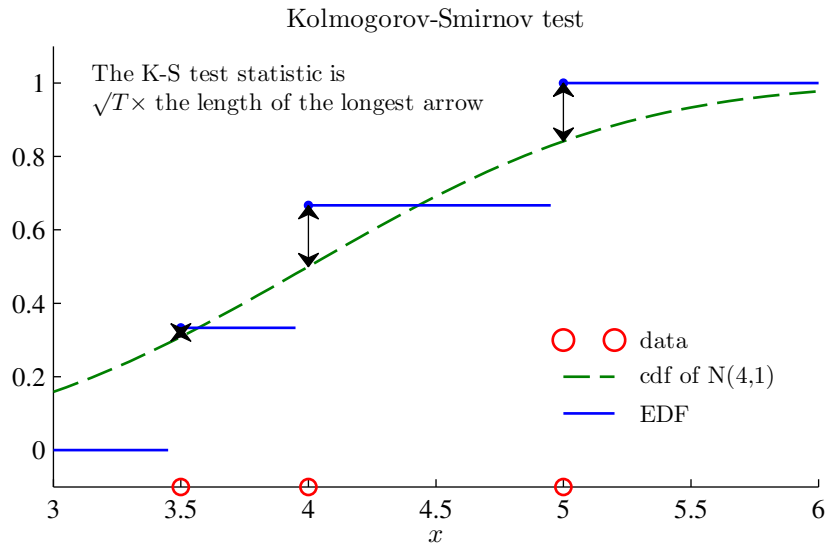


Figure 3.6: K-S test

Example 3.4 *(Kolmogorov-Smirnov test statistic) Figure 3.5 also shows the cumulative distribution function (cdf) of a normally distributed variable. The test statistic (3.6) 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 $\text{EDF}(x) = F(x)$ if $\sqrt{T}D_t > c$, where c is a critical value which can be calculated from

$$\lim_{T\to\infty} \Pr\left(\sqrt{T}D_T \leq c\right) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2c^2}. \tag{3.7}$$

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 3.7*. 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

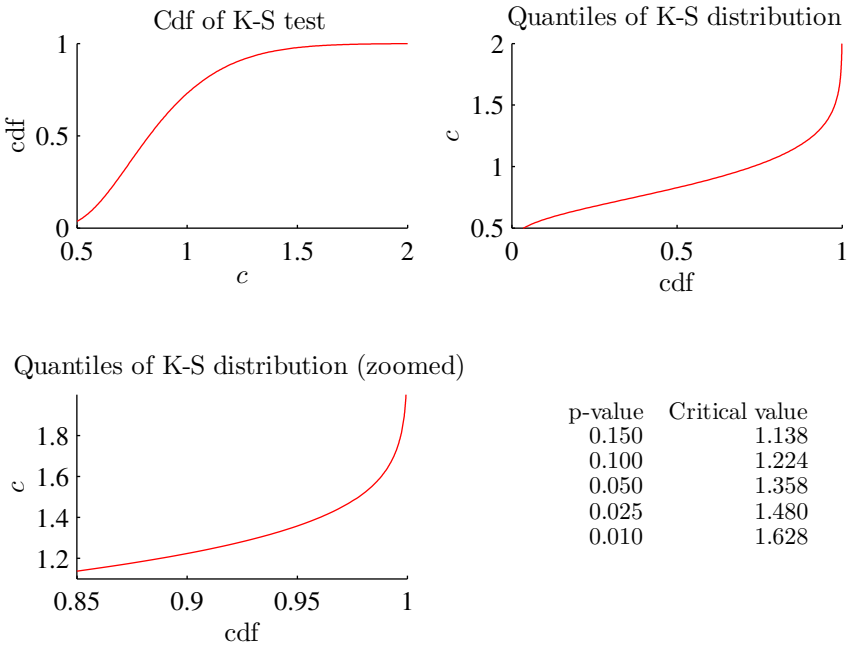


Figure 3.7: Distribution of the Kolmogorov-Smirnov test statistics, $\sqrt{T}D_T$

predicts that the fraction p_i should be in category i, with $\sum_{i=1}^{K} p_i = 1$. Then

$$\sum_{i=1}^{K} \frac{(N_i - Tp_i)^2}{Tp_i} \sim \chi^2_{K-1}. \tag{3.8}$$

There is a corresponding test for comparing two empirical distributions.

3.1.5 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 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)

$$f(x_t; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi) = (1 - \pi)\phi(x_t; \mu_1, \sigma_1^2) + \pi\phi(x_t; \mu_2, \sigma_2^2), \tag{3.9}$$

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

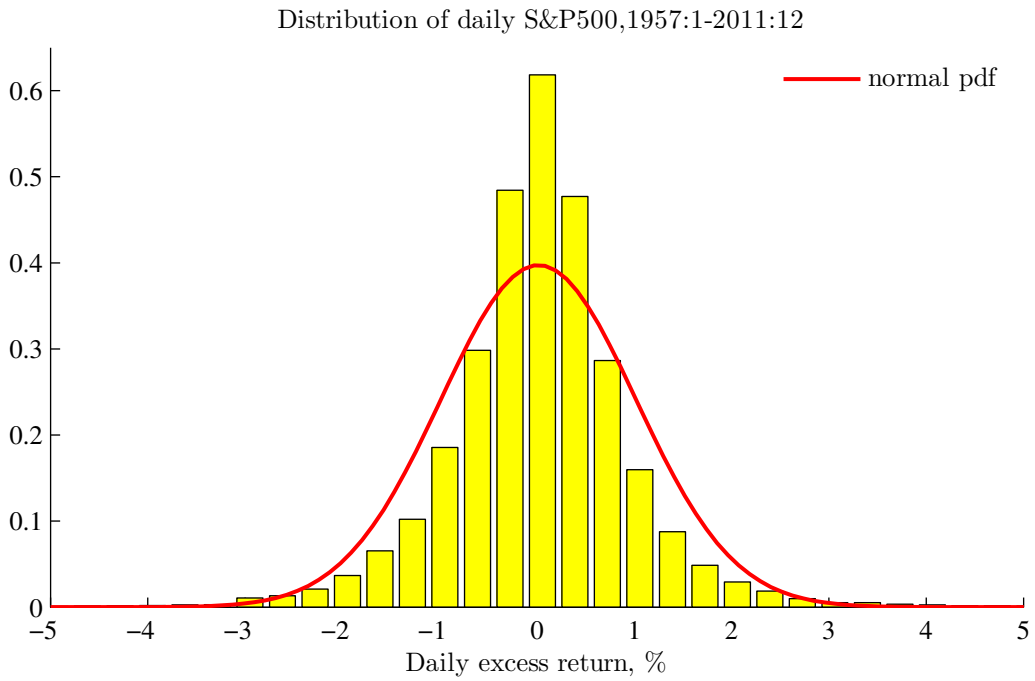


Figure 3.8: Histogram of returns and a fitted normal distribution

thus contains five parameters: the means and the variances of the two components and their relative weight (π).

See Figures 3.8–3.10 for an illustration.

Remark 3.5 *(Estimation of the mixture normal pdf) With 2 mixture components, the log likelihood is just*

$$LL = \sum_{t=1}^{T} \ln f(x_t; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi),$$

where $f()$ is the pdf in (3.9) 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:
(1) Guess values of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ and π. For instance, pick $\mu_1 = x_1$, $\mu_2 = x_2$, $\sigma_1^2 = \sigma_2^2 = \text{Var}(x_t)$ and $\pi = 0.5$.
(2) Calculate

$$\gamma_t = \frac{\pi\phi(x_t; \mu_2, \sigma_2^2)}{(1-\pi)\phi(x_t; \mu_1, \sigma_1^2) + \pi\phi(x_t; \mu_2, \sigma_2^2)} \text{ for } t = 1, \ldots, T.$$

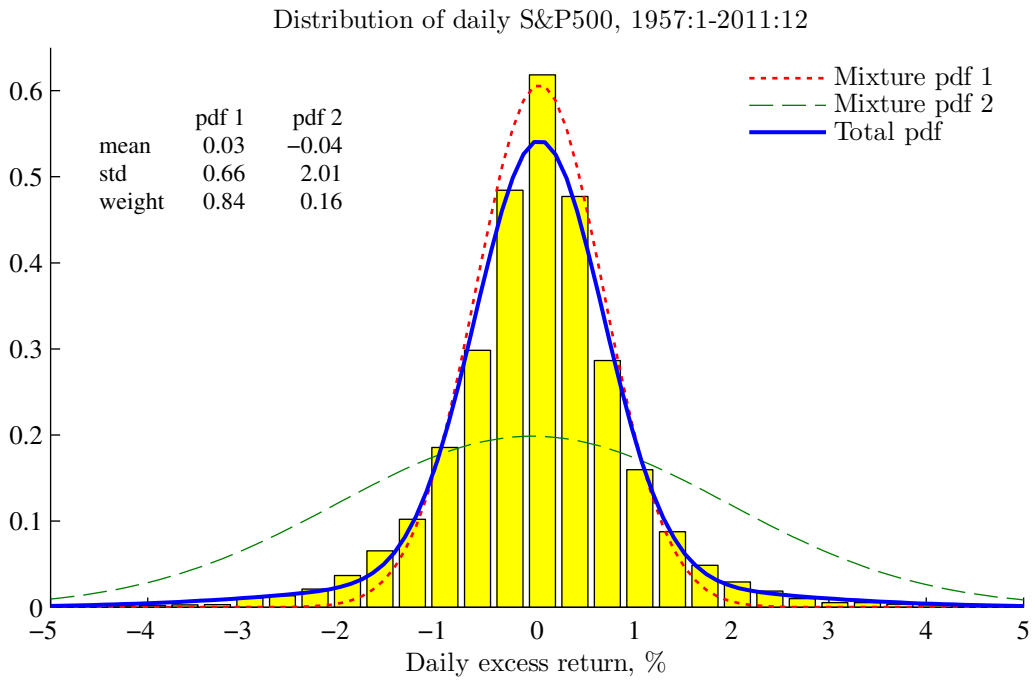


Figure 3.9: Histogram of returns and a fitted mixture normal distribution

(3) Calculate (in this order)

$$\mu_1 = \frac{\sum_{t=1}^{T}(1-\gamma_t)x_t}{\sum_{t=1}^{T}(1-\gamma_t)}, \sigma_1^2 = \frac{\sum_{t=1}^{T}(1-\gamma_t)(x_t-\mu_1)^2}{\sum_{t=1}^{T}(1-\gamma_t)},$$
$$\mu_2 = \frac{\sum_{t=1}^{T}\gamma_t x_t}{\sum_{t=1}^{T}\gamma_t}, \sigma_2^2 = \frac{\sum_{t=1}^{T}\gamma_t(x_t-\mu_2)^2}{\sum_{t=1}^{T}\gamma_t}, \text{ and}$$
$$\pi = \sum_{t=1}^{T}\gamma_t/T.$$

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) iteration.

3.1.6 Kernel Density Estimation

Reference: Silverman (1986)

A histogram is just a count of the relative number of observations that fall in (pre-

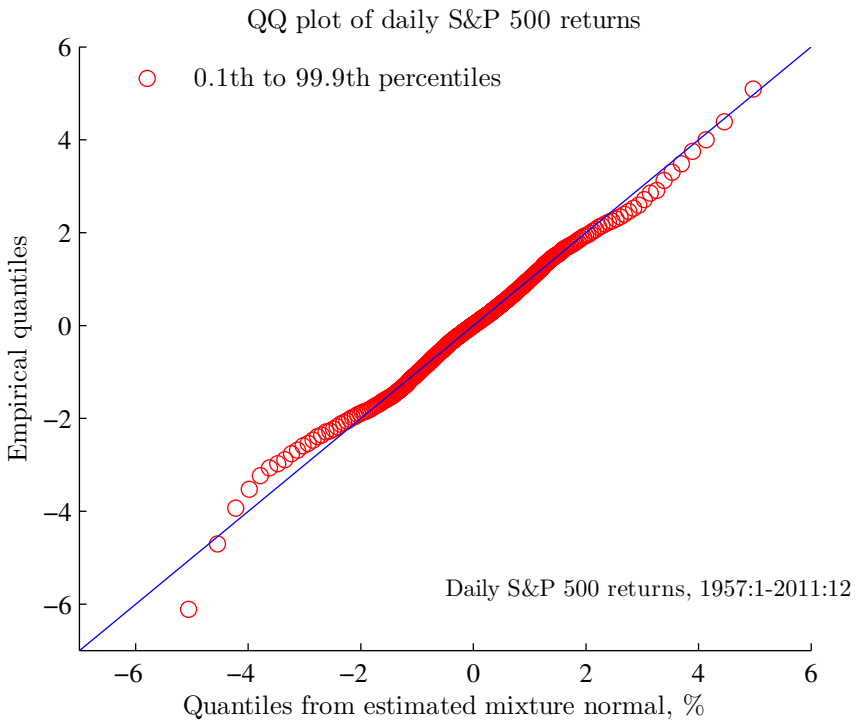


Figure 3.10: Quantiles of daily S&P returns

specified) 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 a wide, then the scaled histogram at the point x (say, $x = 2.3$) can be defined as

$$g(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{a}\delta(x_t \text{ is in bin}_i), \text{ where} \tag{3.10}$$

$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

Note that the area under $g(x)$ indeed integrates to unity.

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

$$g(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{h}\delta\left(\left|\frac{x_t - x}{h}\right| \leq 1/2\right). \tag{3.11}$$

In fact, that $\frac{1}{h}\delta(|x_t - x| \leq h/2)$ is the pdf value of a uniformly distributed variable (over the interval $x - h/2$ to $x + h/2$). This shows that our estimate of the pdf (here: the histogram) can be thought of as a average of hypothetical pdf values of the data in the neighbourhood of x. However, we can gain efficiency and get a smoother (across x values) estimate by using another density function 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. In fact, the $N(0, h^2)$ is often used. The kernel density estimator of the pdf at some point x is then

$$\hat{f}(x) = \frac{1}{T}\sum\nolimits_{t=1}^{T}\frac{1}{h\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_t - x}{h}\right)^2\right]. \tag{3.12}$$

Notice that the function in the summation is the density function of a $N(x, h^2)$ distribution.

The value $h = 1.06\,\mathrm{Std}(x_t)T^{-1/5}$ is sometimes recommended, since it can be shown to be the optimal choice (in MSE sense) if data is normally distributed and the gaussian kernel is used. The bandwidth h could also be chosen by a leave-one-out cross-validation technique.

See Figure 3.12 for an example and Figure 3.13 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

$$\sqrt{Th}[\hat{f}(x) - \mathrm{E}\,\hat{f}(x)] \to^d N\left[0, \frac{1}{2\sqrt{\pi}}f(x)\right], \tag{3.13}$$

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

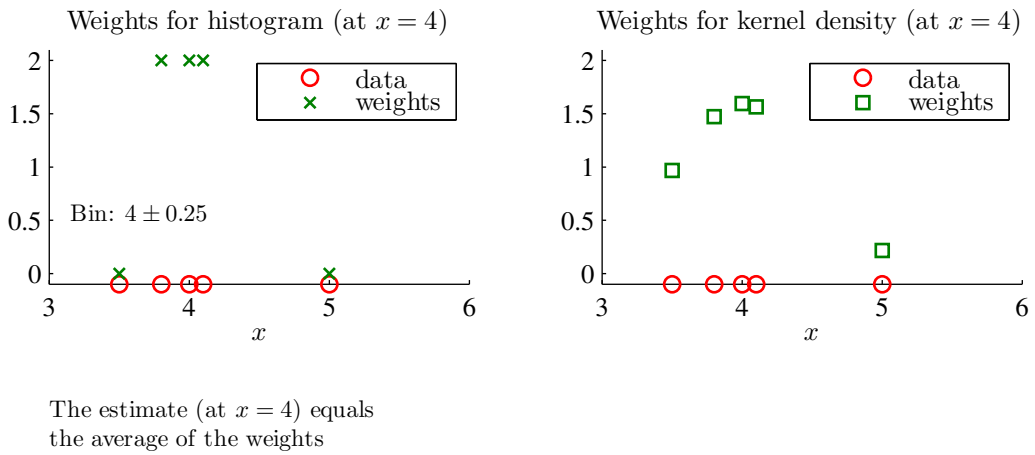


Figure 3.11: Calculation of the pdf at $x = 4$

Gaussian kernel as

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{(2\pi)^{d/2}|H^2\hat{\Omega}|^{1/2}} \exp\left[-\frac{1}{2}(x_t - x)'(H^2\hat{\Omega})^{-1}(x_t - x)\right]. \quad (3.14)$$

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 3.6 *((3.14) with $d = 1$) With just one variable, (3.14) becomes*

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{H\,\mathrm{Std}(x_t)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - x}{H\,\mathrm{Std}(x_t)}\right)^2\right],$$

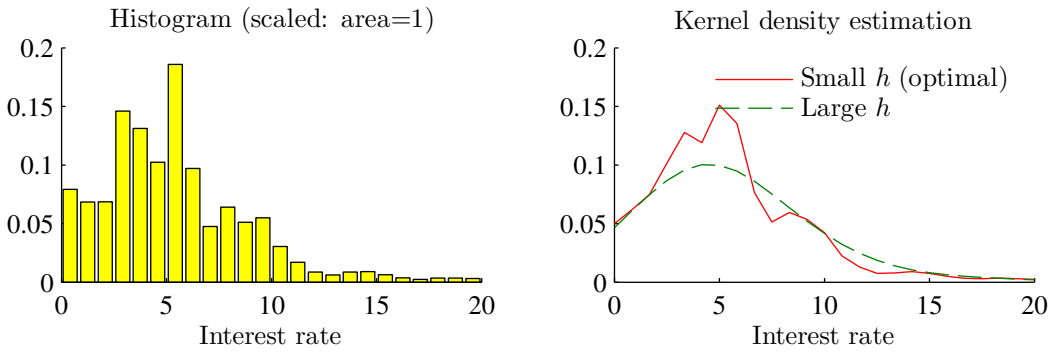
which is the same as (3.12) if $h = H\,\mathrm{Std}(x_t)$.

3.1.7 "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 (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 3.14–3.15 for an illustration.



Daily federal funds rates 1954:7-2011:12
K-S (against $\mathrm{N}(\mu, \sigma^2)$) : $\sqrt{T}D = 14.6$
Skewness: 1.1
kurtosis: 5.0
Bera-Jarque: 7774.9

Figure 3.12: Federal funds rate

3.2 Estimating Risk-neutral Distributions from Options

Reference: Breeden and Litzenberger (1978); Cox and Ross (1976), Taylor (2005) 16, Jackwerth (2000), Söderlind and Svensson (1997a) and Söderlind (2000)

3.2.1 The Breeden-Litzenberger Approach

A European call option price with strike price X has the price

$$C = \mathrm{E}\, M \max(0, S - X), \tag{3.15}$$

where M is the nominal discount factor and S is the price of the underlying asset at the expiration date of the option k periods from now.

We have seen that the price of a derivative is a discounted risk-neutral expectation of the derivative payoff. For the option it is

$$C = B_k \mathrm{E}^* \max(0, S - X), \tag{3.16}$$

where E^* is the risk-neutral expectation.

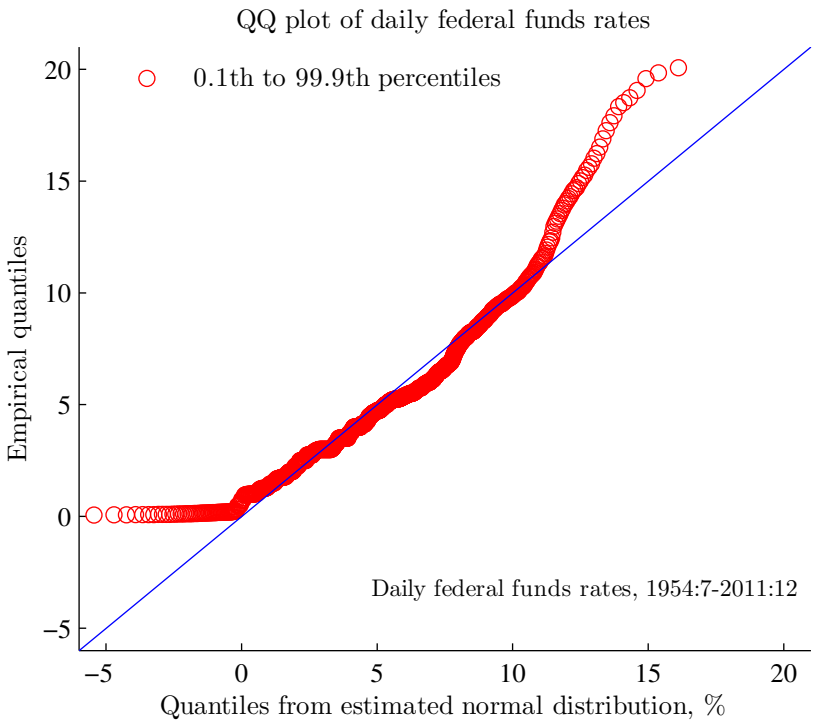


Figure 3.13: Federal funds rate

Example 3.7 *(Call prices, three states) Suppose that S only can take three values: 90, 100, and 110; and that the risk-neutral probabilities for these events are: 0.5, 0.4, and 0.1, respectively. We consider three European call option contracts with the strike prices 89, 99, and 109. From (3.16) their prices are (if $B = 1$)*

$$\begin{aligned} C\,(X = 89) &= 0.5(90-89) + 0.4(100-89) + 0.1(110-89) = 7 \\ C\,(X = 99) &= 0.5\times 0 + 0.4(100-99) + 0.1(110-99) = 1.5 \\ C\,(X = 109) &= 0.5\times 0 + 0.4\times 0 + 0.1(110-109) = 0.1. \end{aligned}$$

Clearly, with information on the option prices, we could in this case back out what the probabilities are.

(3.16) can also be written as

$$C = \exp(-ik)\int_X^\infty (S-X)\,h^*(S)\,dS, \tag{3.17}$$

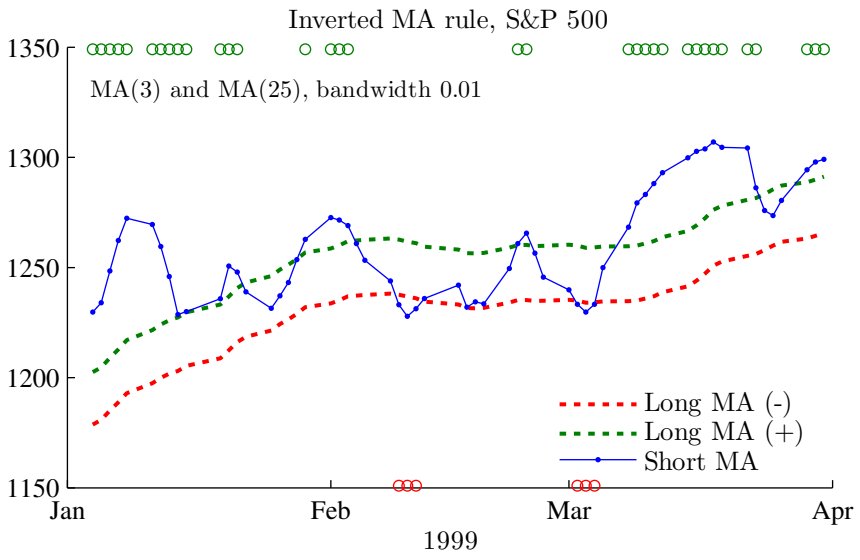


Figure 3.14: Examples of trading rules

where i is the per period (annualized) interest rate so $\exp(-ik) = B_k$ and $h^*(S)$ is the (univariate) risk-neutral probability density function of the underlying price (not its log). Differentiating (3.17) with respect to the strike price and rearranging gives the risk-neutral distribution function

$$\Pr^*(S \leq X) = 1 + \exp(ik)\frac{\partial C(X)}{\partial X}. \tag{3.18}$$

Proof. Differentiating the call price with respect to the strike price gives

$$\frac{\partial C}{\partial X} = -\exp(-ik)\int_X^\infty h^*(S)\,dS = -\exp(-ik)\Pr^*(S > X).$$

Use $\Pr^*(S > X) = 1 - \Pr^*(S \leq X)$. ■

Differentiating once more gives the risk-neutral probability density function of S at $S = X$

$$\text{pdf}^*(X) = \exp(ik)\frac{\partial^2 C(X)}{\partial X^2}. \tag{3.19}$$

Figure 3.16 shows some data and results for German bond options on one trading date. (A change of variable approach is used to show the distribution of the log asset price.)

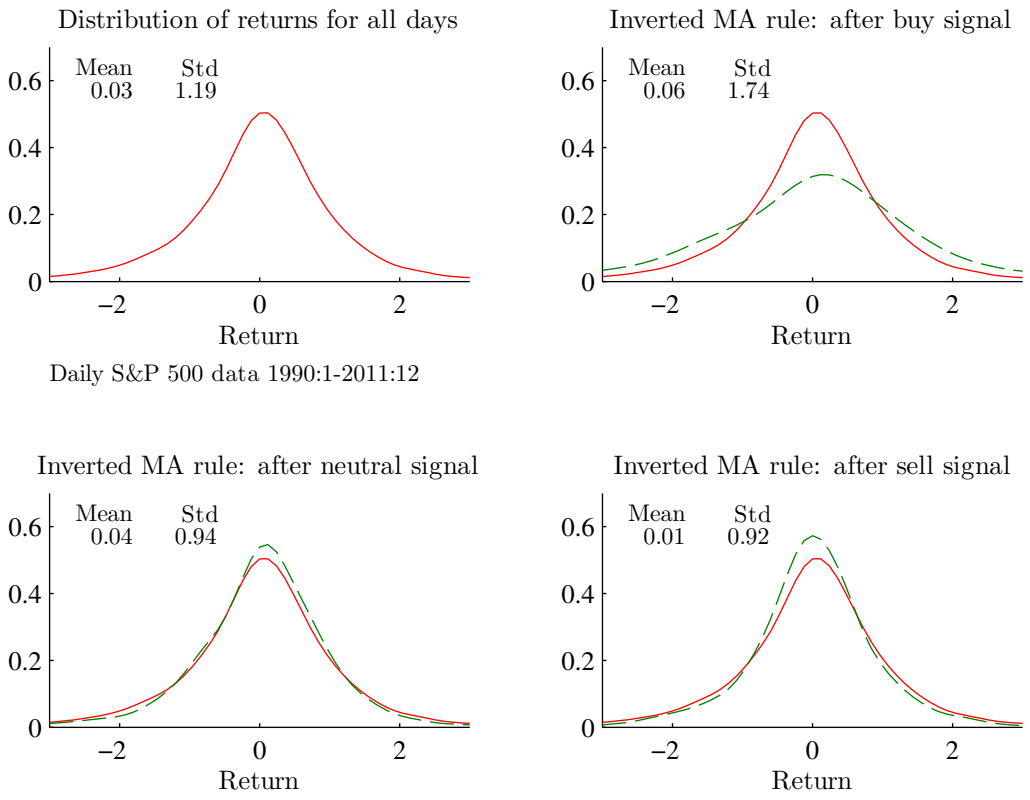


Figure 3.15: Examples of trading rules

A difference quotient approximation of the derivative in (3.18)

$$\frac{\partial C}{\partial X} \approx \frac{1}{2}\left[\frac{C\left(X_{i+1}\right)-C\left(X_i\right)}{X_{i+1}-X_i}+\frac{C\left(X_i\right)-C\left(X_{i-1}\right)}{X_i-X_{i-1}}\right] \tag{3.20}$$

gives the approximate distribution function. The approximate probability density function, obtained by a second-order difference quotient

$$\frac{\partial^2 C}{\partial X^2} \approx \left[\frac{C\left(X_{i+1}\right)-C\left(X_i\right)}{X_{i+1}-X_i}-\frac{C\left(X_i\right)-C\left(X_{i-1}\right)}{X_i-X_{i-1}}\right] / \left[\frac{1}{2}\left(X_{i+1}-X_{i-1}\right)\right] \tag{3.21}$$

is also shown. The approximate distribution function is decreasing in some intervals, and the approximate density function has some negative values and is very jagged. This could possibly be explained by some aberrations of the option prices, but more likely by the approximation of the derivatives: changing approximation method (for instance, from centred to forward difference quotient) can have a strong effect on the results, but

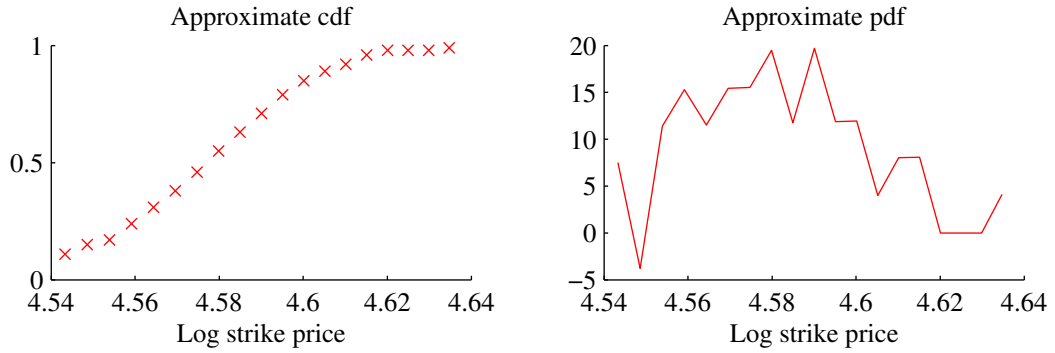


Figure 3.16: Bund options 6 April 1994. Options expiring in June 1994.

all methods seem to generate strange results in some interval. This suggests that it might be important to estimate an explicit distribution. That is, to impose enough restrictions on the results to guarantee that they are well behaved.

3.2.2 Mixture of Normals

A flexible way of estimating an explicit distribution is to assume that the distribution of the logs of M and S, conditional on the information today, is a mixture of n bivariate normal distributions (see Söderlind and Svensson (1997b)). Let $\phi(x;\mu,\Omega)$ denote a normal multivariate density function over x with mean vector μ and covariance matrix Ω. The weight of the j^{th} normal distribution is $\alpha^{(j)}$, so the probability density function, pdf, of $\ln M$ and $\ln S$ is assumed to be

$$\text{pdf}\left(\begin{bmatrix}\ln M\\ \ln S\end{bmatrix}\right)=\sum_{j=1}^{n}\alpha^{(j)}\phi\left(\begin{bmatrix}\ln M\\ \ln S\end{bmatrix};\begin{bmatrix}\mu_m^{(j)}\\ \mu_s^{(j)}\end{bmatrix},\begin{bmatrix}\sigma_{mm}^{(j)} & \sigma_{ms}^{(j)}\\ \sigma_{ms}^{(j)} & \sigma_{ss}^{(j)}\end{bmatrix}\right), \tag{3.22}$$

with $\sum_{j=1}^{n}\alpha^{(j)}=1$ and $\alpha^{(j)}\geq 0$. One interpretation of mixing normal distributions is that they represent different macro economic 'states', where the weight is interpreted as the probability of state j.

Let Φ (.) be the standardized (univariate) normal distribution function. If $\mu_m^{(j)}=\mu_m$ and $\sigma_{mm}^{(j)}=\sigma_{mm}$ in (3.22), then the marginal distribution of the log SDF is gaussian

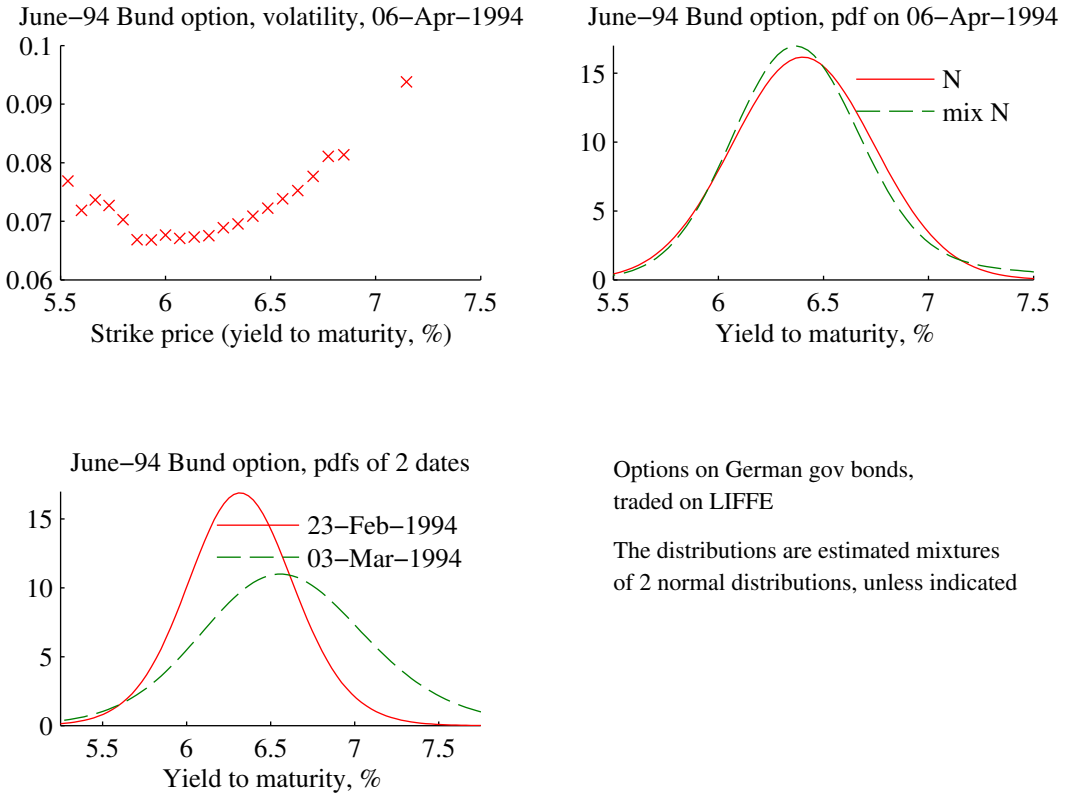


Figure 3.17: Bund options 23 February and 3 March 1994. Options expiring in June 1994.

(while that of the underlying asset price is not). In this case the European call option price (3.15) has a closed form solution in terms of the spot interest rate, strike price, and the parameters of the bivariate distribution[1]

$$C = \exp(-ik)\sum_{j=1}^{n}\alpha^{(j)}\left[\exp\left(\mu_s^{(j)}+\sigma_{ms}^{(j)}+\frac{1}{2}\sigma_{ss}^{(j)}\right)\Phi\left(\frac{\mu_s^{(j)}+\sigma_{ms}^{(j)}+\sigma_{ss}^{(j)}-\ln X}{\sqrt{\sigma_{ss}^{(j)}}}\right)\right.$$
$$\left. - X\Phi\left(\frac{\mu_s^{(j)}+\sigma_{ms}^{(j)}-\ln X}{\sqrt{\sigma_{ss}^{(j)}}}\right)\right]. \tag{3.23}$$

[1]Without these restrictions, $\alpha^{(j)}$ in (3.23) is replaced by $\tilde{\alpha}^{(j)} = \alpha^{(j)}\exp(\bar{m}^{(j)} + \sigma_{mm}^{(j)}/2)/\sum_{j=1}^{n}\alpha^{(j)}\exp(\mu_m^{(j)} + \sigma_{mm}^{(j)}/2)$. In this case, $\tilde{\alpha}^{(j)}$, not $\alpha^{(j)}$, will be estimated from option data.

(For a proof, see Söderlind and Svensson (1997b).) Notice that this is like using the physical distribution, but with $\mu_s^{(j)} + \sigma_{ms}^{(j)}$ instead of $\mu_s^{(j)}$.

Notice also that this is a weighted average of the option price that would hold in each state

$$C = \sum_{j=1}^{n} \alpha^{(j)} C^{(j)}. \tag{3.24}$$

(See Ritchey (1990) and Melick and Thomas (1997).)

A forward contract written in t stipulates that, in period τ, the holder of the contract gets one asset and pays F. This can be thought of as an option with a zero strike price and no discounting—and it is also the mean of the riskneutral distribution. The forward price then follows directly from (3.23) as

$$F = \sum_{j=1}^{n} \alpha^{(j)} \exp\left(\mu_s^{(j)} + \sigma_{ms}^{(j)} + \frac{\sigma_{ss}^{(j)}}{2}\right). \tag{3.25}$$

There are several reasons for assuming a mixture of normal distributions. First, non-parametric methods often generate strange results, so we need to assume *some* parametric distribution. Second, it gives closed form solutions for the option and forward prices, which is very useful in the estimation of the parameters. Third, it gives the Black-Scholes model as a special case when $n = 1$.

To see the latter, let $n = 1$ and use the forward price from (3.25), $F = \exp(\mu_s + \sigma_{ms} + \sigma_{ss}/2)$, in the option price (3.23) to get

$$C = \exp(-ik) F \Phi\left(\frac{\ln F/X + \sigma_{ss}/2}{\sqrt{\sigma_{ss}}}\right) - \exp(-ik) X \Phi\left(\frac{\ln F/X - \sigma_{ss}/2}{\sqrt{\sigma_{ss}}}\right), \tag{3.26}$$

which is indeed Black's formula.

We want to estimate the marginal distribution of the future asset price, S. From (3.22), it is a mixture of univariate normal distributions with weights $\alpha^{(j)}$, means $\mu_s^{(j)}$, and variances $\sigma_{ss}^{(j)}$. The basic approach is to back out these parameters from data on option and forward prices by exploiting the pricing relations (3.23)–(3.25). For that we need data on at least at many different strike prices as there are parameters to estimate.

Remark 3.8 *Figures 3.16–3.17 show some data and results (assuming a mixture of two normal distributions) for German bond options around the announcement of the very high money growth rate on 2 March 1994..*

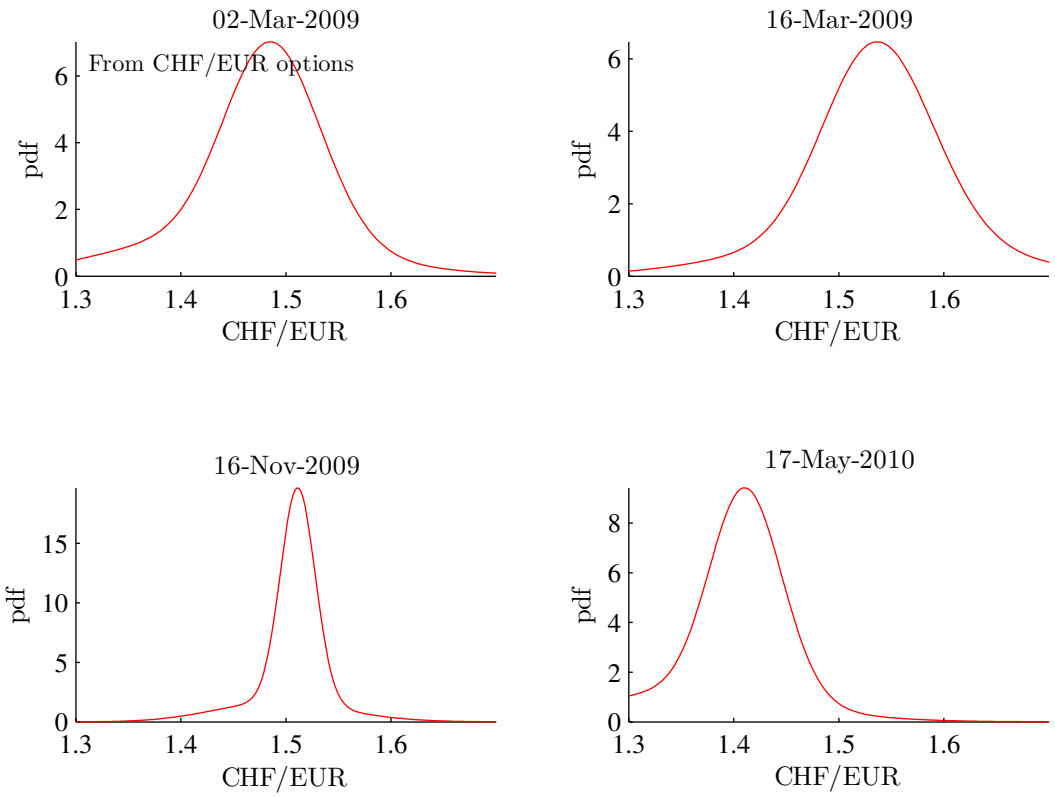


Figure 3.18: Riskneutral distribution of the CHF/EUR exchange rate

Remark 3.9 *Figures 3.18–3.20 show results for the CHF/EUR exchange rate around the period of active (Swiss) central bank interventions on the currency market.*

Remark 3.10 *(Robust measures of the standard deviation and skewness) Let P_α be the αth quantile (for instance, quantile 0.1) of a distribution. A simple robust measure of the standard deviation is just the difference between two symmetric quantile,*

$$\text{Std} = P_{1-\alpha} - P_\alpha,$$

where it is assumed that $\alpha < 0.5$. Sometimes this measure is scaled so it would give the right answer for a normal distribution. For instance, with $\alpha = 0.1$, the measure would be divided by 2.56 and for $\alpha = 0.25$ by 1.35.

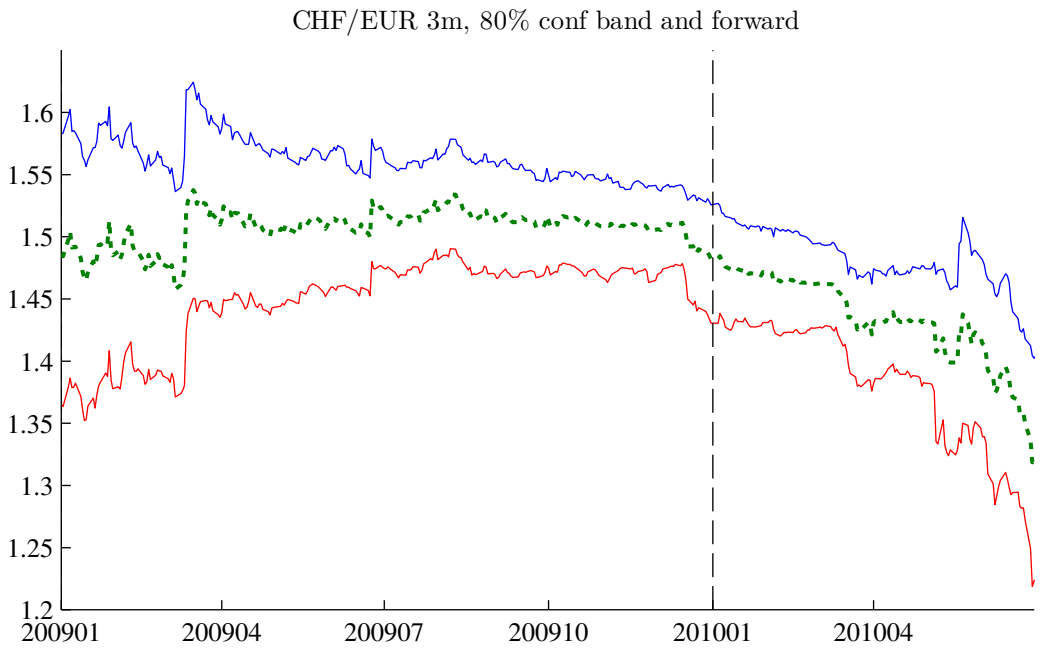


Figure 3.19: Riskneutral distribution of the CHF/EUR exchange rate

One of the classical robust skewness measures was suggested by Hinkley

$$Skew = \frac{(P_{1-\alpha} - P_{0.5}) - (P_{0.5} - P_{\alpha})}{P_{1-\alpha} - P_{\alpha}}.$$

This skewness measure can only take on values between -1 *(when* $P_{1-\alpha} = P_{0.5}$*) and* 1 *(when* $P_{\alpha} = P_{0.5}$*). When the median is just between the two percentiles (*$P_{0.5} = (P_{1-\alpha} + P_{\alpha})/2$*), then it is zero.*

3.3 Threshold Exceedance and Tail Distribution*

Reference: McNeil, Frey, and Embrechts (2005) 7

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

$$X = -R, \tag{3.27}$$

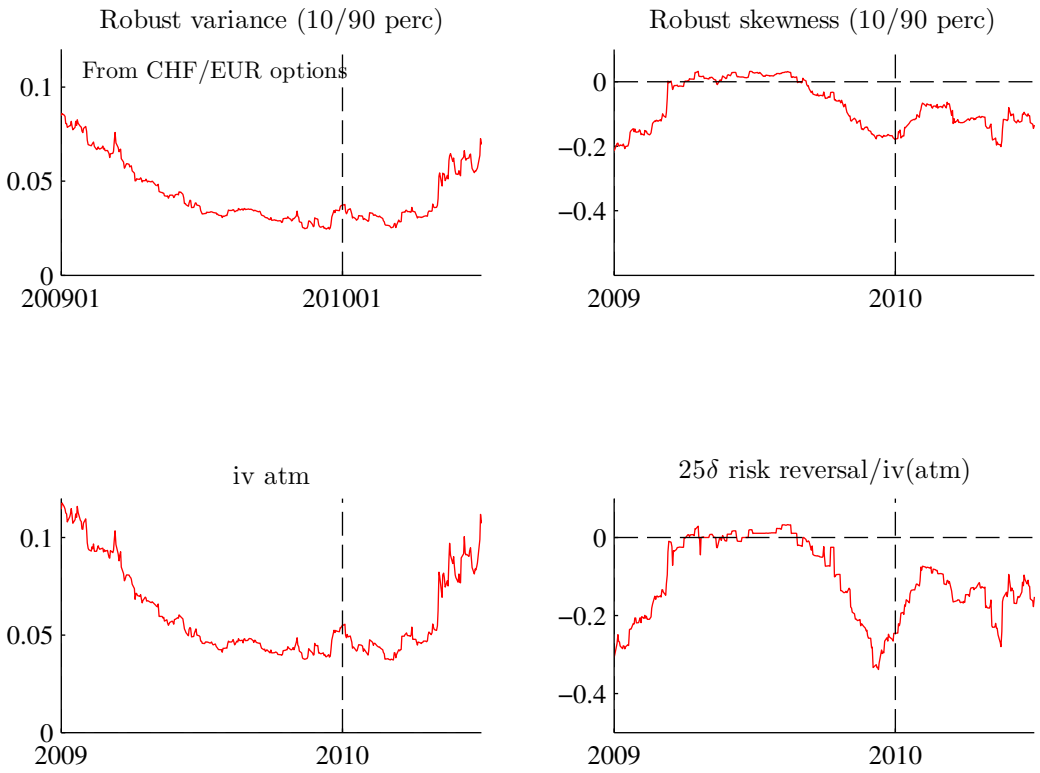


Figure 3.20: Riskneutral distribution of the CHF/EUR exchange rate

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, 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. See *Figure 3.21* for an illustration.

Remark 3.11 *(Cdf and pdf of the generalized Pareto distribution) The generalized Pareto distribution is described by a scale parameter ($\beta > 0$) and a shape parameter (ξ). The cdf* ($\Pr(Z \leq z)$*, where Z is the random variable and z is a value) is*

$$G(z) = \begin{cases} 1 - (1 + \xi z/\beta)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-z/\beta) & \xi = 0, \end{cases}$$

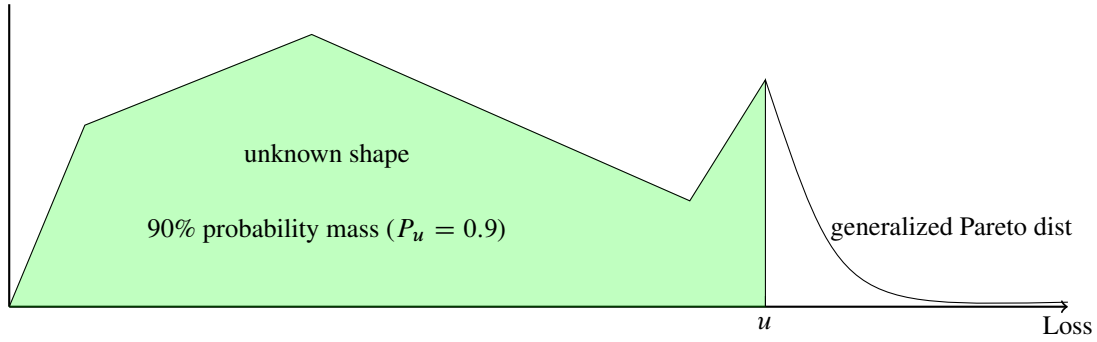


Figure 3.21: Loss distribution

for $0 \leq z$ *if* $\xi \geq 0$ *and* $z \leq -\beta/\xi$ *in case* $\xi < 0$. *The pdf is therefore*

$$g(z) = \begin{cases} \frac{1}{\beta}(1+\xi z/\beta)^{-1/\xi-1} & \text{if } \xi \neq 0 \\ \frac{1}{\beta}\exp(-z/\beta) & \xi = 0. \end{cases}$$

The mean is defined (finite) if $\xi < 1$ *and is then* $\mathrm{E}(z) = \beta/(1-\xi)$. *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 3.22 for an illustration.*

Remark 3.12 *(Random number from a generalized Pareto distribution*) By inverting the Cdf, we can notice that if* u *is uniformly distributed on* $(0, 1]$, *then we can construct random variables with a GPD by*

$$\begin{aligned} z &= \tfrac{\beta}{\xi}[(1-u)^{-\xi} - 1] \quad \text{if } \xi \neq 0 \\ z &= -\ln(1-u)\beta \qquad \xi = 0. \end{aligned}$$

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 $X \leq u$. Then, the cdf of the loss for values greater than the threshold ($\Pr(X \leq x)$ for $x > u$) can be written

$$F(x) = P_u + G(x-u)(1-P_u), \text{ for } x > u, \tag{3.28}$$

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.

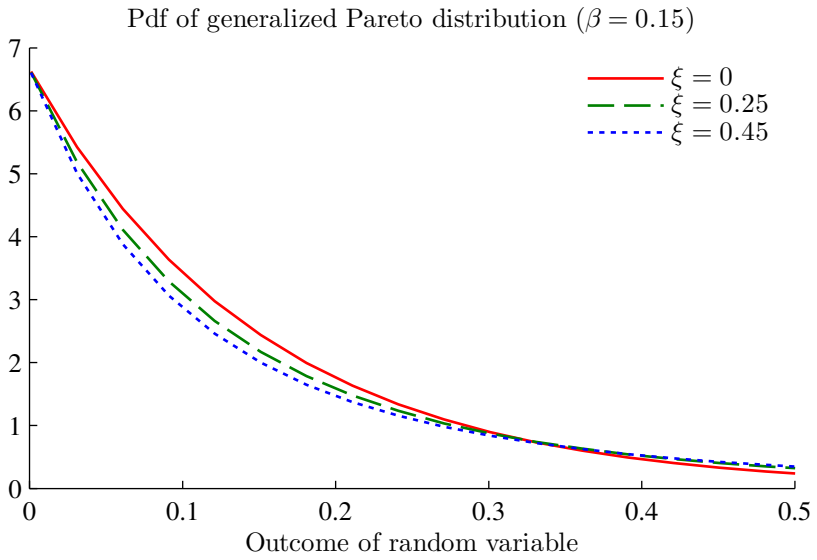


Figure 3.22: Generalized Pareto distributions

Clearly, the pdf is

$$f(x) = g(x-u)(1-P_u), \text{ for } x > u, \tag{3.29}$$

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 Figure 3.24 for an illustration.

It is often to calculate the *tail probability* $\Pr(X > x)$, which in the case of the cdf in (3.28) is

$$1 - F(x) = (1 - P_u)[1 - G(x-u)], \tag{3.30}$$

where $G(z)$ is the cdf of the generalized Pareto distribution.

The VaR_α (say, $\alpha = 0.95$) is the α-th quantile of the loss distribution

$$\text{VaR}_\alpha = \text{cdf}_X^{-1}(\alpha), \tag{3.31}$$

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

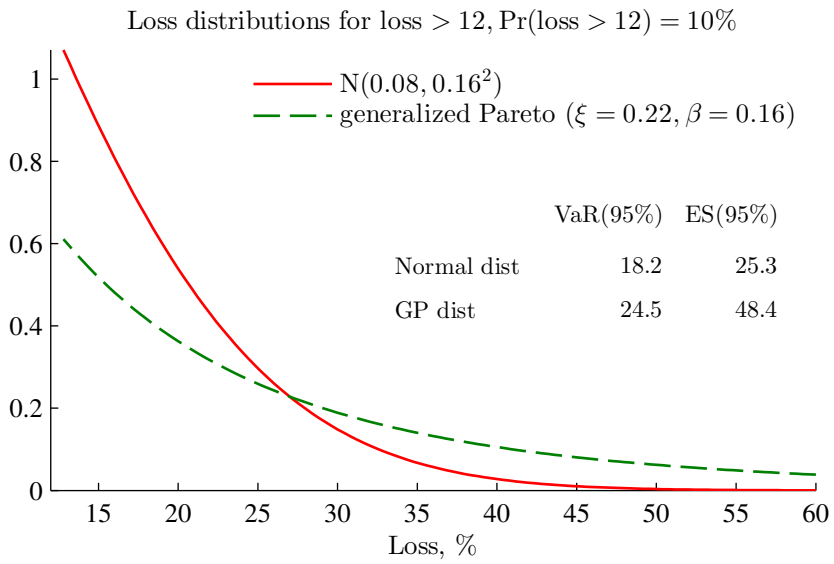


Figure 3.23: Comparison of a normal and a generalized Pareto distribution for the tail of losses

equals α

$$\Pr(X \leq \text{VaR}_\alpha) = \alpha. \tag{3.32}$$

(Equivalently, the $\Pr(X > \text{VaR}_\alpha) = 1 - \alpha$.)

Assuming α is higher than P_u (so $\text{VaR}_\alpha \geq u$), the cdf (3.28) together with the form of the generalized Pareto distribution give

$$\text{VaR}_\alpha = \begin{cases} u + \frac{\beta}{\xi}\left[\left(\frac{1-\alpha}{1-P_u}\right)^{-\xi} - 1\right] & \text{if } \xi \neq 0 \\ u - \beta \ln\left(\frac{1-\alpha}{1-P_u}\right) & \xi = 0 \end{cases}, \text{ for } \alpha \geq P_u. \tag{3.33}$$

Proof. (of (3.33)) Set $F(x) = \alpha$ in (3.28) and use $z = x - u$ in the cdf from Remark 3.11 and solve for x. ■

If we assume $\xi < 1$ (to make sure that the mean is finite), then straightforward integration using (3.29) shows that the expected shortfall is

$$\begin{aligned} \text{ES}_\alpha &= \text{E}(X \mid X \geq \text{VaR}_\alpha) \\ &= \frac{\text{VaR}_a}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \text{ for } \alpha > P_u \text{ and } \xi < 1. \end{aligned} \tag{3.34}$$

Let $\upsilon = \text{VaR}_\alpha$ and then subtract υ from both sides of the expected shortfall to get the expected exceedance of the loss over another threshold $\upsilon > u$

$$\begin{aligned} e(\upsilon) &= \text{E}\,(X - \upsilon | X > \upsilon) \\ &= \frac{\xi\upsilon}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \text{ for } \upsilon > u \text{ and } \xi < 1. \end{aligned} \tag{3.35}$$

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 3.24 for an illustration). This provides a way of assessing which distribution that best fits the tail of the historical histogram.

Remark 3.13 *(Expected exceedance from a normal distribution) If* $X \sim N(\mu, \sigma^2)$*, then*

$$\begin{aligned} \text{E}(X - \upsilon | X > \upsilon) &= \mu + \sigma \frac{\phi(\upsilon_0)}{1 - \Phi(\upsilon_0)} - \upsilon, \\ \textit{with } \upsilon_0 &= (\upsilon - \mu)/\sigma \end{aligned}$$

where $\phi()$ *and* Φ *are the pdf and cdf of a* $N(0,1)$ *variable respectively.*

The expected exceedance over υ is often compared with an empirical estimate of the same thing: the mean of $X_t - \upsilon$ for those observations where $X_t > \upsilon$

$$\begin{aligned} \hat{e}(\upsilon) &= \frac{\sum_{t=1}^{T}(X_t - \upsilon)\delta(X_t > \upsilon)}{\sum_{t=1}^{T}(X_t > \upsilon)}, \text{ where} \\ \delta(q) &= \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases} \end{aligned} \tag{3.36}$$

If it is found that $\hat{e}(\upsilon)$ is increasing (more or less) linearly with the threshold level (υ), then it is reasonable to model the tail of the distribution from that point as a generalized Pareto distribution.

The estimation of the parameters of the distribution (ξ and β) is typically done by maximum likelihood. Alternatively, A comparison of the empirical exceedance (3.36) with the theoretical (3.35) 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)

$$\hat{e}(v_i) = a + bv_i + \varepsilon_i. \tag{3.37}$$

Then, the theoretical exceedance (3.35) for a given starting point of the GPD u is related to this regression according to

$$\begin{aligned} a &= \frac{\beta - \xi u}{1-\xi} \text{ and } b = \frac{\xi}{1-\xi}, \text{ or} \\ \xi &= \frac{b}{1+b} \text{ and } \beta = a(1-\xi) + \xi u. \end{aligned} \tag{3.38}$$

See Figure 3.25 for an illustration.

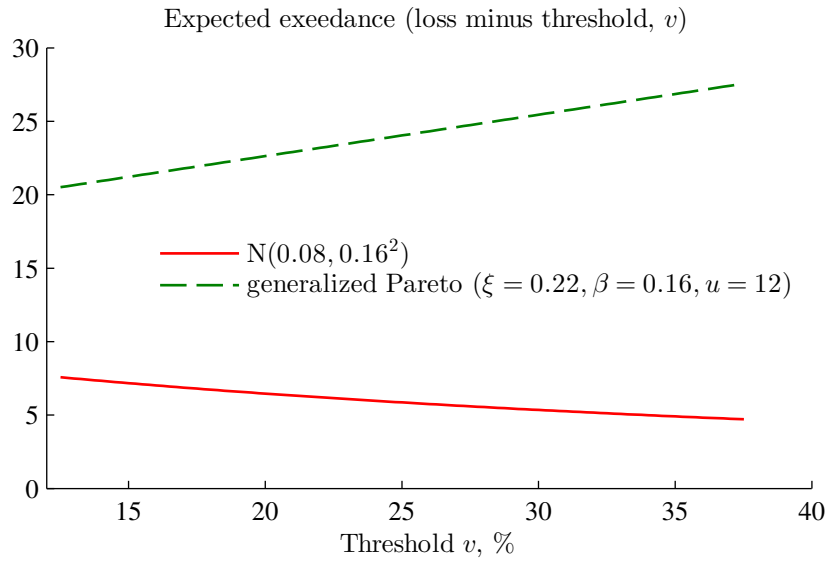


Figure 3.24: Expected exceedance, normal and generalized Pareto distribution

Remark 3.14 *(Log likelihood function of the loss distribution) Since we have assumed that the threshold exceedance $(X-u)$ has a generalized Pareto distribution, Remark 3.11 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).

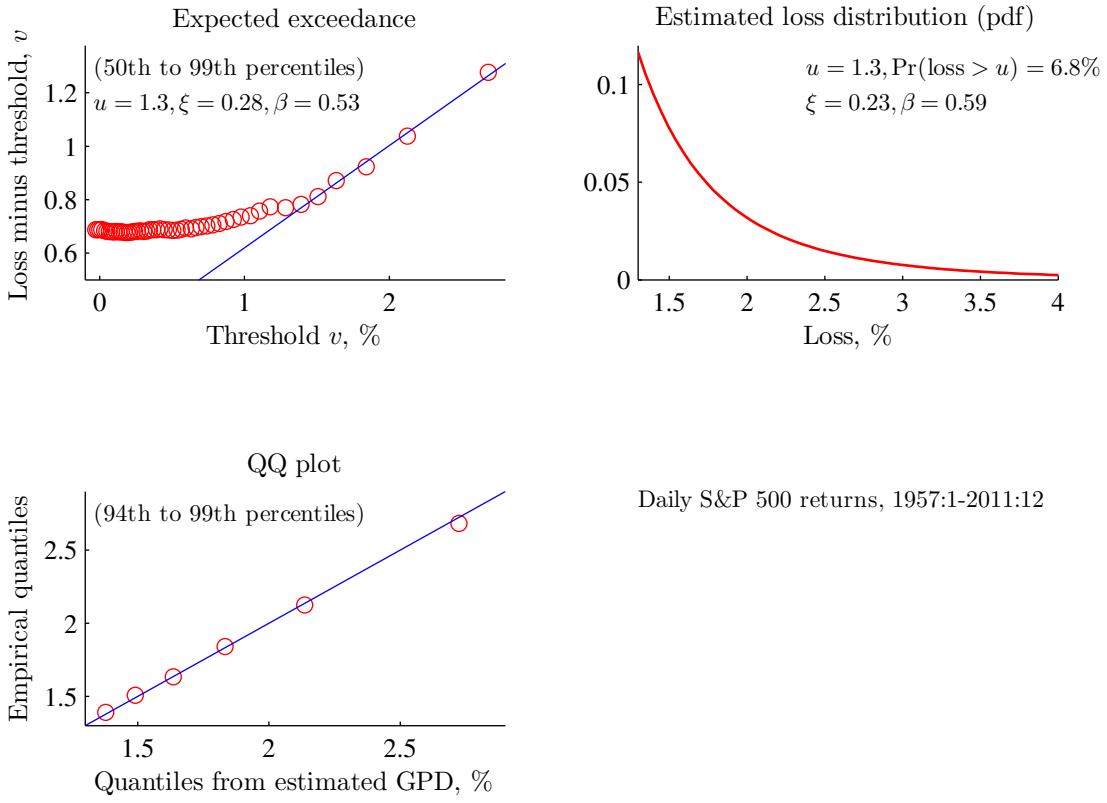


Figure 3.25: Results from S&P 500 data

Example 3.15 *(Estimation of the generalized Pareto distribution on S&P daily returns). Figure 3.25 (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.

3.4 Exceedance Correlations*

Reference: Ang and Chen (2002)

It is often argued that most assets are more strongly correlated in down markets than in up markets. If so, diversification may not be such a powerful tool as what we would otherwise believe.

A straightforward way of examining this is to calculate the correlation of two returns(x and y, say) for specific intervals. For instance, we could specify that x_t should be between h_1 and h_2 and y_t between k_1 and k_2

$$\text{Corr}(x_t, y_t | h_1 < x_t \leq h_2, k_1 < y_t \leq k_2). \tag{3.39}$$

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 3.26.

In contrast, Figures 3.27–3.28 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.

3.5 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

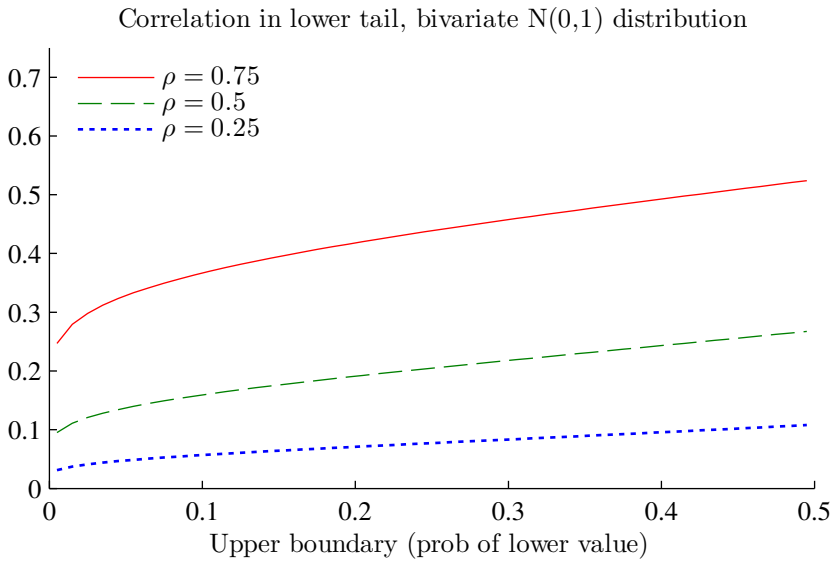


Figure 3.26: Correlation in lower tail when data is drawn from a normal distribution with correlation ρ

in finance, but we sometimes need to go beyond the correlation—to capture non-linear relations. It also turns out to be easier to calibrate/estimate copulas (see below) by using other measures of dependency.

Spearman's rank correlation (called Spearman's rho) 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.

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

$$\begin{array}{cccc} x_t & \text{rank}(x_t) & y_t & \text{rank}(y_t) \\ 2 & 2.5 & 7 & 2 \\ 10 & 4 & 8 & 3 \\ -3 & 1 & 2 & 1 \\ 2 & 2.5 & 10 & 4 \end{array} \tag{3.40}$$

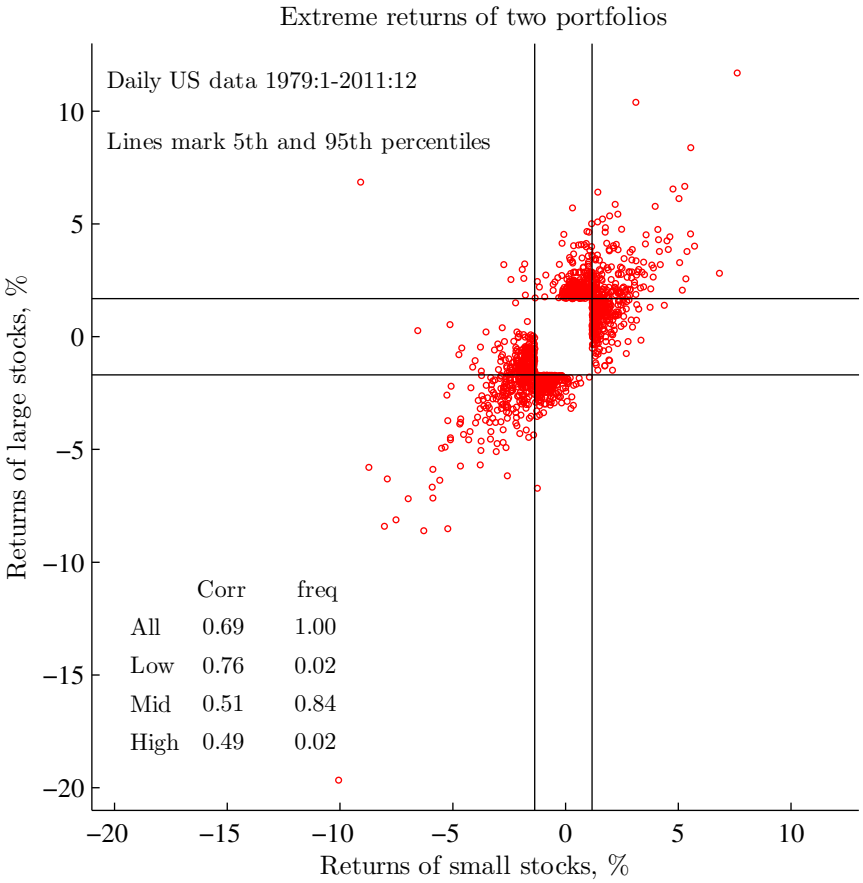


Figure 3.27: Correlation of two portfolios

In the second step, simply estimate the correlation of the ranks of two variables

$$\text{Spearman's } \rho = \text{Corr}[\text{rank}(x_t), \text{rank}(y_t)]. \tag{3.41}$$

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, $d_t = \text{rank}(x_t) - \text{rank}(y_t)$, $\rho = 1 - 6\Sigma_{t=1}^{T} d_t^2/(T^3 - T)$. It gives the same result if there are no tied ranks.) See Figure 3.29 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

$$\sqrt{T-1}\hat{\rho} \to^d N(0,1). \tag{3.42}$$

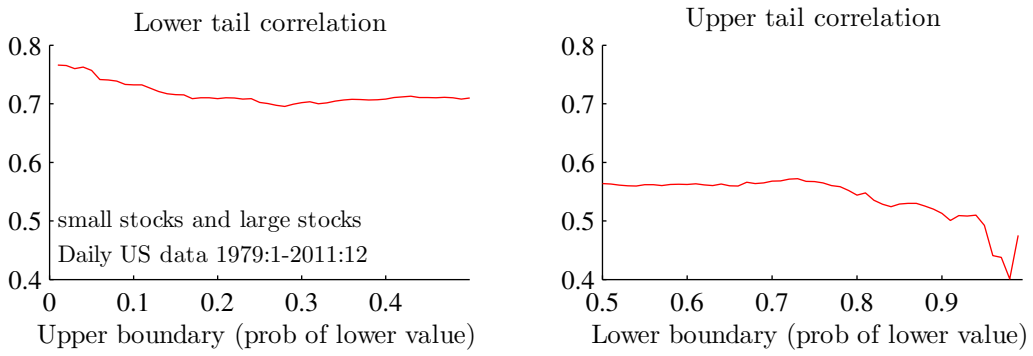


Figure 3.28: Correlation in the tails for two portfolios

(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 3.16 *(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* $\text{Corr}[F_X(X), F_Y(Y)]$ *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

$$\begin{array}{cc} \underline{\text{Changes of } x} & \underline{\text{Changes of } y} \\ x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \\ x_3 - x_2 & y_3 - y_2, \end{array} \tag{3.43}$$

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

$$\begin{aligned} ij \text{ is concordant if } (x_j - x_i)(y_j - y_i) > 0 \\ ij \text{ is discordant if } (x_j - x_i)(y_j - y_i) < 0. \end{aligned} \tag{3.44}$$

Finally, we count the number of concordant (T_c) and discordant (T_d) pairs and calculate

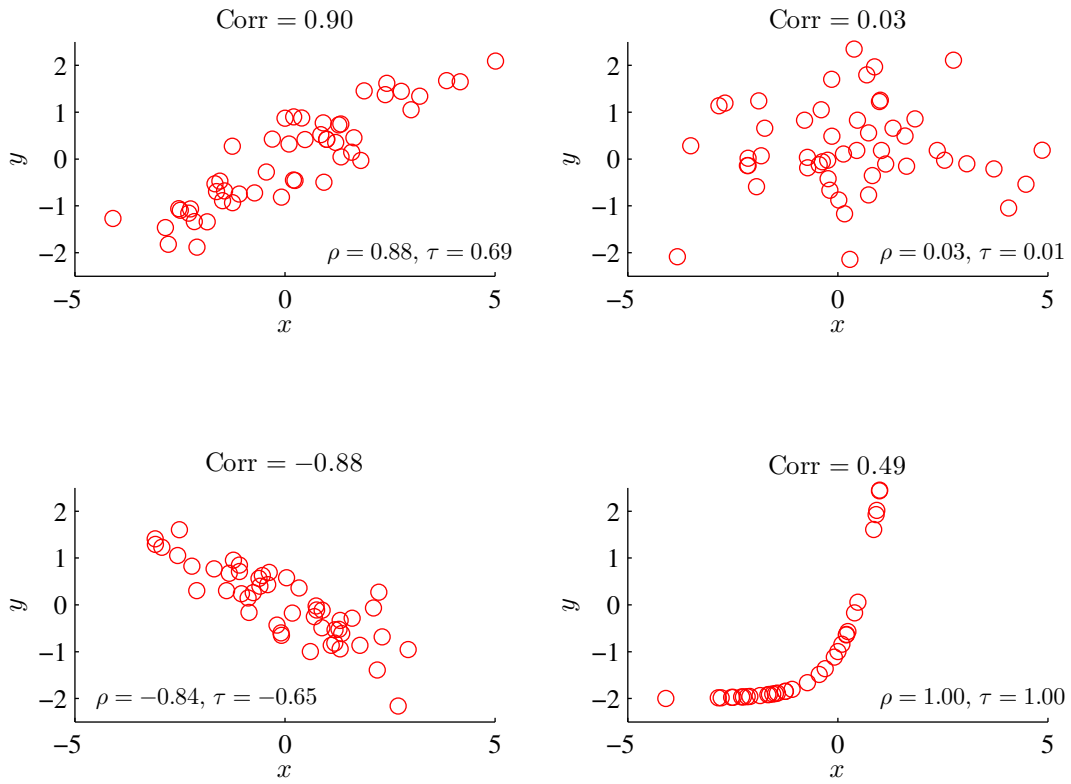


Figure 3.29: Illustration of correlation and rank correlation

Kendall's tau as

$$\text{Kendall's } \tau = \frac{T_c - T_d}{T(T-1)/2}. \tag{3.45}$$

It can be shown that

$$\text{Kendall's } \tau \to^d N\left(0, \frac{4T+10}{9T(T-1)}\right), \tag{3.46}$$

so it is straightforward to test τ by a t-test.

Example 3.17 *(Kendall's tau) Suppose the data is*

x	y
2	7
10	9
−3	10.

We then get the following changes

$$\begin{array}{cc} \underline{\textit{Changes of } x} & \underline{\textit{Changes of } y} \\ x_2 - x_1 = 10 - 2 = 8 & y_2 - y_1 = 9 - 7 = 2 \quad \textit{concordant} \\ x_3 - x_1 = -3 - 2 = -5 & y_3 - y_1 = 10 - 7 = 3 \quad \textit{discordant} \\ x_3 - x_2 = -3 - 10 = -13 & y_3 - y_2 = 10 - 9 = 1, \quad \textit{discordant.} \end{array}$$

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

$$\text{Spearman's rho} = \frac{6}{\pi}\arcsin(\rho/2) \approx \rho \tag{3.47}$$

$$\text{Kendall's tau } = \frac{2}{\pi}\arcsin(\rho). \tag{3.48}$$

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 3.30. Clearly, when data is not normally distributed, then these measures can give distinctly different answers.

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

$$G_\alpha = \Pr(x \le \xi_{x,\alpha}, y \le \xi_{y,\alpha}), \tag{3.49}$$

$\xi_{x,\alpha}$ and $\xi_{y,\alpha}$ are α-quantile of the x- and y-distribution respectively.

In practice, this can be estimated from data by first finding the empirical α-quantiles ($\hat{\xi}_{x,\alpha}$ and $\hat{\xi}_{y,\alpha}$) by simply sorting the data and then picking out the value of observation αT of this sorted list (do this individually for x and y). Then, calculate the estimate

$$\hat{G}_\alpha = \frac{1}{T}\sum_{t=1}^{T}\delta_t, \text{ where} \tag{3.50}$$

$$\delta_t = \begin{cases} 1 \text{ if } x_t \le \hat{\xi}_{x,\alpha} \text{ and } y_t \le \hat{\xi}_{y,\alpha} \\ 0 \text{ otherwise.} \end{cases}$$

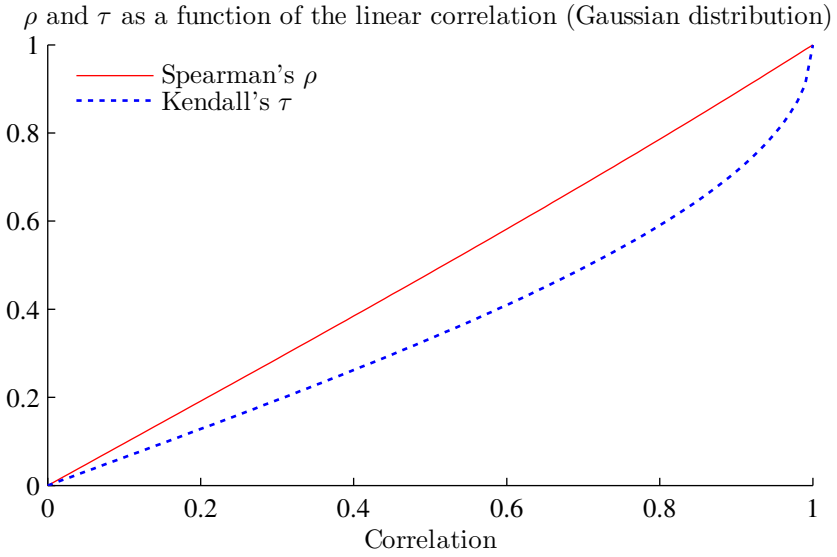


Figure 3.30: Spearman's rho and Kendall's tau if data has a bivariate normal distribution

See Figure 3.31 for an illustration based on a joint normal distribution.

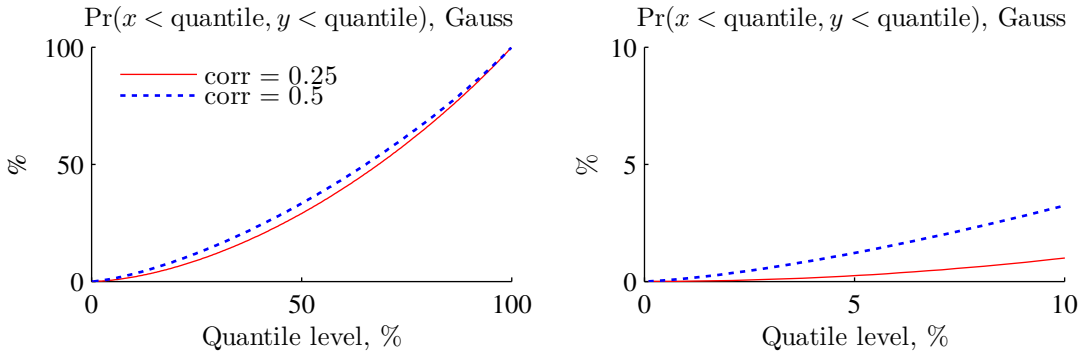


Figure 3.31: Probability of joint low returns, bivariate normal distribution

3.6 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.

3.6.1 Multivariate Distributions and Copulas

Any pdf can also be written as

$$f_{1,2}(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2), \text{ with} \quad (3.51)$$
$$u_i = F_i(x_i),$$

where $c()$ is a *copula density* function and $u_i = F_i(x_i)$ is the cdf value as in (3.1). The extension to three or more random variables is straightforward.

Equation (3.51) 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 3.18 *(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 3.19 *(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

$$F_{1,2}(x_1, x_2) = \int_{s=-\infty}^{x_1} \int_{t=-\infty}^{x_2} f_{1,2}(s,t)dsdt.$$

(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 3.32 for an illustration.

Remark 3.20 *(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 $f_1(x_1) = \int_{x_2=-\infty}^{\infty} f_{1,2}(x_1, x_2)dx_2$.*

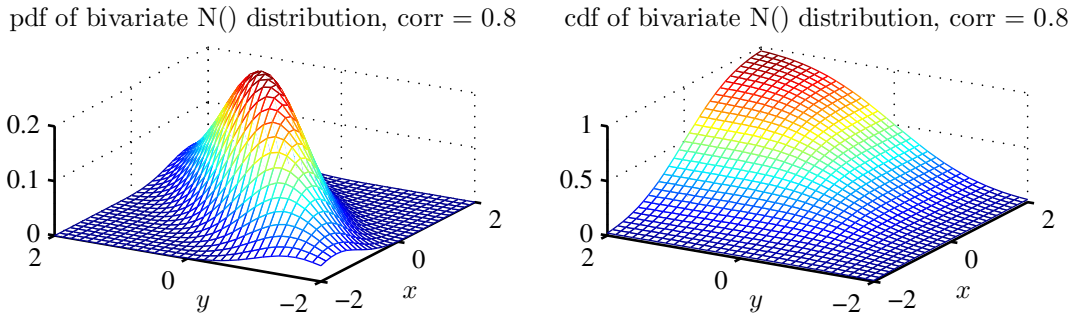


Figure 3.32: Bivariate normal distributions

Remark 3.21 *(Joint pdf and copula density, n variables) For n variables (3.51) 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 3.22 *(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 (3.51) 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 $C()$.

3.6.2 The Gaussian and Other Copula Densities

The *Gaussian copula density function* is

$$c(u_1, u_2) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2\xi_1^2 - 2\rho\xi_1\xi_2 + \rho^2\xi_2^2}{2(1-\rho^2)}\right), \text{ with} \tag{3.52}$$
$$\xi_i = \Phi^{-1}(u_i),$$

where $\Phi^{-1}()$ is the inverse of an $N(0,1)$ distribution. Notice that when using this function in (3.51) 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 3.33 for an illustration.

A zero correlation ($\rho = 0$) makes the copula density (3.52) 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

$$\rho = \text{Spearman's rho}, \tag{3.53}$$

as suggested by (3.47).

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 (3.50). 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 3.31 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 3.23 *(The Gaussian copula function*) The distribution function corresponding to the Gaussian copula density (3.52) 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 (3.52) 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 3.24 *(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

$$c(u_1, u_2) = (-1 + u_1^{-\alpha} + u_2^{-\alpha})^{-2-1/\alpha}(u_1 u_2)^{-\alpha-1}(1 + \alpha), \tag{3.54}$$

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 3.33 for an illustration.

For the Clayton copula we have

$$\text{Kendall's } \tau = \frac{\alpha}{\alpha + 2}, \text{ so} \tag{3.55}$$

$$\alpha = \frac{2\tau}{1 - \tau}. \tag{3.56}$$

The easiest way to calibrate a Clayton copula is therefore to set the parameter α according to (3.56).

Figure 3.34 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 3.31 for the Gaussian copula, see (3.47)–(3.48). 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 3.25 *(Multivariate Clayton copula density*) The Clayton copula density for n variables is*

$$c(u) = \left(1 - n + \sum_{i=1}^{n} u_i^{-\alpha}\right)^{-n-1/\alpha} \left(\prod_{i=1}^{n} u_i\right)^{-\alpha-1} \left(\prod_{i=1}^{n} [1 + (i-1)\alpha]\right).$$

Remark 3.26 *(Clayton copula function*) The copula function (the cdf) corresponding to (3.54) 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 (3.1).

2. Calculate the copula density as follows (for the Gaussian or Clayton copulas, respectively):

   (a) for the Gaussian copula (3.52)

      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 (3.54)

      i. assume (or estimate/calibrate) an α to use in the Clayton copula (typically based on Kendall's τ as in (3.56))

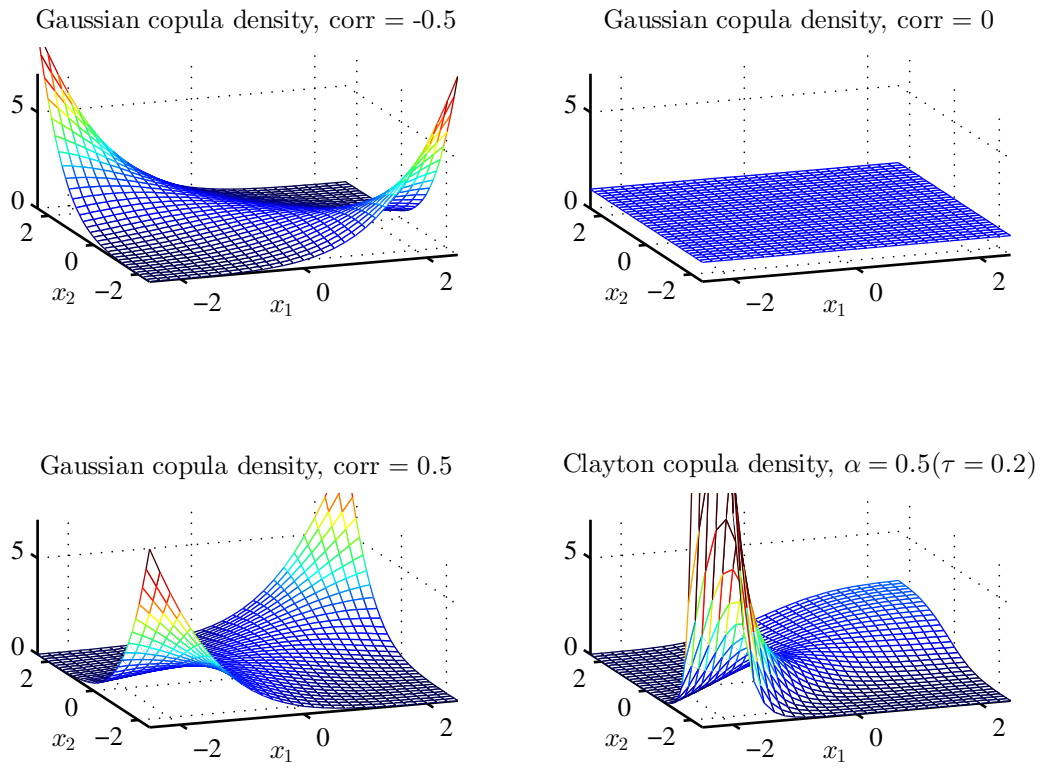


Figure 3.33: Copula densities (as functions of x_i)

ii. calculate the copula density value $c(u_1, u_2)$

3. Combine the marginal pdfs and the copula density as in (3.51), $f_{1,2}(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2)$, where $u_i = F_i(x_i)$ is the cdf value according to the marginal distribution of variable i.

See Figures 3.35–3.36 for illustrations.

Remark 3.27 *(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 ($X_2 \leq F_2^{-1}(q)$)*

$$\Lambda_l = \Pr[X_1 \leq F_1^{-1}(q) | X_2 \leq F_2^{-1}(q)],$$

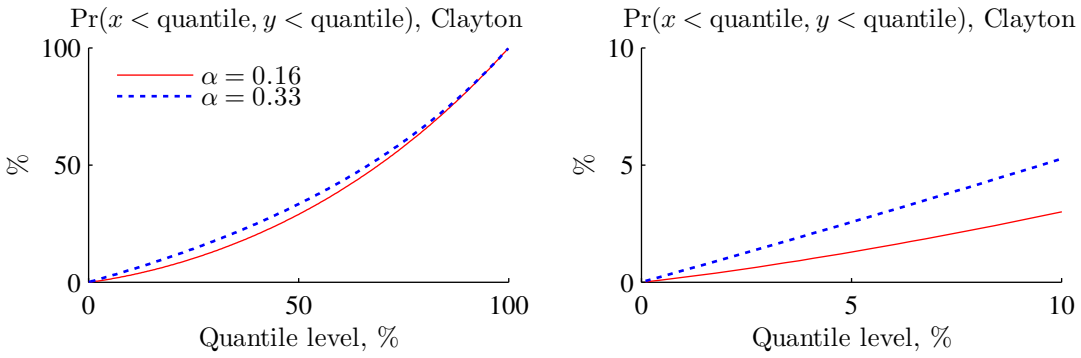


Figure 3.34: Probability of joint low returns, Clayton copula

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.

3.7 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 3.37–3.39 provide an illustration.

Remark 3.28 *Figure 3.37 suggests that the joint occurrence (of these two assets) of re-*

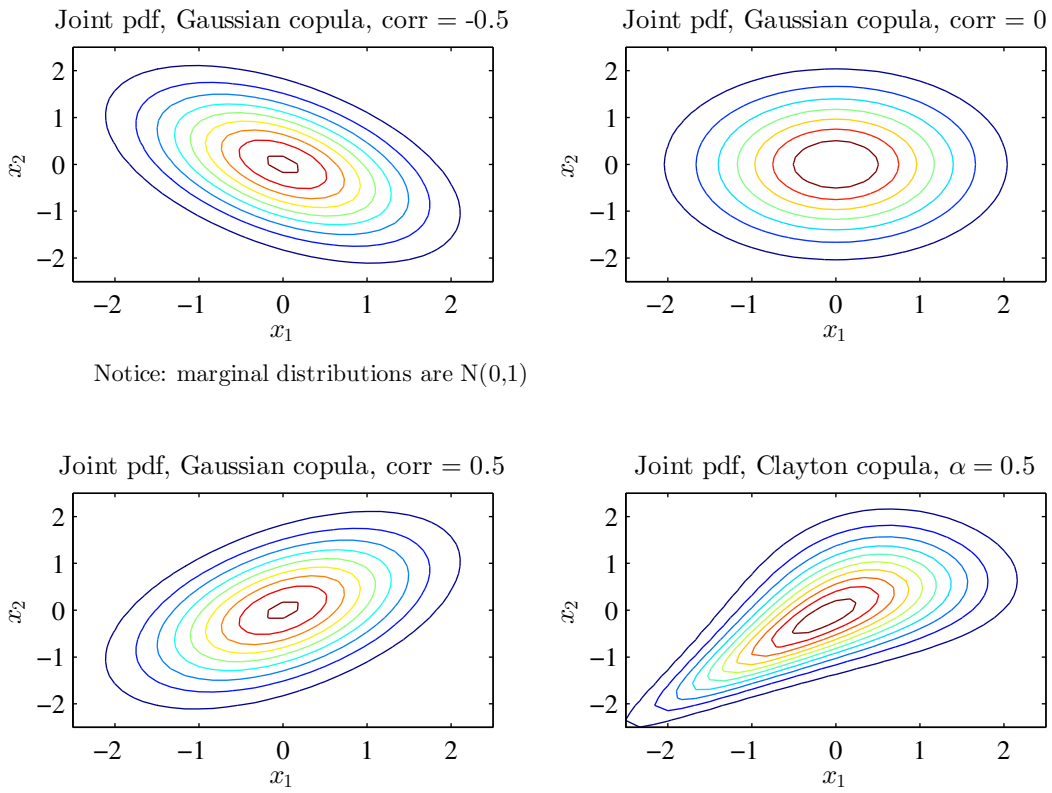


Figure 3.35: Contours of bivariate pdfs

ally negative returns happens more often than the estimated normal distribution would 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 (3.39) 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 $u_i =$

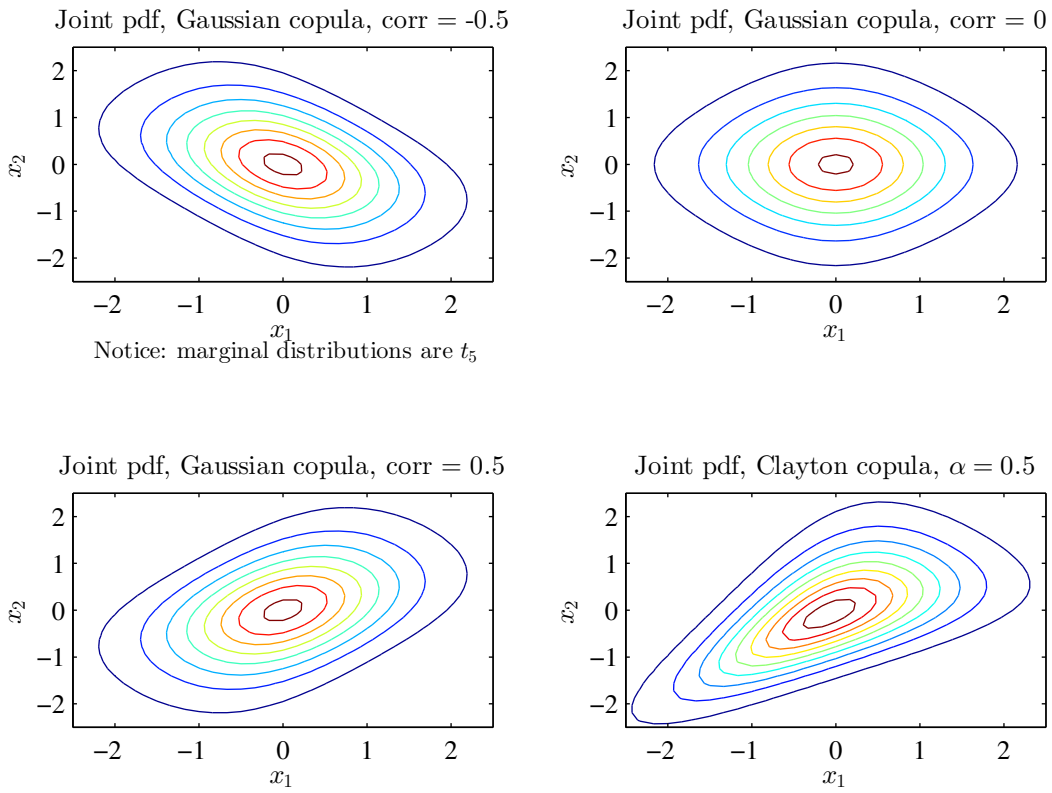


Figure 3.36: Contours of bivariate pdfs

$F_i(x_i)$ in (3.51) as

$$x_i = F_i^{-1}(u_i), \tag{3.57}$$

where $F_i^{-1}()$ is the inverse of the cdf.

Remark 3.29 *(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 3.30 *(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 3.31 *(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*

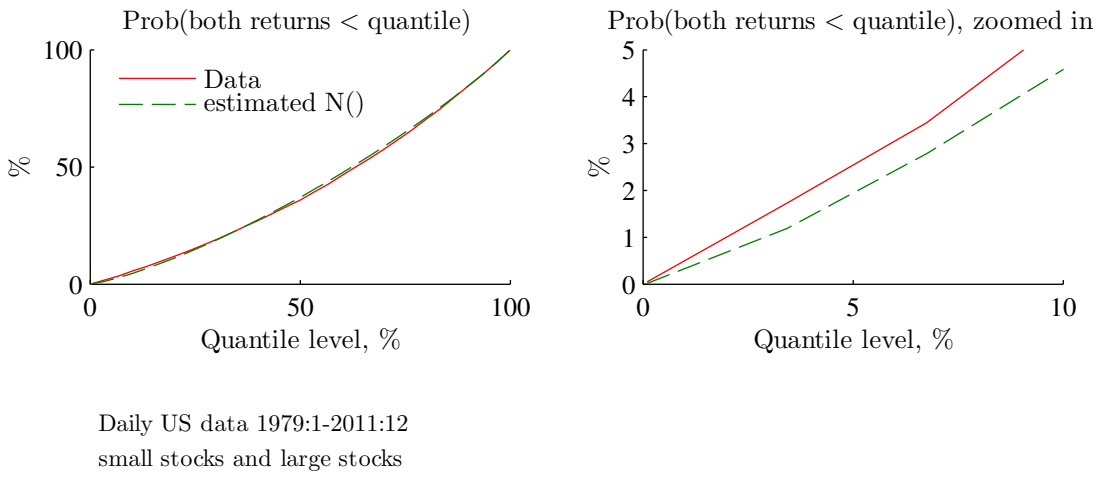


Figure 3.37: Probability of joint low returns

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, 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 3.40–3.41 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.

Breeden, D., and R. Litzenberger, 1978, "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51, 621–651.

Cox, J. C., and S. A. Ross, 1976, "The Valuation of Options for Alternative Stochastic Processes," *Journal of Financial Economics*, 3, 145–166.

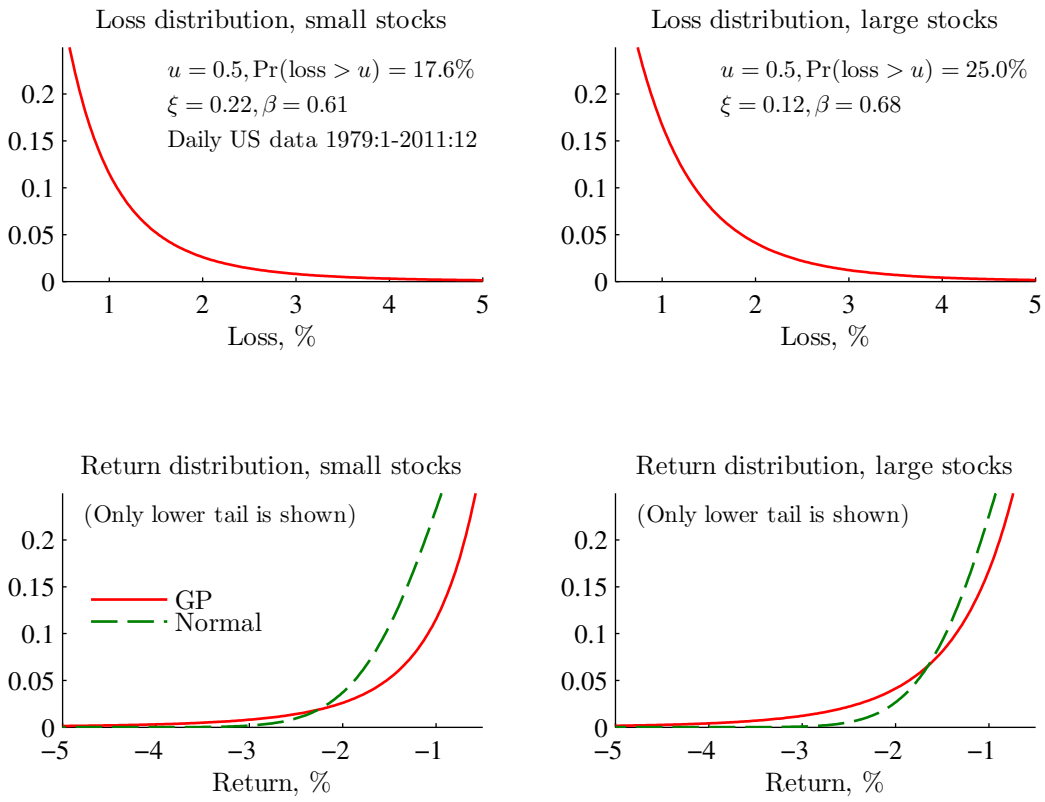


Figure 3.38: Estimation of marginal loss distributions

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.

Jackwerth, J. C., 2000, "Recovering risk aversion from option prices and realized returns," *Review of Financial Studies*, 13, 433–451.

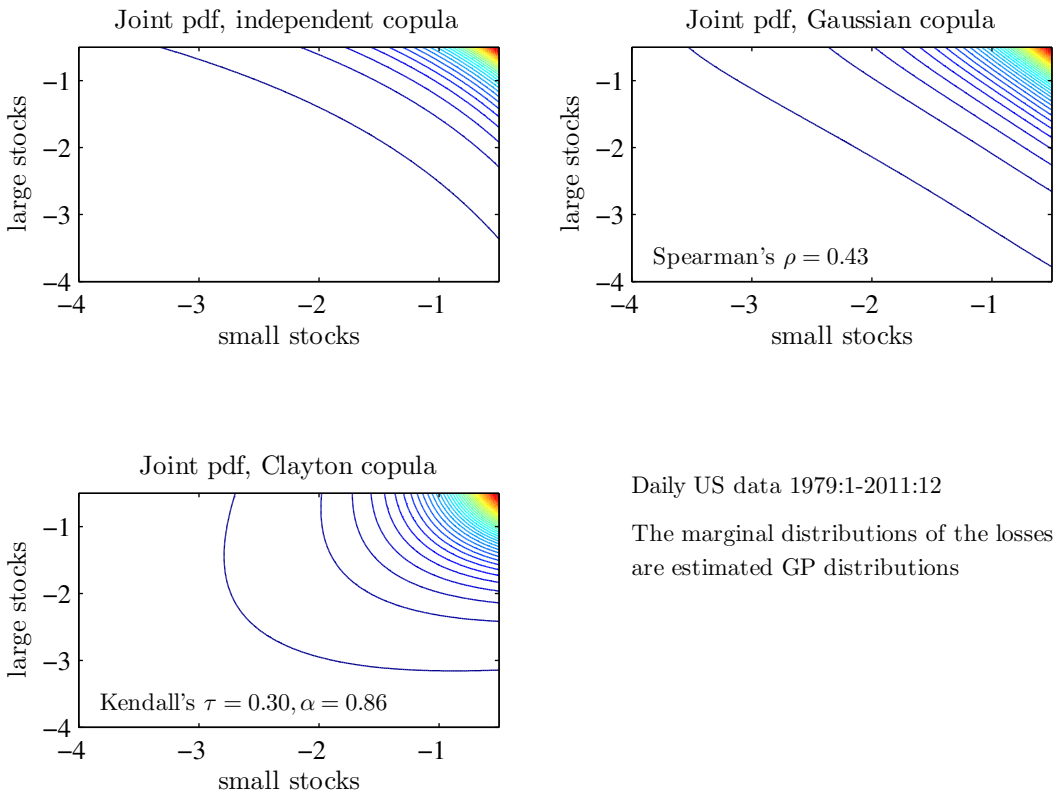


Figure 3.39: Joint pdfs with different copulas

Jondeau, E., S.-H. Poon, and M. Rockinger, 2007, *Financial Modeling under Non-Gaussian Distributions*, Springer.

Lo, A. W., H. Mamaysky, and J. Wang, 2000, "Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation," *Journal of Finance*, 55, 1705–1765.

McNeil, A. J., R. Frey, and P. Embrechts, 2005, *Quantitative risk management*, Princeton University Press.

Melick, W. R., and C. P. Thomas, 1997, "Recovering an Asset's Implied PDF from Options Prices: An Application to Crude Oil During the Gulf Crisis," *Journal of Financial and Quantitative Analysis*, 32, 91–115.

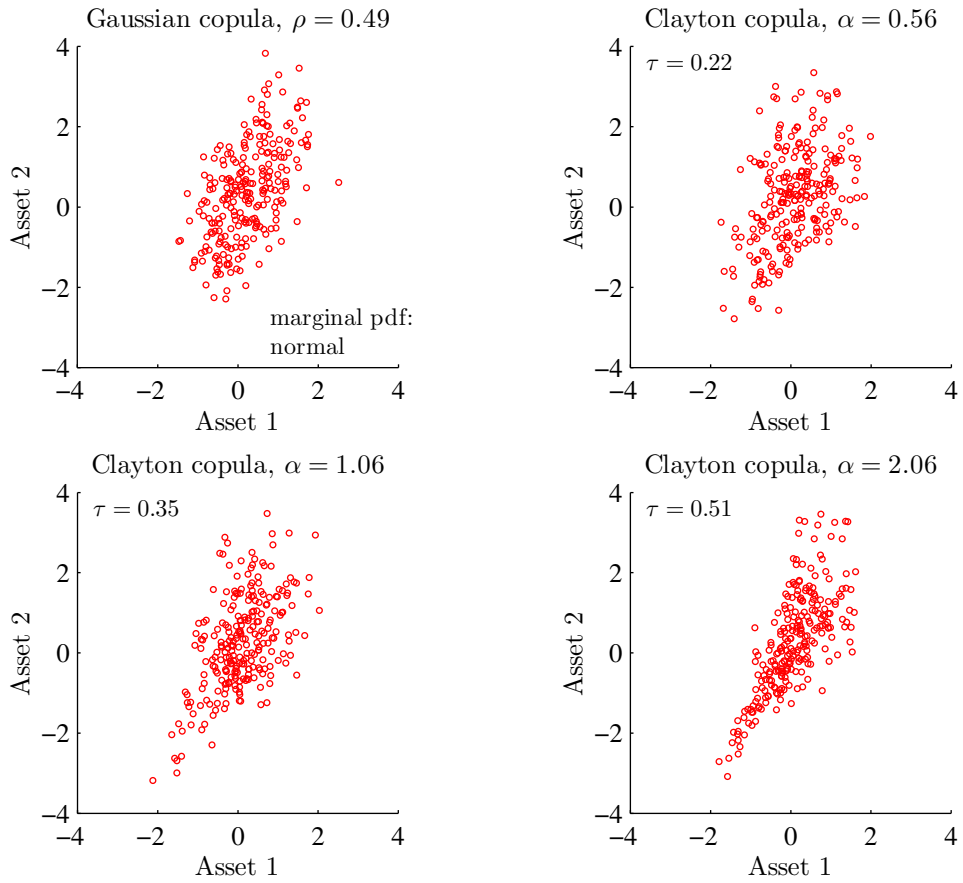


Figure 3.40: Example of scatter plots of two asset returns drawn from different copulas

Mittelhammer, R. C., 1996, *Mathematical statistics for economics and business*, Springer-Verlag, New York.

Ritchey, R. J., 1990, "Call option valuation for discrete normal mixtures," *Journal of Financial Research*, 13, 285–296.

Silverman, B. W., 1986, *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Söderlind, P., 2000, "Market expectations in the UK before and after the ERM crisis," *Economica*, 67, 1–18.

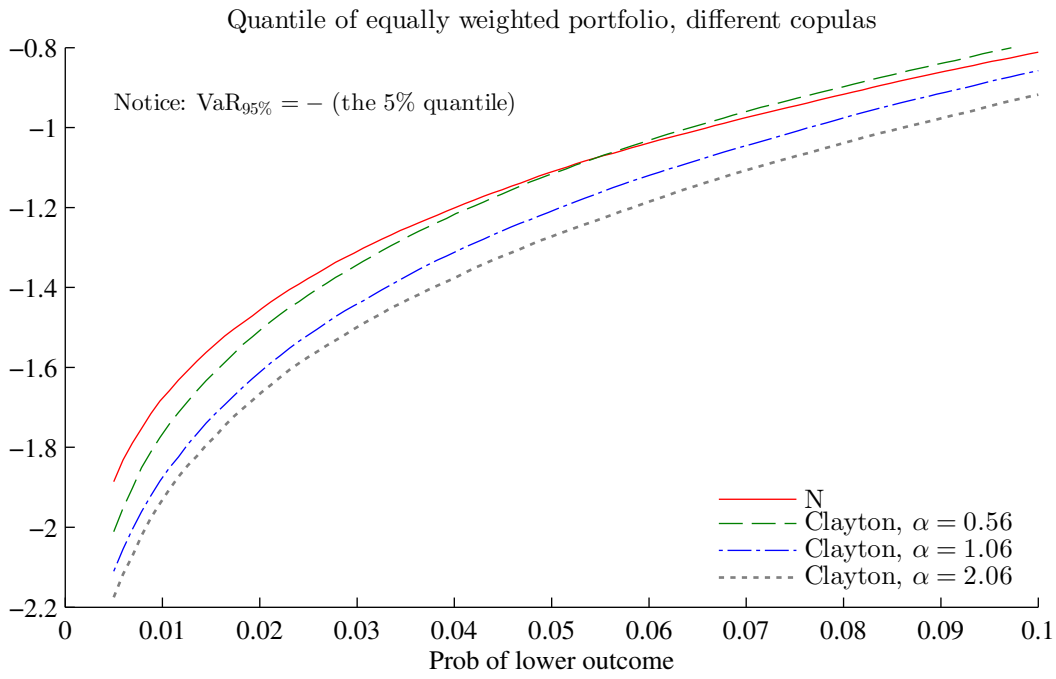


Figure 3.41: Quantiles of an equally weighted portfolio of two asset returns drawn from different copulas

Söderlind, P., and L. E. O. Svensson, 1997a, "New techniques to extract market expectations from financial instruments," *Journal of Monetary Economics*, 40, 383–420.

Söderlind, P., and L. E. O. Svensson, 1997b, "New techniques to extract market expectations from financial instruments," *Journal of Monetary Economics*, 40, 383–429.

Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.

4 Predicting Asset Returns

Sections denoted by a star (*) is not required reading.

Reference: Cochrane (2005) 20.1; Campbell, Lo, and MacKinlay (1997) 2 and 7; Taylor (2005) 5–7

4.1 A Little Financial Theory and Predictability

The traditional interpretation of autocorrelation in asset returns is that there are some "irrational traders." For instance, feedback trading would create positive short term autocorrelation in returns. If there are non-trivial market imperfections, then predictability can be used to generate economic profits. If there are no important market imperfections, then predictability of excess returns should be thought of as predictable movements in risk premia.

To see illustrate the latter, let R^e_{t+1} be the excess return on an asset. The canonical asset pricing equation then says

$$\mathrm{E}_t\, m_{t+1} R^e_{t+1} = 0, \tag{4.1}$$

where m_{t+1} is the stochastic discount factor.

Remark 4.1 *(A consumption-based model) Suppose we want to maximize the expected discounted sum of utility* $\mathrm{E}_t\, \Sigma_{s=0}^{\infty} \beta^s u(c_{t+s})$. *Let* Q_t *be the consumer price index in* t. *Then, we have*

$$m_{t+1} = \begin{cases} \beta \frac{u'(c_{t+1})}{u'(c_t)} \frac{Q_t}{Q_{t+1}} & \textit{if returns are nominal} \\ \beta \frac{u'(c_{t+1})}{u'(c_t)} & \textit{if returns are real.} \end{cases}$$

We can rewrite (4.1) (using $\mathrm{Cov}(x, y) = \mathrm{E}\, xy - \mathrm{E}\, x\, \mathrm{E}\, y$) as

$$\mathrm{E}_t\, R^e_{t+1} = -\mathrm{Cov}_t(m_{t+1}, R^e_{t+1}) / \mathrm{E}_t\, m_{t+1}. \tag{4.2}$$

This says that the expected excess return will vary if risk (the covariance) does. If there is some sort of reasonable relation between beliefs and the properties of actual returns (not

necessarily full rationality), then we should not be too surprised to find predictability.

Example 4.2 *(Epstein-Zin utility function) Epstein and Zin (1991) define a certainty equivalent of future utility as* $Z_t = [\mathrm{E}_t(U_{t+1}^{1-\gamma})]^{1/(1-\gamma)}$ *where* γ *is the risk aversion—and then use a CES aggregator function to govern the intertemporal trade-off between current consumption and the certainty equivalent:* $U_t = [(1-\delta)C_t^{1-1/\psi} + \delta Z_t^{1-1/\psi}]^{1/(1-1/\psi)}$ *where* ψ *is the elasticity of intertemporal substitution. If returns are iid (so the consumption-wealth ratio is constant), then it can be shown that this utility function has the same pricing implications as the CRRA utility, that is,*

$$\mathrm{E}[(C_t/C_{t-1})^{-\gamma} R_t] = \textit{constant.}$$

(See Söderlind (2006) for a simple proof.)

Example 4.3 *(Portfolio choice with predictable returns) Campbell and Viceira (1999) specify a model where the log return of the only risky asset follows the time series process*

$$r_{t+1} = r_f + x_t + u_{t+1},$$

where r_f *is a constant riskfree rate,* u_{t+1} *is unpredictable, and the state variable follows (constant suppressed)*

$$x_{t+1} = \phi x_t + \eta_{t+1},$$

where η_{t+1} *is also unpredictable. Clearly,* $\mathrm{E}_t(r_{t+1} - r_f) = x_t$. $\mathrm{Cov}_t(u_{t+1}, \eta_{t+1})$ *can be non-zero. For instance, with* $\mathrm{Cov}_t(u_{t+1}, \eta_{t+1}) < 0$*, a high return (*$u_{t+1} > 0$*) is typically associated with an expected low future return (*x_{t+1} *is low since* $\eta_{t+1} < 0$*) With Epstein-Zin preferences, the portfolio weight on the risky asset is (approximately) of the form*

$$v_t = a_0 + a_1 x_t,$$

where a_0 *and* a_1 *are complicated expression (in terms of the model parameters—can be calculated numerically). There are several interesting results. First, if returns are not predictable (*x_t *is constant since* η_{t+1} *is), then the portfolio choice is constant. Second, when returns are predictable, but the relative risk aversion is unity (no intertemporal hedging), then* $v_t = 1/(2\gamma) + x_t/[\gamma \,\mathrm{Var}_t(u_{t+1})]$*. Third, with a higher risk aversion and* $\mathrm{Cov}_t(u_{t+1}, \eta_{t+1}) < 0$*, there is a positive hedging demand for the risky asset: it pays off (today) when the future investment opportunities are poor.*

Example 4.4 *(Habit persistence) The habit persistence model of Campbell and Cochrane (1999) has a CRRA utility function, but the argument is the difference between consumption and a habit level, $C_t - X_t$, instead of just consumption. The habit is parameterized in terms of the "surplus ratio" $S_t = (C_t - X_t)/C_t$. The log surplus ratio.(s_t)is assumed to be a non-linear AR(1)*

$$s_t = \phi s_{t-1} + \lambda(s_{t-1})\Delta c_t.$$

It can be shown (see Söderlind (2006)) that if $\lambda(s_{t-1})$ is a constant λ and the excess return is unpredictable (by s_t) then the habit persistence model is virtually the same as the CRRA model, but with $\gamma(1+\lambda)$ as the "effective" risk aversion.

Example 4.5 *(Reaction to news and the autocorrelation of returns) Let the log asset price, p_t, be the sum of a random walk and a temporary component (with perfectly correlated innovations, to make things simple)*

$$\begin{aligned} p_t &= u_t + \theta\varepsilon_t, \text{ where } u_t = u_{t-1} + \varepsilon_t \\ &= u_{t-1} + (1+\theta)\varepsilon_t. \end{aligned}$$

Let $r_t = p_t - p_{t-1}$ be the log return. It is straightforward to calculate that

$$\text{Cov}(r_{t+1}, r_t) = -\theta(1+\theta)\,\text{Var}(\varepsilon_t),$$

so $0 < \theta < 1$ (initial overreaction of the price) gives a negative autocorrelation. See Figure 4.1 for the impulse responses with respect to a piece of news, ε_t.

4.2 Autocorrelations

Reference: Campbell, Lo, and MacKinlay (1997) 2

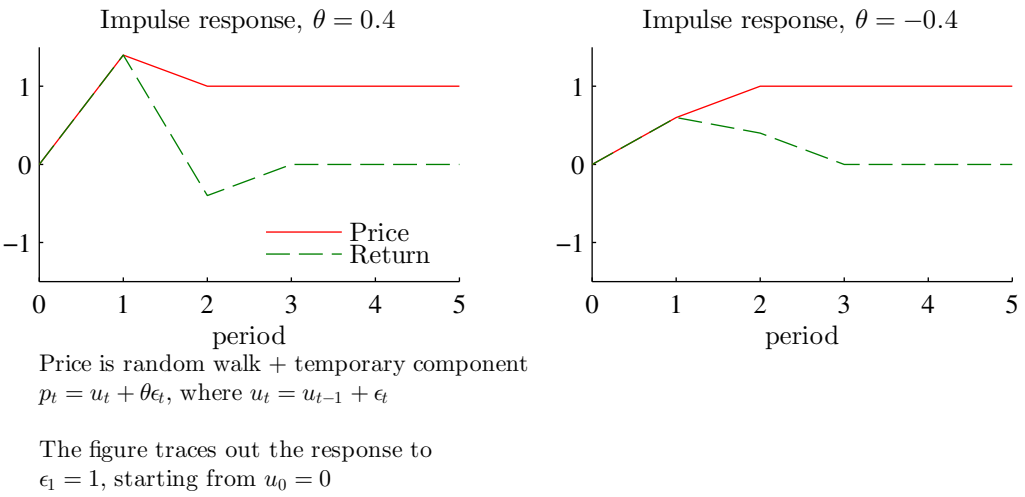


Figure 4.1: Impulse reponses when price is random walk plus temporary component

4.2.1 Autocorrelation Coefficients and the Box-Pierce Test

The autocovariances of the r_t process can be estimated as

$$\hat{\gamma}_s = \frac{1}{T}\sum_{t=1+s}^{T}(r_t - \bar{r})(r_{t-s} - \bar{r})', \tag{4.3}$$

$$\text{with } \bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t. \tag{4.4}$$

(We typically divide by T even though there are only $T - s$ observations to estimate γ_s from.) Autocorrelations are then estimated as

$$\hat{\rho}_s = \hat{\gamma}_s/\hat{\gamma}_0. \tag{4.5}$$

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

$$\sqrt{T}\begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \to^d N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right). \tag{4.6}$$

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 4.6 *(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 $|\hat{\rho}_1| > 1.65/\sqrt{1000} \approx 0.053$.*

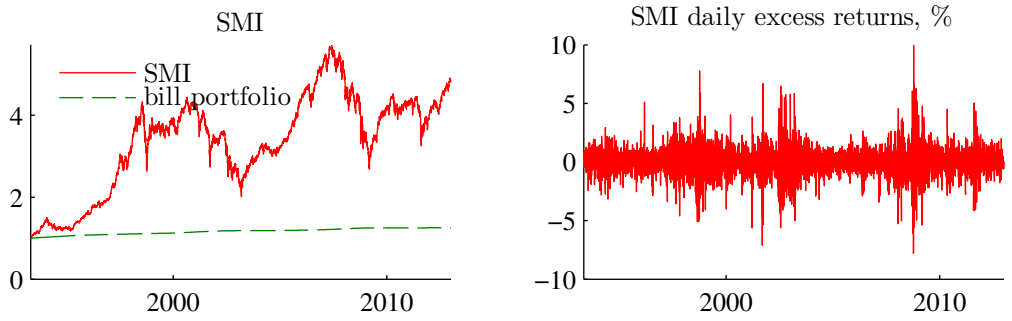
The *Box-Pierce test* follows directly from the result in (4.6), since it shows that $\sqrt{T}\hat{\rho}_i$ and $\sqrt{T}\hat{\rho}_j$ are iid N(0,1) variables. Therefore, the sum of the square of them is distributed as an χ^2 variable. The test statistic typically used is

$$Q_L = T\sum_{s=1}^{L}\hat{\rho}_s^2 \to^d \chi_L^2. \tag{4.7}$$

Example 4.7 *(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 (4.7), L, should be guided by theoretical considerations, but it may also be wise to try different values. There is clearly a trade off: too few lags may miss a significant high-order autocorrelation, but too many lags can destroy the power of the test (as the test statistic is not affected much by increasing L, but the critical values increase).

The main problem with these tests is that the assumptions behind the results in (4.6) may not be reasonable. For instance, data may be heteroskedastic. One way of handling these issues is to make use of the GMM framework. (Alternatively, the results in Taylor (2005) are useful.) Moreover, care must be taken so that for, instance, time aggregation doesn't introduce serial correlation.



Daily SMI data, 1993:5-2012:12
1st order autocorrelation of returns (daily, weekly, monthly): 0.03 -0.11 0.04
1st order autocorrelation of absolute returns (daily, weekly, monthly): 0.29 0.31 0.19

Figure 4.2: Time series properties of SMI

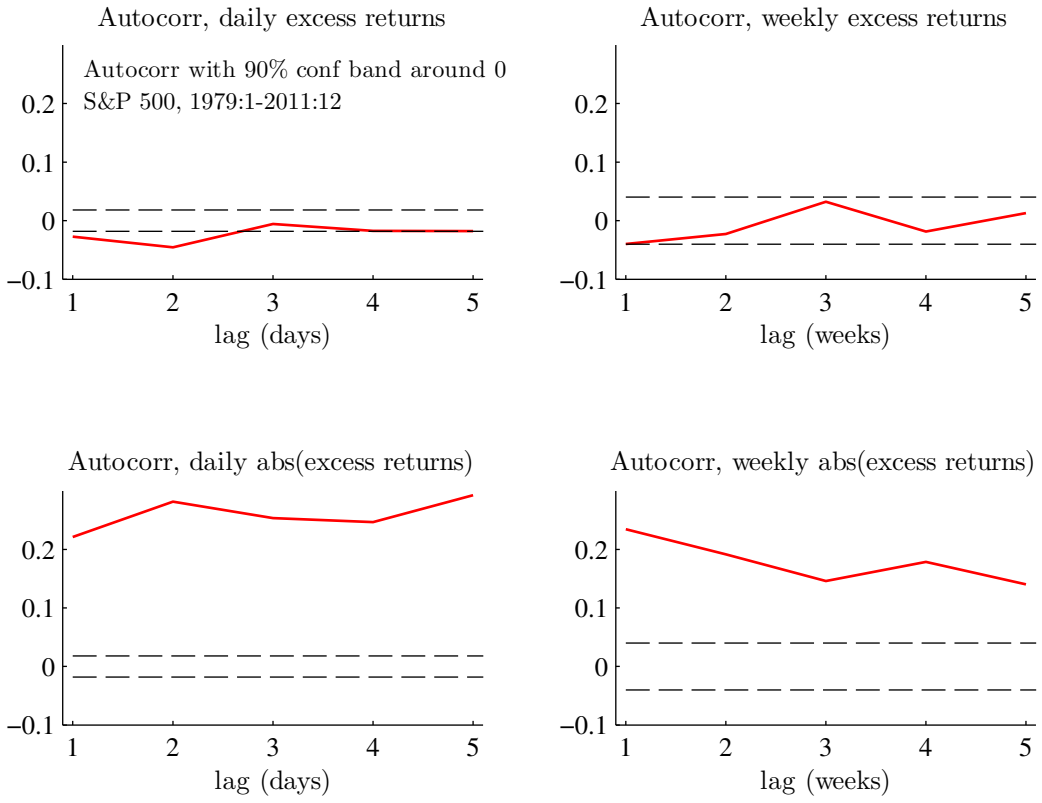


Figure 4.3: Predictability of US stock returns

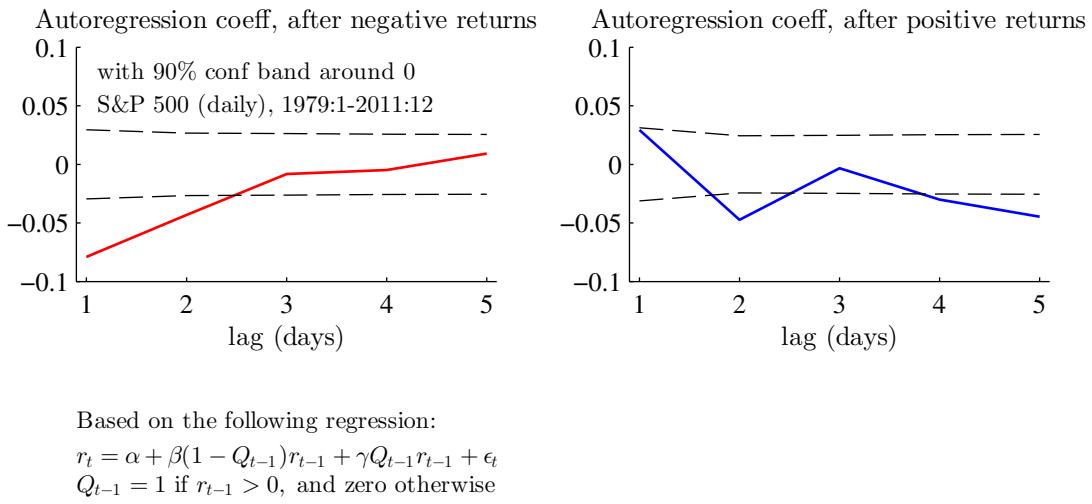


Figure 4.4: Predictability of US stock returns, results from a regression with interactive dummies

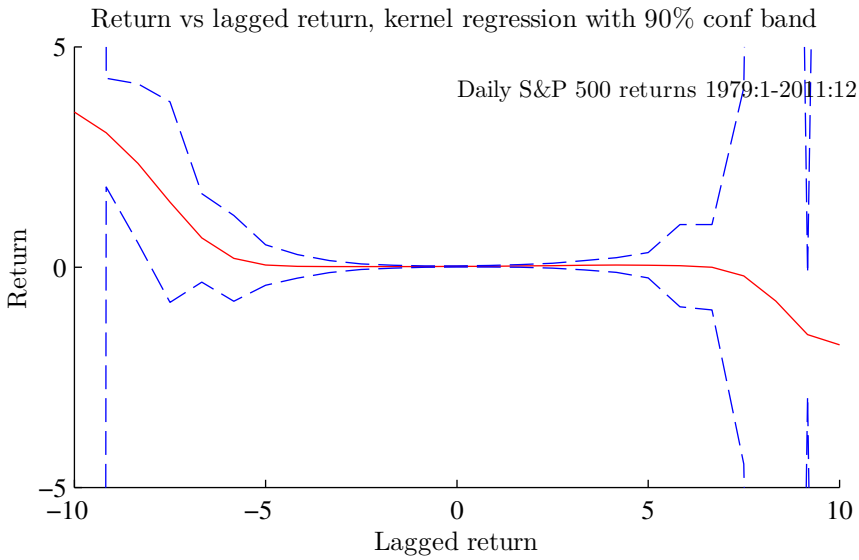


Figure 4.5: Non-parametric regression with confidence bands

4.2.2 GMM Test of Autocorrelation*

This section discusses how GMM can be used to test if a series is autocorrelated. The analysis focuses on first-order autocorrelation, but it is straightforward to extend it to

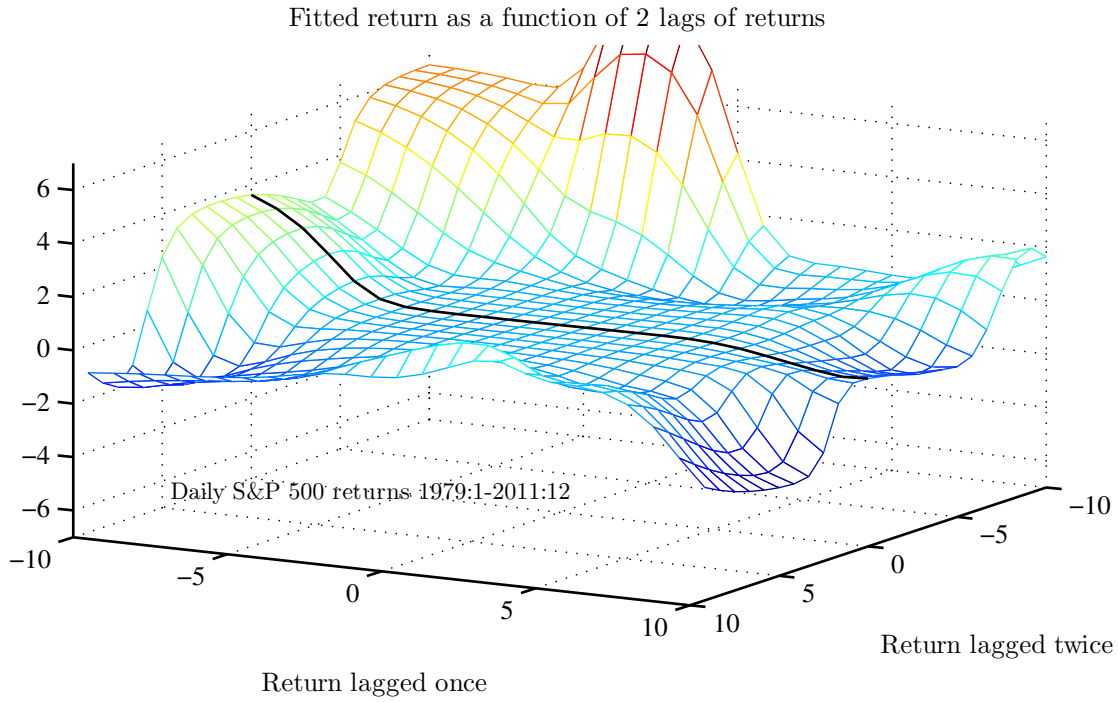


Figure 4.6: Non-parametric regression with two regressors

higher-order autocorrelation.

Consider a scalar random variable x_t with a zero mean (it is easy to extend the analysis to allow for a non-zero mean). Consider the moment conditions

$$g_t(\beta) = \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ so } \bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ with } \beta = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}. \tag{4.8}$$

σ^2 is the variance and ρ the first-order autocorrelation so $\rho\sigma^2$ is the first-order autocovariance. We want to test if $\rho = 0$. We could proceed along two different routes: estimate ρ and test if it is different from zero or set ρ to zero and then test overidentifying restrictions.

We are able to arrive at simple expressions for these tests—provided we are willing to make strong assumptions about the data generating process. (These tests then typically coincide with classical tests like the Box-Pierce test.) One of the strong points of GMM is that we could perform similar tests without making strong assumptions—provided we use a correct estimator of the asymptotic covariance matrix of the moment conditions.

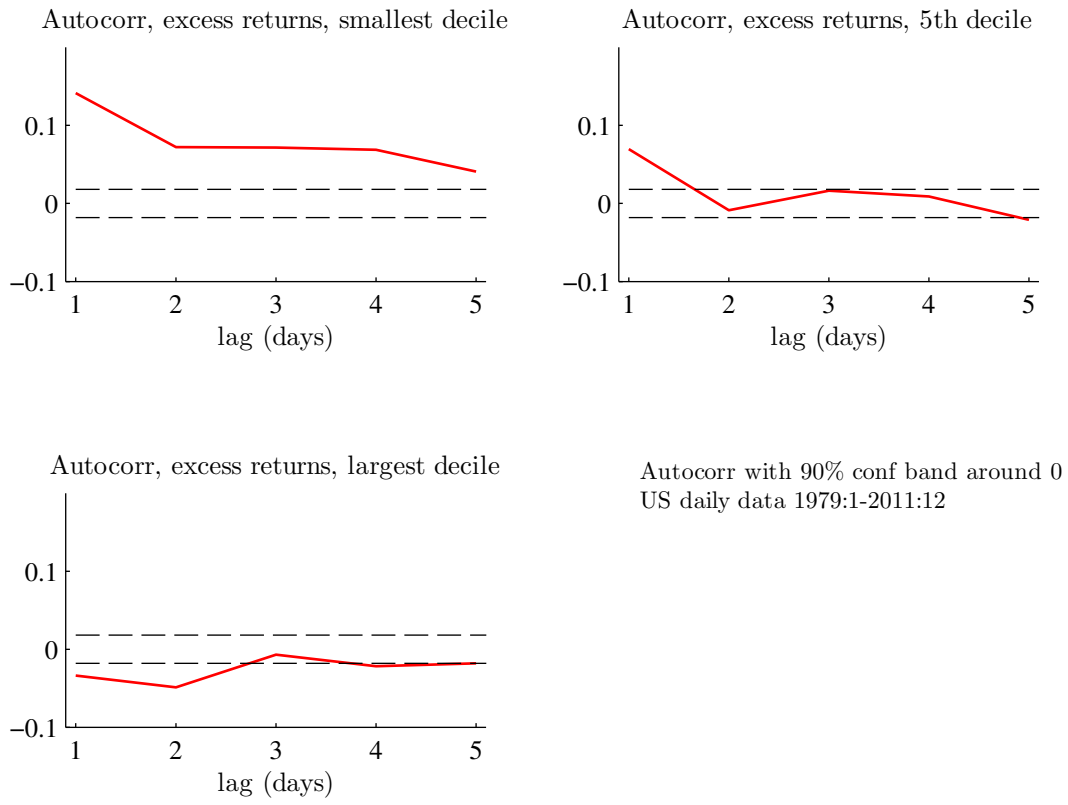


Figure 4.7: Predictability of US stock returns, size deciles

Remark 4.8 *(Box-Pierce as an Application of GMM) (4.8) is an exactly identified system so the weight matrix does not matter, so the asymptotic distribution is*

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), \text{ where } V = \left(D_0' S_0^{-1} D_0\right)^{-1},$$

where D_0 is the Jacobian of the moment conditions and S_0 the covariance matrix of the moment conditions (at the true parameter values). We have

$$D_0 = \operatorname{plim} \begin{bmatrix} \partial \bar{g}_1(\beta_0)/\partial\sigma^2 & \partial \bar{g}_1(\beta_0)/\partial\rho \\ \partial \bar{g}_2(\beta_0)/\partial\sigma^2 & \partial \bar{g}_2(\beta_0)/\partial\rho \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -\rho & -\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix},$$

since $\rho = 0$ (the true value). The definition of the covariance matrix is

$$S_0 = \mathrm{E}\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} g_t(\beta_0)\right]\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} g_t(\beta_0)\right]'.$$

Assume that there is no autocorrelation in $g_t(\beta_0)$ (which means, among other things, that volatility, x_t^2, is not autocorrelated). We can then simplify as

$$S_0 = \mathrm{E}\, g_t(\beta_0) g_t(\beta_0)'.$$

This assumption is stronger than assuming that $\rho = 0$, but we make it here in order to illustrate the asymptotic distribution. Moreover, assume that x_t is iid $N(0, \sigma^2)$. In this case (and with $\rho = 0$ imposed) we get

$$\begin{aligned} S_0 &= \mathrm{E} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}' = \mathrm{E} \begin{bmatrix} (x_t^2 - \sigma^2)^2 & (x_t^2 - \sigma^2) x_t x_{t-1} \\ (x_t^2 - \sigma^2) x_t x_{t-1} & (x_t x_{t-1})^2 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\, x_t^4 - 2\sigma^2 \mathrm{E}\, x_t^2 + \sigma^4 & 0 \\ 0 & \mathrm{E}\, x_t^2 x_{t-1}^2 \end{bmatrix} = \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}. \end{aligned}$$

To make the simplification in the second line we use the facts that $\mathrm{E}\, x_t^4 = 3\sigma^4$ if $x_t \sim N(0, \sigma^2)$, and that the normality and the iid properties of x_t together imply $\mathrm{E}\, x_t^2 x_{t-1}^2 = \mathrm{E}\, x_t^2 \mathrm{E}\, x_{t-1}^2$ and $\mathrm{E}\, x_t^3 x_{t-1} = \mathrm{E}\, \sigma^2 x_t x_{t-1} = 0$. Combining gives

$$\begin{aligned} \mathrm{Cov}\left(\sqrt{T} \begin{bmatrix} \hat{\sigma}^2 \\ \hat{\rho} \end{bmatrix} \right) &= \left(D_0' S_0^{-1} D_0 \right)^{-1} \\ &= \left(\begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix}' \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

This shows that $\sqrt{T}\hat{\rho} \to^d N(0, 1)$.

4.2.3 Autoregressions

An alternative way of testing autocorrelations is to estimate an AR model

$$r_t = c + a_1 r_{t-1} + a_2 r_{t-2} + ... + a_p r_{t-p} + \varepsilon_t, \tag{4.9}$$

and then test if all the slope coefficients are zero with a χ^2 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. To account for heteroskedasticity and other problems, it can make sense to estimate the covariance matrix of the parameters by an estimator like Newey-West.

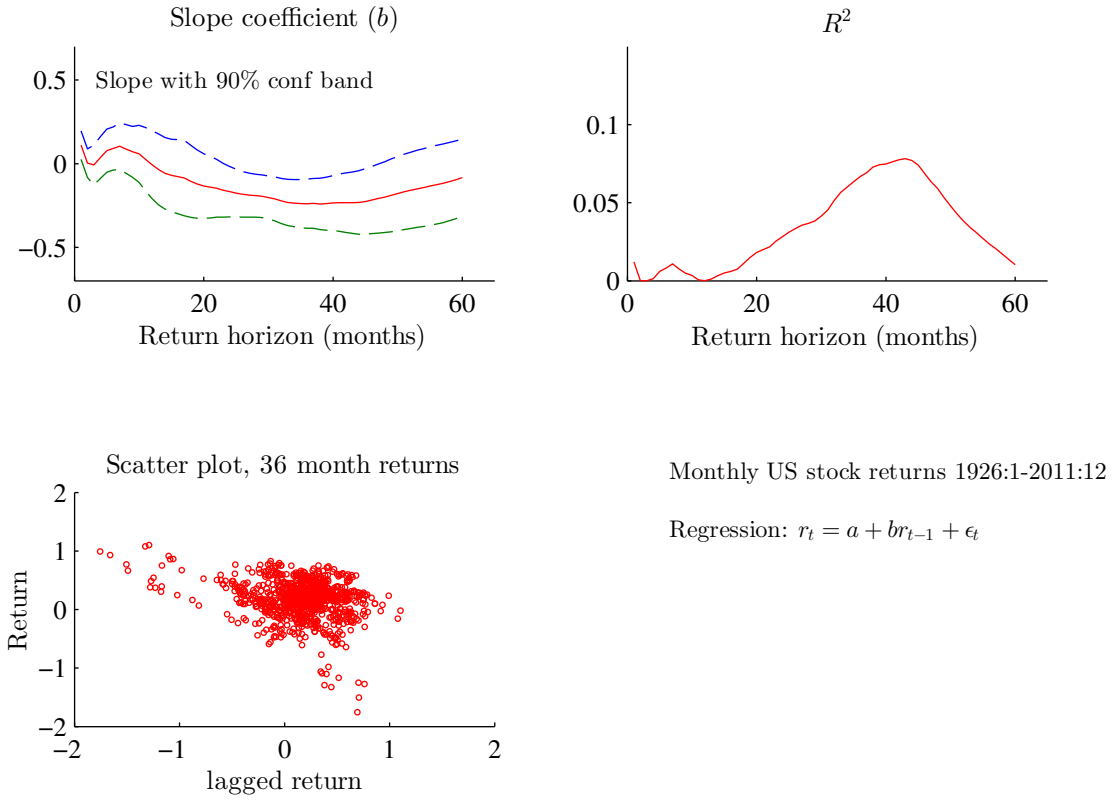


Figure 4.8: Predictability of US stock returns

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

$$r_t = c + a\delta(r_{t-1} \leq 0)r_{t-1} + b\delta(r_{t-1} > 0)r_{t-1}, \text{ where } \delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases} \tag{4.10}$$

See Figure 4.4 for an illustration. Also see Figures 4.5–4.6 for non-parametric estimates.

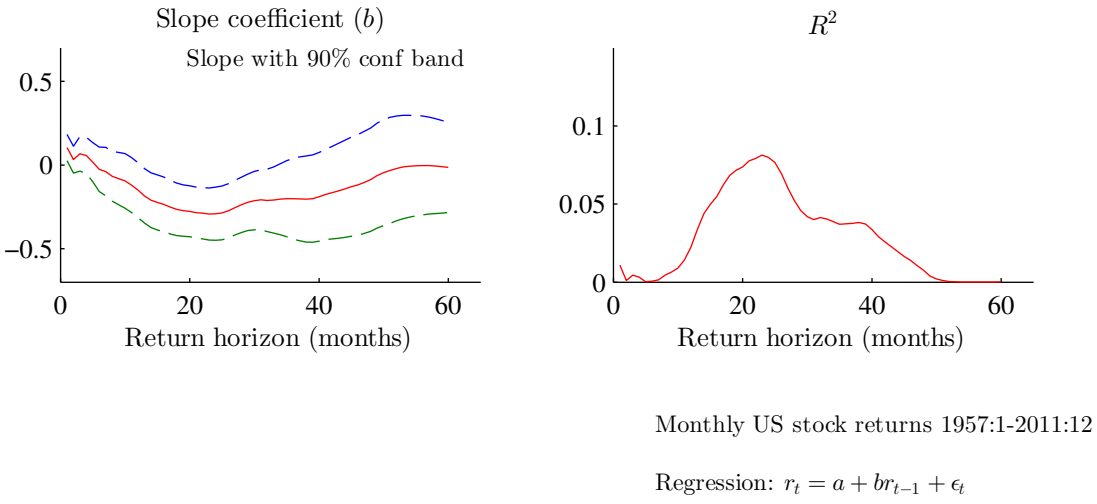


Figure 4.9: Predictability of US stock returns

4.2.4 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), $x_t = a_1 x_{t-1} + a_2 x_{t-2} + \varepsilon_t$, we have

$$\begin{bmatrix} \text{Cov}(x_t, x_t) \\ \text{Cov}(x_{t-1}, x_t) \\ \text{Cov}(x_{t-2}, x_t) \end{bmatrix} = \begin{bmatrix} \text{Cov}(x_t, a_1 x_{t-1} + a_2 x_{t-2} + \varepsilon_t) \\ \text{Cov}(x_{t-1}, a_1 x_{t-1} + a_2 x_{t-2} + \varepsilon_t) \\ \text{Cov}(x_{t-2}, a_1 x_{t-1} + a_2 x_{t-2} + \varepsilon_t) \end{bmatrix}, \text{ or}$$

$$\begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} a_1\gamma_1 + a_2\gamma_2 + \text{Var}(\varepsilon_t) \\ a_1\gamma_0 + a_2\gamma_1 \\ a_1\gamma_1 + a_2\gamma_0 \end{bmatrix}. \tag{4.11}$$

To transform to autocorrelation, divide through by γ_0. The last two equations are then

$$\begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2\rho_1 \\ a_1\rho_1 + a_2 \end{bmatrix} \text{ or } \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} a_1/(1-a_2) \\ a_1^2/(1-a_2) + a_2 \end{bmatrix}. \tag{4.12}$$

If we know the parameters of the AR(2) model (a_1, a_2, and $\text{Var}(\varepsilon_t)$), 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 that 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.

4.2.5 Variance Ratios

The 2-period variance ratio is the ratio of $\text{Var}(r_t + r_{t-1})$ to $2\,\text{Var}(r_t)$

$$VR_2 = \frac{\text{Var}(r_t + r_{t-1})}{2\,\text{Var}(r_t)} \tag{4.13}$$

$$= 1 + \rho_1, \tag{4.14}$$

where ρ_s is the sth autocorrelation. If r_t is not serially correlated, then this variance ratio is unity; a value above one indicates positive serial correlation and a value below one indicates negative serial correlation.

Proof. (of (4.14)) Let r_t have a zero mean (or be demeaned), so $\text{Cov}(r_t, r_{t-s}) = \text{E}\, r_t r_{t-s}$. We then have

$$\begin{aligned} VR_2 &= \frac{\text{E}(r_t + r_{t-1})^2}{2\,\text{E}\, r_t^2} \\ &= \frac{\text{Var}(r_t) + \text{Var}(r_{t-1}) + 2\,\text{Cov}(r_t, r_{t-1})}{2\,\text{Var}(r_t)} \\ &= \frac{1 + 1 + 2\rho_1}{2}, \end{aligned}$$

which gives (4.14). ■

We can also consider longer variance ratios, where we sum q observations in the numerator and then divide by $q\,\text{Var}(r_t)$. In fact, it can be shown that we have

$$VR_q = \frac{\text{Var}\left(\sum_{s=0}^{q-1} r_{t-s}\right)}{q\,\text{Var}(r_t)} \tag{4.15}$$

$$= \sum_{s=-(q-1)}^{q-1} \left(1 - \frac{|s|}{q}\right) \rho_s \text{ or} \tag{4.16}$$

$$= 1 + 2\sum_{s=1}^{q-1} \left(1 - \frac{s}{q}\right) \rho_s. \tag{4.17}$$

The third line exploits the fact that the autocorrelation (and autocovariance) function is symmetric around zero, so $\rho_{-s} = \rho_s$. (We could equally well let the summation in (4.16) and (4.17) run from $-q$ to q since the weight $1-|s|/q$, is zero for that lag.) It is immediate that no autocorrelation means that $VR_q = 1$ for all q. If all autocorrelations are non-positive, $\rho_s \leq 0$, then $VR_q \leq 1$, and vice versa.

Example 4.9 *(VR_3) For $q = 3$, (4.15)–(4.17) are*

$$\begin{aligned} VR_3 &= \frac{\text{Var}\left(r_t + r_{t-1} + r_{t-2}\right)}{3\,\text{Var}(r_t)} \\ &= \frac{1}{3}\rho_{-2} + \frac{2}{3}\rho_{-1} + 1 + \frac{2}{3}\rho_1 + \frac{1}{3}\rho_2 \\ &= 1 + 2\left(\frac{2}{3}\rho_1 + \frac{1}{3}\rho_2\right). \end{aligned}$$

Proof. (of (4.16)) The numerator in (4.15) is

$$\begin{aligned} \text{Var}(r_t + r_{t-1} + \ldots + r_{t-q+1}) = q\,\text{Var}(r_t) + 2(q-1)\,\text{Cov}(r_t, r_{t-1}) + 2(q-2)\,\text{Cov}(r_t, r_{t-2}) + \ldots \\ + 2\,\text{Cov}(r_t, r_{t-q+1}). \end{aligned}$$

For instance, for $q = 3$

$$\begin{aligned} \text{Var}(r_t + r_{t-1} + r_{t-2}) = \text{Var}(r_t) + \text{Var}(r_{t-1}) + \text{Var}(r_{t-2}) + \\ 2\,\text{Cov}(r_t, r_{t-1}) + 2\,\text{Cov}(r_{t-1}, r_{t-2}) + \\ 2\,\text{Cov}(r_t, r_{t-2}). \end{aligned}$$

Assume that variances and covariances are constant over time. Divide by $q\,\text{Var}(r_t)$ to get

$$VR_q = 1 + 2\left(1 - \frac{1}{q}\right)\rho_1 + 2\left(1 - \frac{2}{q}\right)\rho_2 + \ldots + 2\frac{1}{q}\rho_{q-1}.$$

■

Example 4.10 *(Variance ratio of an AR(1)) When $r_t = a\,r_{t-1} + \varepsilon_t$ where ε_t is iid white noise (and r_t has a zero mean or is demeaned), then*

$$\begin{aligned} VR_2 &= 1 + a \text{ and} \\ VR_3 &= 1 + \frac{4}{3}a + \frac{2}{3}a^2. \end{aligned}$$

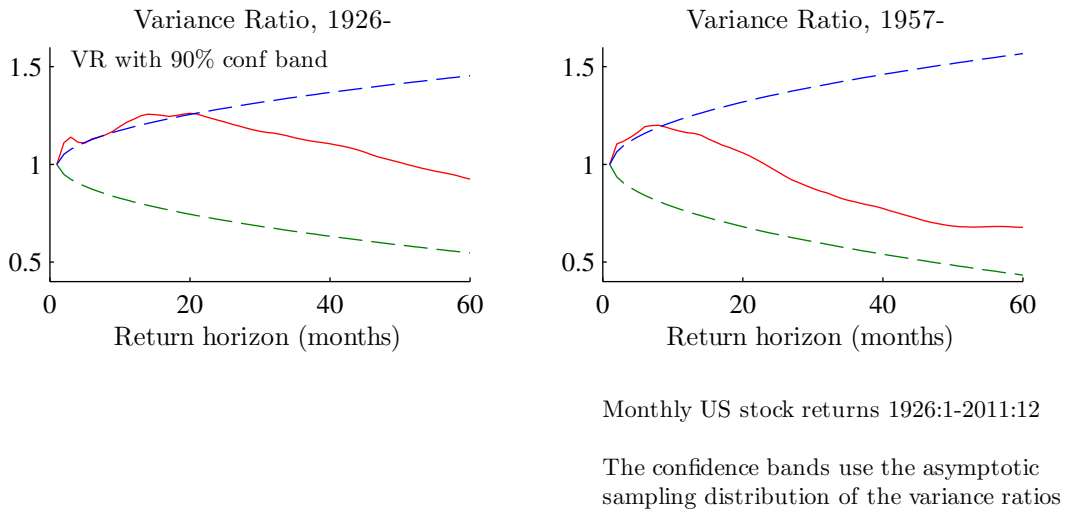


Figure 4.10: Variance ratios of US stock returns

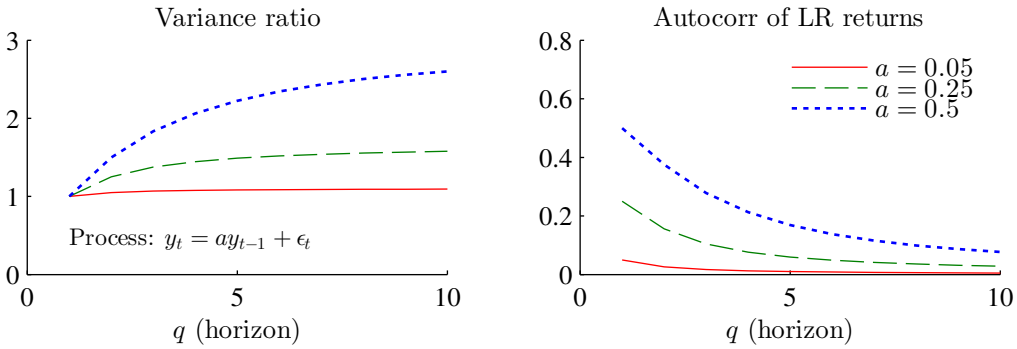


Figure 4.11: Variance ratio and long run autocorrelation of an AR(1) process

See Figure 4.11 for a numerical example.

The estimation of VR_q is done by replacing the population variances in (4.15) with the sample variances, or the autocorrelations in (4.17) by the sample autocorrelations.

The sampling distribution of $\widehat{VR}_q$ under the null hypothesis that there is no autocorrelation follows from the sampling distribution of the autocorrelation coefficient. Rewrite (4.17) as

$$\sqrt{T}\left(\widehat{VR}_q - 1\right) = 2\sum_{s=1}^{q-1}\left(1 - \frac{s}{q}\right)\sqrt{T}\hat{\rho}_s. \tag{4.18}$$

If the assumptions behind (4.6) are satisfied, then we have that, under the null hypothesis of no autocorrelation, (4.18) is a linear combination of (asymptotically) uncorrelated $N(0,1)$ variables (the $\sqrt{T}\hat{\rho}_s$). It then follows that

$$\sqrt{T}\left(\widehat{VR}_q - 1\right) \to^d N\left[0, \sum_{s=1}^{q-1} 4\left(1 - \frac{s}{q}\right)^2\right]. \tag{4.19}$$

Example 4.11 *(Distribution of $\widehat{VR}_2$ and $\widehat{VR}_3$) We have*

$$\sqrt{T}\left(\widehat{VR}_2 - 1\right) \to^d N(0,1) \text{ and } \sqrt{T}\left(\widehat{VR}_3 - 1\right) \to^d N(0, 20/9).$$

These distributional results depend on the assumptions behind the results in (4.6). One way of handling deviations from those assumptions is to estimate the autocorrelations and their covariance matrix with GMM, alternatively, the results in Taylor (2005) can be used.

See Figure 4.10 for an illustration.

4.2.6 Long-Run Autoregressions

Consider an AR(1) of two-period sums of non-overlapping (log) returns

$$r_{t+1} + r_{t+2} = a + b_2(r_{t-1} + r_t) + \varepsilon_{t+2}. \tag{4.20}$$

Notice that it is important that dependent variable and the regressor are non-overlapping (don't include the return for the same period)—otherwise we are likely to find spurious autocorrelation. The least squares population regression coefficient is

$$b_2 = \frac{\text{Cov}(r_{t+1} + r_{t+2}, r_{t-1} + r_t)}{\text{Var}(r_{t-1} + r_t)} \tag{4.21}$$

$$= \frac{1}{VR_2}\frac{\rho_1 + 2\rho_2 + \rho_3}{2}. \tag{4.22}$$

Proof. (of (4.22)) Multiply and divide (4.21) by $2\,\text{Var}(r_t)$

$$b_2 = \frac{2\,\text{Var}(r_t)}{\text{Var}(r_{t-1} + r_t)}\frac{\text{Cov}(r_{t+1} + r_{t+2}, r_{t-1} + r_t)}{2\,\text{Var}(r_t)}.$$

The first term is $1/VR_2$. The numerator of the second term is

$$\text{Cov}(r_{t+1} + r_{t+2}, r_{t-1} + r_t) = \text{Cov}(r_{t+1}, r_{t-1}) + \text{Cov}(r_{t+1}, r_t) + \text{Cov}(r_{t+2}, r_{t-1}) + \text{Cov}(r_{t+2}, r_t),$$

so the second term simplifies to

$$\frac{1}{2}\left(\rho_2 + \rho_1 + \rho_3 + \rho_2\right).$$

■

The general pattern that emerges from these expressions is that the slope coefficient in an AR(1) of (non-overlapping) long-run returns

$$\sum_{s=1}^{q} r_{t+s} = a + b_q \sum_{s=1}^{q} r_{t+s-q} + \varepsilon_{t+q} \tag{4.23}$$

is

$$b_q = \frac{1}{VR_q} \sum_{s=-(q-1)}^{q-1} \left(1 - \frac{|s|}{q}\right) \rho_{q+s}. \tag{4.24}$$

Note that the autocorrelations are displaced by the amount q. As for the variance ratio, the summation could run from $-q$ to q instead, since the weight, $1-|s|/q$, is zero for that lag.

Equation (4.24) shows that the variance ratio and the AR(1) coefficient of long-run returns are closely related. A bit of manipulation (and using the fact that $\rho_{-s} = \rho_s$) shows that

$$1 + b_q = \frac{VR_{2q}}{VR_q}. \tag{4.25}$$

If the variance ratio increases with the horizon, then this means that the long-run returns are positively autocorrelated.

Example 4.12 *(Long-run autoregression of an AR(1)) When $r_t = a r_{t-1} + \varepsilon_t$ where ε_t is iid white noise, then the variance ratios are as in Example (4.10), and we know that $\rho_{q+s} = a^{q+s}$. From (4.22) we then have*

$$\begin{aligned} b_2 &= \frac{1}{VR_2} \frac{a + 2a^2 + a^3}{2} \\ &= \frac{1}{1+a} \frac{a + 2a^2 + a^3}{2}. \end{aligned}$$

See Figure 4.11 for a numerical example. For future reference, note that we can simplify to get $b_2 = (1+a)\,a/2$.

Example 4.13 *(Trying (4.25) on an AR(1)) From Example (4.10) we have that*

$$\begin{aligned}\frac{VR_4}{VR_2} - 1 &= \frac{1 + \frac{3}{2}a + a^2 + \frac{1}{2}a^3}{1 + a} - 1 \\ &= \frac{1}{2}(1 + a)\, a,\end{aligned}$$

which is b_2 in Example 4.12.

Using All Data Points in Long-Run Autoregressions?*

Inference of the slope coefficient in long-run autoregressions like (4.20) must be done with care. While it is clear that the dependent variable and the regressor must be for non-overlapping periods, there is still the issue of whether we should use all available data points or not.

Suppose one-period returns actually are serially uncorrelated and have zero means (to simplify)

$$r_t = u_t, \text{ where } u_t \text{ is iid with } \mathrm{E}\, u_u = 0, \tag{4.26}$$

and that we are studying two-periods returns. One possibility is to use $r_{t+1} + r_{t+2}$ as the first observation and $r_{t+3} + r_{t+4}$ as the second observation: no common period. This clearly halves the sample size, but has an advantage when we do inference. To see that, notice that two successive observations are then

$$r_{t+1} + r_{t+2} = a + b_2 (r_{t-1} + r_t) + \varepsilon_{t+2} \tag{4.27}$$

$$r_{t+3} + r_{t+4} = a + b_2 (r_{t+1} + r_{t+2}) + \varepsilon_{t+4}. \tag{4.28}$$

If (4.26) is true, then $a = b_2 = 0$ and the residuals are

$$\varepsilon_{t+2} = u_{t+1} + u_{t+2} \tag{4.29}$$

$$\varepsilon_{t+4} = u_{t+3} + u_{t+4}, \tag{4.30}$$

which are uncorrelated.

Compare this to the case where we use all data. Two successive observations are then

$$r_{t+1} + r_{t+2} = a + b_2 (r_{t-1} + r_t) + \varepsilon_{t+2} \tag{4.31}$$

$$r_{t+2} + r_{t+3} = a + b_2 (r_t + r_{t+1}) + \varepsilon_{t+3}. \tag{4.32}$$

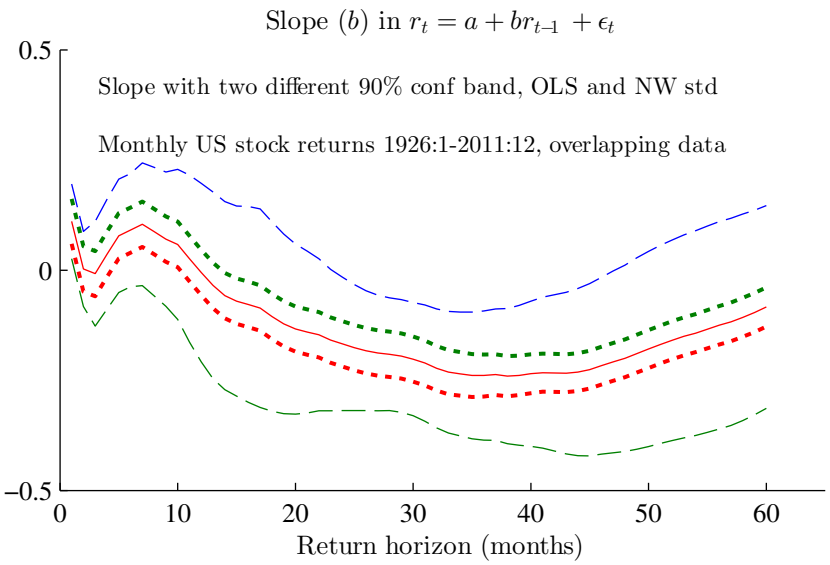


Figure 4.12: Slope coefficient, LS vs Newey-West standard errors

As before, if (4.26) is true, then $a = b_2 = 0$ (so there is no problem with the point estimates), but the residuals are

$$\varepsilon_{t+2} = u_{t+1} + \underbrace{u_{t+2}} \tag{4.33}$$

$$\varepsilon_{t+3} = \underbrace{u_{t+2}} + u_{t+3}, \tag{4.34}$$

which are correlated since u_{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. See Figure 4.12 for an illustration.

4.3 Multivariate (Auto-)correlations

4.3.1 Momentum or Contrarian Strategy?

A momentum strategy invests in assets that have performed well recently—and often goes short in those that have underperformed. See 4.13 for an empirical illustration.

To formalize this, let there be N assets with with returns R, with means and autoco-

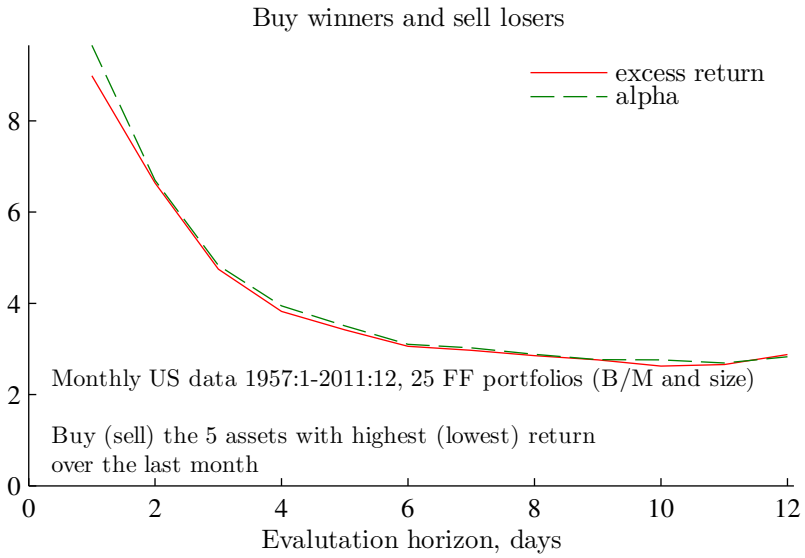


Figure 4.13: Performance of momentum investing

variance matrix

$$\begin{aligned} \mathrm{E}\, R &= \mu \text{ and} \\ \Gamma(k) &= \mathrm{E}[(R_t - \mu)(R_{t-k} - \mu)']. \end{aligned} \tag{4.35}$$

Example 4.14 *($\Gamma(k)$ with two assets) We have*

$$\Gamma(k) = \begin{bmatrix} \mathrm{Cov}(R_{1,t}, R_{1,t-k}) & \mathrm{Cov}(R_{1,t}, R_{2,t-k}) \\ \mathrm{Cov}(R_{2,t}, R_{1,t-k}) & \mathrm{Cov}(R_{2,t}, R_{2,t-k}) \end{bmatrix}.$$

Define the equal weighted market portfolio return as simply

$$R_{mt} = \frac{1}{N}\sum_{i=1}^{N} R_{it} = \mathbf{1}'R_t/N \tag{4.36}$$

with the corresponding mean return

$$\mu_m = \frac{1}{N}\sum_{i=1}^{N} \mu_i = \mathbf{1}'\mu/N. \tag{4.37}$$

A *momentum strategy* could (for instance) use the portfolio weights

$$w_t(k) = \frac{R_{t-k} - R_{mt-k}}{N}, \tag{4.38}$$

which basically says that $w_{it}(k)$ is positive for assets with an above average return k periods back. Notice that the weights sum to zero, so this is a zero cost portfolio. However, the weights differ from fixed weights (for instance, put $1/5$ into the best 5 assets, and $-1/5$ into the 5 worst assets) since the overall size of the exposure ($1'|w_t|$) changes over time. A large dispersion of the past returns means large positions and vice versa. To analyse a contrarian strategy, reverse the sign of (4.38).

The profit from this strategy is

$$\pi_t(k) = \sum_{i=1}^{N} \underbrace{\frac{R_{it-k} - R_{mt-k}}{N}}_{w_{it}} R_{it} = \sum_{i=1}^{N} \frac{R_{it-k}R_{it}}{N} - R_{mt-k}R_{mt}, \tag{4.39}$$

where the last term uses the fact that $\Sigma_{i=1}^{N} R_{mt-k}R_{it}/N = R_{mt-k}R_{mt}$.

The expected value is

$$\mathrm{E}\,\pi_t(k) = -\frac{1}{N^2}\left[1'\Gamma(k)1 - \mathrm{tr}\Gamma(k)\right] + \frac{N-1}{N^2}\mathrm{tr}\Gamma(k) + \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mu_m)^2, \tag{4.40}$$

where the $1'\Gamma(k)1$ sums all the elements of $\Gamma(k)$ and $\mathrm{tr}\Gamma(k)$ sums the elements along the main diagonal. (See below for a proof.) To analyse a contrarian strategy, reverse the sign of (4.40).

With a random walk, $\Gamma(k) = 0$, then (4.40) shows that the momentum strategy wins money: the first two terms are zero, while the third term contributes to a positive performance. The reason is that the momentum strategy (on average) invests in assets with high average returns ($\mu_i > \mu_m$).

The first term of (4.40) sums all elements in the autocovariance matrix and then subtracts the sum of the diagonal elements—so it only depends on the sum of the cross-covariances, that is, how a return is correlated with the lagged return of other assets. In general, negative cross-covariances benefit a momentum strategy. To see why, suppose a high lagged return on asset 1 predicts a low return on asset 2, but asset 2 cannot predict asset 1 ($\mathrm{Cov}(R_{2,t}, R_{1,t-k}) < 0$ and $\mathrm{Cov}(R_{1,t}, R_{2,t-k}) = 0$). This helps the momentum strategy since we have a negative portfolio weight of asset 2 (since it performed relatively

poorly in the previous period).

Example 4.15 *((4.40) with 2 assets) Suppose we have*

$$\Gamma(k) = \begin{bmatrix} \text{Cov}(R_{1,t}, R_{1,t-k}) & \text{Cov}(R_{1,t}, R_{2,t-k}) \\ \text{Cov}(R_{2,t}, R_{1,t-k}) & \text{Cov}(R_{2,t}, R_{2,t-k}) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -0.1 & 0 \end{bmatrix}.$$

Then

$$-\frac{1}{N^2}\left[1'\Gamma(k)1 - tr\Gamma(k)\right] = -\frac{1}{2^2}\left[-0.1 - 0\right] = 0.025, \text{ and}$$
$$\frac{N-1}{N^2} tr\Gamma(k) = \frac{2-1}{2} \times 0 = 0,$$

so the sum of the first two terms of (4.40) is positive (good for a momentum strategy). For instance, suppose $R_{1,t-k} > 0$, then $R_{2,t}$ tends to be low which is good (we have a negative portfolio weight on asset 2).

The second term of (4.40) depends only on own autocovariances, that is, how a return is correlated with the lagged return of the same asset. If these own autocovariances are (on average) positive, then a strongly performing asset in $t-k$ tends to perform well in t, which helps a momentum strategy (as the strongly performing asset is overweighted).

See Figure 4.15 for an illustration based on Figure 4.14.

Example 4.16 *Figure 4.15 shows that a momentum strategy works reasonably well on daily data on the 25 FF portfolios. While the cross-covariances have a negative influence (because they are mostly positive), they are dominated by the (on average) positive auto-correlations. The correlation matrix is illustrated in Figure 4.14. In short, the small firms (asset 1-5) are correlated with the lagged returns of most assets, while large firms are not.*

Example 4.17 *((4.40) with 2 assets) With*

$$\Gamma(k) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

then

$$-\frac{1}{N^2}\left[1'\Gamma(k)1 - tr\Gamma(k)\right] = -\frac{1}{2^2}\left(0.2 - 0.2\right) = 0, \text{ and}$$
$$\frac{N-1}{N^2} tr\Gamma(k) = \frac{2-1}{2} \times (0.1 + 0.1) = 0.05,$$

(Auto−)correlation matrix, daily FF returns 1979:1−2011:12

	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	0.12	0.11	0.08	0.07	0.06	0.15	0.11	0.09	0.08	0.07	0.16	0.12	0.08	0.07	0.07	0.17	0.11	0.08	0.06	0.05	0.14	0.11	0.08	0.06	0.05
2	0.09	0.08	0.06	0.05	0.04	0.12	0.09	0.06	0.06	0.05	0.13	0.10	0.06	0.06	0.05	0.13	0.09	0.07	0.04	0.04	0.11	0.09	0.06	0.05	0.04
3	0.06	0.05	0.03	0.02	0.02	0.09	0.06	0.03	0.03	0.03	0.09	0.07	0.04	0.03	0.03	0.10	0.07	0.05	0.02	0.03	0.09	0.07	0.05	0.04	0.03
4	0.05	0.05	0.03	0.02	0.02	0.08	0.05	0.03	0.03	0.03	0.09	0.06	0.03	0.03	0.03	0.09	0.06	0.04	0.02	0.02	0.09	0.07	0.04	0.03	0.03
5	0.09	0.08	0.07	0.07	0.07	0.11	0.09	0.07	0.07	0.08	0.11	0.10	0.07	0.07	0.07	0.12	0.09	0.08	0.06	0.06	0.10	0.09	0.07	0.06	0.06
6	0.08	0.07	0.05	0.05	0.04	0.12	0.09	0.07	0.06	0.06	0.13	0.10	0.07	0.07	0.06	0.14	0.10	0.09	0.06	0.06	0.14	0.12	0.08	0.07	0.07
7	0.03	0.03	0.02	0.01	0.01	0.07	0.04	0.03	0.02	0.02	0.08	0.06	0.03	0.03	0.03	0.09	0.07	0.05	0.02	0.02	0.09	0.07	0.05	0.04	0.03
8	0.01	0.01	0.00	−0.00	−0.00	0.04	0.03	0.01	0.01	0.01	0.05	0.04	0.02	0.02	0.02	0.06	0.05	0.03	0.01	0.02	0.07	0.05	0.03	0.03	0.02
9	0.01	0.01	−0.00	−0.01	−0.01	0.04	0.02	0.01	0.00	0.01	0.05	0.04	0.02	0.02	0.02	0.06	0.05	0.03	0.01	0.01	0.06	0.05	0.03	0.02	0.02
10	0.03	0.03	0.02	0.01	0.02	0.05	0.04	0.02	0.02	0.03	0.05	0.05	0.03	0.04	0.03	0.06	0.05	0.05	0.03	0.03	0.06	0.05	0.03	0.03	0.03
11	0.06	0.05	0.04	0.04	0.03	0.10	0.08	0.06	0.06	0.05	0.11	0.09	0.07	0.07	0.06	0.13	0.09	0.08	0.06	0.05	0.14	0.11	0.08	0.07	0.07
12	0.04	0.04	0.04	0.03	0.03	0.08	0.06	0.05	0.05	0.05	0.09	0.08	0.06	0.06	0.05	0.10	0.09	0.08	0.05	0.05	0.11	0.10	0.08	0.07	0.06
13	0.03	0.03	0.03	0.03	0.02	0.06	0.05	0.04	0.04	0.04	0.07	0.07	0.05	0.06	0.05	0.08	0.08	0.07	0.05	0.05	0.09	0.09	0.06	0.06	0.06
14	0.02	0.03	0.02	0.02	0.02	0.05	0.04	0.03	0.03	0.04	0.05	0.05	0.04	0.05	0.04	0.06	0.06	0.05	0.04	0.04	0.07	0.06	0.05	0.05	0.04
15	0.02	0.03	0.02	0.02	0.02	0.05	0.05	0.04	0.04	0.04	0.05	0.06	0.05	0.05	0.05	0.06	0.07	0.06	0.04	0.05	0.07	0.07	0.05	0.06	0.06
16	0.02	0.03	0.02	0.02	0.01	0.06	0.05	0.04	0.04	0.04	0.08	0.07	0.04	0.05	0.04	0.09	0.06	0.05	0.03	0.03	0.11	0.09	0.06	0.06	0.05
17	0.04	0.04	0.04	0.03	0.03	0.07	0.07	0.06	0.06	0.06	0.09	0.08	0.06	0.07	0.06	0.10	0.09	0.07	0.05	0.05	0.11	0.10	0.08	0.08	0.07
18	0.03	0.04	0.03	0.03	0.03	0.06	0.06	0.05	0.05	0.05	0.07	0.07	0.06	0.06	0.05	0.08	0.08	0.07	0.05	0.05	0.10	0.09	0.07	0.07	0.06
19	0.03	0.03	0.03	0.03	0.03	0.05	0.05	0.04	0.05	0.05	0.06	0.06	0.05	0.06	0.05	0.07	0.07	0.06	0.05	0.05	0.08	0.07	0.06	0.06	0.06
20	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.03	0.03	0.04	0.05	0.05	0.04	0.05	0.04	0.06	0.06	0.05	0.04	0.04	0.06	0.06	0.04	0.05	0.05
21	−0.05	−0.05	−0.05	−0.05	−0.06	−0.03	−0.04	−0.04	−0.04	−0.04	−0.02	−0.03	−0.04	−0.04	−0.04	−0.02	−0.03	−0.04	−0.05	−0.05	−0.00	−0.02	−0.04	−0.03	−0.04
22	−0.04	−0.04	−0.04	−0.04	−0.05	−0.02	−0.03	−0.03	−0.03	−0.03	−0.01	−0.02	−0.03	−0.02	−0.03	−0.01	−0.02	−0.03	−0.04	−0.04	−0.00	−0.01	−0.03	−0.03	−0.03
23	−0.02	−0.02	−0.03	−0.02	−0.03	−0.01	−0.01	−0.01	−0.01	−0.00	0.00	0.00	−0.01	−0.00	−0.01	0.01	0.00	−0.01	−0.02	−0.02	0.01	0.01	−0.00	−0.00	−0.01
24	−0.05	−0.04	−0.05	−0.05	−0.05	−0.03	−0.03	−0.03	−0.03	−0.02	−0.02	−0.02	−0.03	−0.02	−0.03	−0.02	−0.02	−0.03	−0.04	−0.04	−0.01	−0.02	−0.03	−0.02	−0.02
25	−0.04	−0.03	−0.04	−0.03	−0.03	−0.02	−0.02	−0.02	−0.02	−0.01	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.03	−0.03	−0.02	−0.02	−0.03	−0.02	−0.02

Figure 4.14: Illustration of the cross-autocorrelations, $\text{Corr}(R_t, R_{t-k})$, daily FF data. Dark colors indicate high correlations, light colors indicate low correlations.

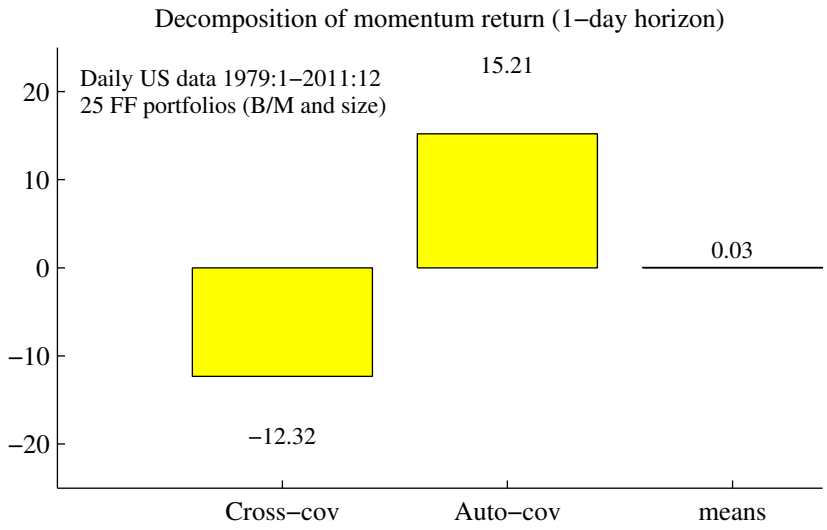


Figure 4.15: Decomposition of return from momentum strategy based on daily FF data

so the sum of the first two terms of (4.40) is positive (good for a momentum strategy).

Proof. (of (4.40)) Take expectations of (4.39) and use the fact that $\mathrm{E}\,xy = \mathrm{Cov}(x, y) + \mathrm{E}\,x\,\mathrm{E}\,y$ to get

$$\mathrm{E}\,\pi_t(k) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{Cov}(R_{it-k}, R_{it}) + \mu_i^2\right] - \left[\mathrm{Cov}(R_{mt-k}, R_{mt}) + \mu_m^2\right].$$

Notice that $\frac{1}{N}\sum_{i=1}^{N}\mathrm{Cov}(R_{it-k}, R_{it}) = \mathrm{tr}\Gamma(k)/N$, where tr denotes the trace. Also, let $\tilde{R} = R - \mu$ and notice that

$$\mathrm{Cov}(R_{mt-k}, R_{mt}) = \mathrm{E}\,\frac{1}{N^2}\left[\left(1'\tilde{R}_t\right)\left(1'\tilde{R}_{it-k}\right)'\right] = \mathrm{E}\,\frac{1}{N^2}\left[1'\tilde{R}_t\tilde{R}_{it-k}'1\right] = \frac{1'\Gamma(k)1}{N^2}.$$

Finally, we note that $\frac{1}{N}\sum_{i=1}^{N}\mu_i^2 - \mu_m^2 = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mu_m)^2$. Together, these results give

$$\mathrm{E}\,\pi_t(k) = -\frac{1'\Gamma(k)1}{N^2} + \frac{1}{N}\mathrm{tr}\Gamma(k) + \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mu_m)^2,$$

which can be rearranged as (4.40). ■

4.4 Other Predictors

There are many other, perhaps more economically plausible, 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.

See Figure 4.16 for an illustration.

4.4.1 Prices and Dividends

The Accounting Identity

Reference: Campbell, Lo, and MacKinlay (1997) 7 and Cochrane (2005) 20.1.

The gross return, R_{t+1}, is defined as

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t}, \text{ so } P_t = \frac{D_{t+1} + P_{t+1}}{R_{t+1}}. \tag{4.41}$$

Substituting for P_{t+1} (and then $P_{t+2}, ...$) gives

$$P_t = \frac{D_{t+1}}{R_{t+1}} + \frac{D_{t+2}}{R_{t+1}R_{t+2}} + \frac{D_{t+3}}{R_{t+1}R_{t+2}R_{t+3}} + \dots \tag{4.42}$$

$$= \sum_{j=1}^{\infty} \frac{D_{t+j}}{\prod_{k=1}^{j} R_{t+k}}, \tag{4.43}$$

provided the discounted value of P_{t+j} goes to zero as $j \to \infty$. This is simply an accounting identity. It is clear that a high price in t must lead to low future returns and/or high future dividends—which (by rational expectations) also carry over to expectations of future returns and dividends.

It is sometimes more convenient to analyze the price-dividend ratio. Dividing (4.42) and (4.43) by D_t gives

$$\frac{P_t}{D_t} = \frac{1}{R_{t+1}}\frac{D_{t+1}}{D_t} + \frac{1}{R_{t+1}R_{t+2}}\frac{D_{t+2}}{D_{t+1}}\frac{D_{t+1}}{D_t} + \frac{1}{R_{t+1}R_{t+2}R_{t+3}}\frac{D_{t+3}}{D_{t+2}}\frac{D_{t+2}}{D_{t+1}}\frac{D_{t+1}}{D_t} + \dots \tag{4.44}$$

$$= \sum_{j=1}^{\infty}\prod_{k=1}^{j} \frac{D_{t+k}/D_{t+k-1}}{R_{t+k}}. \tag{4.45}$$

As with (4.43) it is just an accounting identity. It must therefore also hold in expectations. Since expectations are good (the best?) predictors of future values, we have the implication that the asset price should predict a discounted sum of future dividends, (4.43), and that the price-dividend ratio should predict a discounted sum of future changes in dividends.

Linearizing the Accounting Identity

We now log-linearize the accounting identity (4.45) in order to tie it more closely to the (typically linear) econometrics methods for detecting predictability The result is

$$p_t - d_t \approx \sum_{s=0}^{\infty} \rho^s [(d_{t+1+s} - d_{t+s}) - r_{t+1+s}], \tag{4.46}$$

where $\rho = 1/(1 + \overline{D/P})$ where $\overline{D/P}$ is a steady state dividend-price ration ($\rho = 1/1.04 \approx 0.96$ if $\overline{D/P}$ is 4%).

As before, a high price-dividend ratio must imply future dividend growth and/or low future returns. In the exact solution (4.44), dividends and returns which are closer to the present show up more times than dividends and returns far in the future. In the approximation (4.46), this is captured by giving a higher weight (higher ρ^s).

Proof. (of (4.46)—slow version) Rewrite (4.41) as

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \frac{P_{t+1}}{P_t}\left(1 + \frac{D_{t+1}}{P_{t+1}}\right) \text{ or in logs}$$
$$r_{t+1} = p_{t+1} - p_t + \ln\left[1 + \exp(d_{t+1} - p_{t+1})\right].$$

Make a first order Taylor approximation of the last term around a steady state value of $d_{t+1} - p_{t+1}$, denoted $\overline{d - p}$,

$$\ln\left[1 + \exp(d_{t+1} - p_{t+1})\right] \approx \ln\left[1 + \exp(\overline{d - p})\right] + \frac{\exp(\overline{d - p})}{1 + \exp(\overline{d - p})}\left[d_{t+1} - p_{t+1} - \left(\overline{d - p}\right)\right]$$
$$\approx \text{ constant} + (1 - \rho)\left(d_{t+1} - p_{t+1}\right),$$

where $\rho = 1/[1 + \exp(\overline{d - p})] = 1/(1 + \overline{D/P})$. Combine and forget about the constant. The result is

$$r_{t+1} \approx p_{t+1} - p_t + (1 - \rho)\left(d_{t+1} - p_{t+1}\right)$$
$$= \rho p_{t+1} - p_t + (1 - \rho)\, d_{t+1},$$

where $0 < \rho < 1$. Add and subtract d_t from the right hand side and rearrange

$$r_{t+1} \approx \rho\left(p_{t+1} - d_{t+1}\right) - \left(p_t - d_t\right) + \left(d_{t+1} - d_t\right), \text{ or}$$
$$p_t - d_t \approx \rho\left(p_{t+1} - d_{t+1}\right) + \left(d_{t+1} - d_t\right) - r_{t+1}$$

This is a (forward looking, unstable) difference equation, which we can solve recursively forward. Provided $\lim_{s\to\infty} \rho^s(p_{t+s} - d_{t+s}) = 0$, the solution is (4.46). (Trying to solve for the log price level instead of the log price-dividend ratio is problematic since the condition $\lim_{s\to\infty} \rho^s p_{t+s} = 0$ may not be satisfied.) ■

Dividend-Price Ratio as a Predictor

One of the most successful attempts to forecast long-run return is by using the dividend-price ratio

$$\sum_{s=1}^{q} r_{t+s} = \alpha + \beta_q(d_t - p_t) + \varepsilon_{t+q}. \tag{4.47}$$

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 also Figure 4.16 for an illustration.

By comparing with (4.46), we see that the dividend-ratio in (4.47) is only asked to predict a finite (unweighted) sum of future returns—dividend growth is disregarded. We should therefore expect (4.47) to work particularly well if the horizon is long (high q) and if dividends are stable over time.

From (4.46) we get (from using $\text{Cov}(x, y - z) = \text{Cov}(x, y) - \text{Cov}(x, z)$) that

$$\text{Var}(p_t - d_t) \approx \text{Cov}\left(p_t - d_t, \sum_{s=0}^{\infty} \rho^s \left(d_{t+1+s} - d_{t+s}\right)\right) - \text{Cov}\left(p_t - d_t, \sum_{s=0}^{\infty} \rho^s r_{t+1+s}\right), \tag{4.48}$$

which shows that the variance of the price-dividend ratio can be decomposed into the covariance of price-dividend ratio with future dividend change minus the covariance of price-dividend ratio with future returns. This expression highlights that if $p_t - d_t$ is not constant, then it must forecast dividend growth and/or returns.

The evidence in Cochrane suggests that $p_t - d_t$ does not forecast future dividend growth, so that predictability of future returns explains the variability in the dividend-price ratio. This fits very well into the findings of the R^2 of (4.47). To see that, recall the following fact.

Remark 4.18 *(R^2 from a least squares regression) Let the least squares estimate of β in $y_t = x_t'\beta_0 + u_t$ be $\hat{\beta}$. The fitted values $\hat{y}_t = x_t'\hat{\beta}$. If the regression equation includes a constant, then $R^2 = \widehat{Corr}(y_t, \hat{y}_t)^2$. In a simple regression where $y_t = a + bx_t + u_t$, where x_t is a scalar, $R^2 = \widehat{Corr}(y_t, x_t)^2$.*

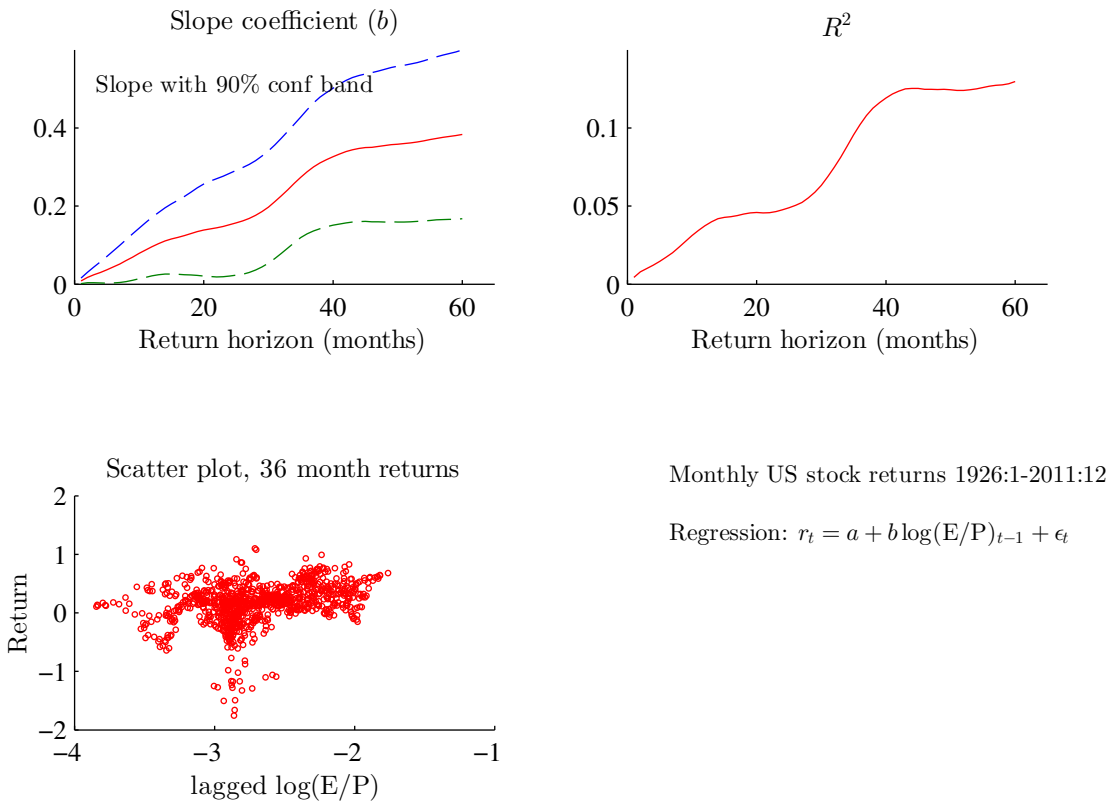


Figure 4.16: Predictability of US stock returns

4.4.2 Predictability but No Autocorrelation

The evidence for US stock returns is that long-run returns may perhaps be predicted by using 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"). Both CLM 7.1.4 and Cochrane 20.1 use small models for discussing this case. The key in these discussions is to make changes in dividends unforecastable, but let the return be forecastable by some state variable ($\mathrm{E}_t\, d_{t+1+s} - \mathrm{E}_t\, d_{t+s} = 0$ and $\mathrm{E}_t\, r_{t+1} = r + x_t$), but in such a way that there is little autocorrelation in returns. By taking expectations of (4.46) we see that price-dividend will then reflect expected future returns and therefore be useful for forecasting.

4.5 Maximally Predictable Portfolio*

As a way to calculate an upper bound on predictability, Lo and MacKinlay (1997) construct maximally predictable portfolios. The weights on the different assets in this portfolio can also help us to understand more about how the predictability works.

Let Z_t be an $n \times 1$ vector of demeaned returns

$$Z_t = R_t - \mathrm{E}\, R_t, \tag{4.49}$$

and suppose that we (somehow) have constructed rational forecasts $\mathrm{E}_{t-1}\, Z_t$ such that

$$Z_t = \mathrm{E}_{t-1}\, Z_t + \varepsilon_t, \text{ where } \mathrm{E}_{t-1}\, \varepsilon_t = 0,\ \mathrm{Var}_{t-1}(\varepsilon_t \varepsilon_t') = \Sigma. \tag{4.50}$$

Consider a portfolio $\gamma' Z_t$. The R^2 from predicting the return on this portfolio is (as usual) the fraction of the variability of $\gamma' Z_t$ that is explained by $\gamma'\, \mathrm{E}_{t-1}\, Z_t$

$$\begin{aligned} R^2(\gamma) &= 1 - \mathrm{Var}(\gamma'\varepsilon_t)/\,\mathrm{Var}(\gamma' Z_t) \\ &= [\mathrm{Var}(\gamma' Z_t) - \mathrm{Var}(\gamma'\varepsilon_t)]/\,\mathrm{Var}(\gamma' Z_t) \\ &= \mathrm{Var}(\gamma'\, \mathrm{E}_{t-1}\, Z_t)/\,\mathrm{Var}(\gamma' Z_t) \\ &= \gamma'\, \mathrm{Cov}(\mathrm{E}_{t-1}\, Z_t)\gamma / \gamma'\, \mathrm{Cov}(Z_t)\gamma. \end{aligned} \tag{4.51}$$

The covariance in the denominator can be calculated directly from data, but the covariance matrix in the numerator clearly depends on the forecasting model we use (to create $\mathrm{E}_{t-1}\, Z_t$).

The portfolio (γ vector) that gives the highest R^2 is the eigenvector (normalized to sum to unity) associated with the largest eigenvalue (also the value of R^2) of $\mathrm{Cov}(Z_t)^{-1}\, \mathrm{Cov}(\mathrm{E}_{t-1}\, Z_t)$.

Example 4.19 *(One forecasting variable) Suppose there is only one predictor, x_{t-1},*

$$Z_t = \beta x_{t-1} + \varepsilon_t,$$

where β is $n \times 1$. This means that $\mathrm{E}_{t-1}\, Z_t = \beta x_{t-1}$, so $\mathrm{Cov}(\mathrm{E}_{t-1}\, Z_t) = \mathrm{Var}(x_{t-1})\beta\beta'$ and that $\mathrm{Cov}(Z_t) = \mathrm{Var}(x_{t-1})\beta\beta' + \Sigma$. We can therefore write (4.51) as

$$R^2(\gamma) = \frac{\gamma'\, \mathrm{Var}(x_{t-1})\beta\beta'\gamma}{\gamma'\, \mathrm{Var}(x_{t-1})\beta\beta'\gamma + \gamma'\Sigma\gamma}.$$

The first order conditions for maximum then gives (this is very similar to the calculations

of the minimum variance portfolio in mean-variance analysis)

$$\gamma = \Sigma^{-1}\beta / \mathbf{1}'\Sigma^{-1}\beta,$$

where $\mathbf{1}$ *is an* $n \times 1$ *vector of ones. In particular, if* Σ *(and therefore* Σ^{-1}*) is diagonal, then the portfolio weight of asset* i *is* β_i *divided by the variance of the forecast error of asset* i*: assets which are hard to predict get smaller weights. We also see that if the sign of* β_i *is different from the sign of* $\mathbf{1}'\Sigma^{-1}\beta$*, then it gets a negative weight. For instance, if* $\mathbf{1}'\Sigma^{-1}\beta > 0$*, so that most assets move in the same direction as* x_{t-1}*, then asset* i *gets a negative weight if it moves in the opposite direction (*$\beta_i < 0$*).*

4.6 Evaluating Forecast Performance

Further reading: Diebold (2001) 11; Stekler (1991); Diebold and Mariano (1995)

To do a solid evaluation of the forecast performance (of some forecaster/forecast method/forecast institute), we need a sample (history) of the forecasts and the resulting forecast errors. The reason is that the forecasting performance for a single period is likely to be dominated by luck, so we can only expect to find systematic patterns by looking at results for several periods.

Let e_t be the forecast error in period t

$$e_t = y_t - \hat{y}_t, \tag{4.52}$$

where $\hat{y}_t$ is the forecast and y_t the actual outcome. (Warning: some authors prefer to work with $\hat{y}_t - y_t$ as the forecast error instead.)

Most statistical forecasting methods are based on the idea of minimizing the sum of squared forecast errors, $\Sigma_{t=1}^{T} e_t^2$. For instance, the least squares (LS) method picks the regression coefficient in

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{4.53}$$

to minimize the sum of squared residuals, $\Sigma_{t=1}^{T} \varepsilon_t^2$. This will, among other things, give a zero mean of the fitted residuals and also a zero correlation between the fitted residual and the regressor.

Evaluation of a forecast often involve extending these ideas to the forecast method, irrespective of whether a LS regression has been used or not. In practice, this means

studying if *(i)* the forecast error, e_t, has a zero mean; *(ii)* the forecast error is uncorrelated to the variables (information) used in constructing the forecast; and *(iii)* to compare the sum (or mean) of squared forecasting errors of different forecast approaches. A non-zero mean of the errors clearly indicates a bias, and a non-zero correlation suggests that the information has not been used efficiently (a forecast error should not be predictable...)

Remark 4.20 *(Autocorrelation of forecast errors*) Suppose we make one-step-ahead forecasts, so we are forming a forecast of y_{t+1} based on what we know in period t. Let $e_{t+1} = y_{t+1} - \mathrm{E}_t\, y_{t+1}$, where $\mathrm{E}_t\, y_{t+1}$ denotes our forecast. If the forecast error is unforecastable, then the forecast errors cannot be autocorrelated, for instance, $\mathrm{Corr}(e_{t+1}, e_t) = 0$. For two-step-ahead forecasts, the situation is a bit different. Let $e_{t+2,t} = y_{t+2} - \mathrm{E}_t\, y_{t+2}$ be the error of forecasting y_{t+2} using the information in period t (notice: a two-step difference). If this forecast error is unforecastable using the information in period t, then the previously mentioned $e_{t+2,t}$ and $e_{t,t-2} = y_t - \mathrm{E}_{t-2}\, y_t$ must be uncorrelated—since the latter is known when the forecast $\mathrm{E}_t\, y_{t+2}$ is formed (assuming this forecast is efficient). However, there is nothing hat guarantees that $e_{t+2,t}$ and $e_{t+1,t-1} = y_{t+1} - \mathrm{E}_{t-1}\, y_{t+1}$ are uncorrected—since the latter contains new information compared to what was known when the forecast $\mathrm{E}_t\, y_{t+2}$ was formed. This generalizes to the following: an efficient h-step-ahead forecast error must have a zero correlation with the forecast error $h-1$ (and more) periods earlier.*

The comparison of forecast approaches/methods is not always a comparison of actual forecasts. Quite often, it is a comparison of a forecast method (or forecasting institute) with some kind of naive forecast like a "no change" or a random walk. The idea of such a comparison is to study if the resources employed in creating the forecast really bring value added compared to a very simple (and inexpensive) forecast.

It is sometimes argued that forecasting methods should not be ranked according to the sum (or mean) squared errors since this gives too much weight to a single large error. Ultimately, the ranking should be done based on the true benefits/costs of forecast errors—which may differ between organizations. For instance, a forecasting agency has a reputation (and eventually customers) to loose, while an investor has more immediate pecuniary losses. Unless the relation between the forecast error and the losses are immediately understood, the ranking of two forecast methods is typically done based on a number of different criteria. The following are often used:

1. mean error, $\Sigma_{t=1}^{T} e_t / T$,
2. mean squared error, $\Sigma_{t=1}^{T} e_t^2 / T$,
3. mean absolute error, $\Sigma_{t=1}^{T} |e_t| / T$,
4. fraction of times that the absolute error of method a smaller than that of method b,
5. fraction of times that method a predicts the direction of change better than method b,
6. profitability of a trading rule based on the forecast (for financial data),
7. results from a regression of the outcomes on two forecasts ($\hat{y}_t^a$ and $\hat{y}_t^b$)

$$y_t = \omega \hat{y}_t^a + \gamma \hat{y}_t^b + \text{ residual},$$

where $\omega = 1$ and $\gamma = 0$ indicates that forecast a contains all the information in b and more.

- A pseudo R^2 defined as $\text{Corr}(y_t, \hat{y}_t)^2$, where y_t is the actual value and $\hat{y}_t$ is the forecast.

As an example, Leitch and Tanner (1991) analyze the profits from selling 3-month T-bill futures when the forecasted interest rate is above futures rate (forecasted bill price is below futures price). The profit from this strategy is (not surprisingly) strongly related to measures of correct direction of change (see above), but (perhaps more surprisingly) not very strongly related to mean squared error, or absolute errors.

Example 4.21 *We want to compare the performance of the two forecast methods a and b. We have the following forecast errors* $(e_1^a, e_2^a, e_3^a) = (-1, -1, 2)$ *and* $(e_1^b, e_2^b, e_3^b) = (-1.9, 0, 1.9)$. *Both have zero means, so there is (in this very short sample) no constant bias. The mean squared errors are*

$$MSE^a = [(-1)^2 + (-1)^2 + 2^2]/3 = 2$$
$$MSE^b = [(-1.9)^2 + 0^2 + 1.9^2]/3 \approx 2.41,$$

so forecast a is better according to the mean squared errors criterion. The mean absolute errors are

$$MAE^a = [|-1| + |-1| + |2|]/3 \approx 1.33$$
$$MAE^b = [|-1.9| + |0| + |1.9|]/3 \approx 1.27,$$

so forecast b is better according to the mean absolute errors criterion. The reason for the difference between these criteria is that forecast b has fewer but larger errors—and the quadratic loss function punishes large errors very heavily. Counting the number of times the absolute error (or the squared error) is smaller, we see that a is better one time (first period), and b is better two times.

To perform formal tests of forecasting superiority a Diebold and Mariano (1995) test is typically performed. For instance to compare the MSE of two methods (a and b), first define

$$g_t = \left(e_t^a\right)^2 - \left(e_t^b\right)^2, \tag{4.54}$$

where e_t^i is the forecasting error of model i. Treating this as a GMM problem, we then test if

$$\mathrm{E}\, g_t = 0, \tag{4.55}$$

by applying a t-test on the same means

$$\frac{\bar{g}}{\mathrm{Std}(\bar{g})} \sim N(0,1), \text{ where } \bar{g} = \Sigma_{t=1}^T d_t / T, \tag{4.56}$$

and where the standard error is typically estimated using Newey-West (or similar) approach. However, when models a and b are nested, then the asymptotic distribution is non-normal so other critical values must be applied (see Clark and McCracken (2001)).

Other evaluation criteria can be used by changing (4.54). For instance, to test the mean absolute errors, use $g_t = |e_t^a| - |e_t^b|$ instead.

Remark 4.22 *From GMM we typically have* $\mathrm{Cov}(\sqrt{T}\bar{g}) = \sum_{s=-\infty}^{\infty} \mathrm{Cov}\left(g_t, g_{t-s}\right)$*, so for a scalar* g_t *wehe have* $\mathrm{Std}(\bar{g}) = \left(\sum_{s=-\infty}^{\infty} \mathrm{Cov}\left(g_t, g_{t-s}\right)/T\right)^{1/2}$*. When data happens to be iid, then this simplifies to* $\mathrm{Std}(\bar{g}) = \sqrt{\mathrm{Var}(g_t)/T} = \mathrm{Std}(g_t)/\sqrt{T}$.

4.7 Spurious Regressions and In-Sample Overfitting

References: Ferson, Sarkissian, and Simin (2003)

4.7.1 Spurious Regressions

Ferson, Sarkissian, and Simin (2003) argue that many prediction equations suffer from "spurious regression" features—and that data mining tends to make things even worse.

Their simulation experiment is based on a simple model where the return predictions are

$$r_{t+1} = \alpha + \delta Z_t + v_{t+1}, \tag{4.57}$$

where Z_t is a regressor (predictor). The true model is that returns follows the process

$$r_{t+1} = \mu + Z_t^* + u_{t+1}, \tag{4.58}$$

where the residual is white noise. In this equation, Z_t^* represents movements in expected returns. The predictors follow a diagonal VAR(1)

$$\begin{bmatrix} Z_t \\ Z_t^* \end{bmatrix} = \begin{bmatrix} \rho & 0 \\ 0 & \rho^* \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ Z_{t-1}^* \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \varepsilon_t^* \end{bmatrix}, \text{ with } \operatorname{Cov}\left(\begin{bmatrix} \varepsilon_t \\ \varepsilon_t^* \end{bmatrix} \right) = \Sigma. \tag{4.59}$$

In the case of a "pure spurious regression," the innovations to the predictors are uncorrelated (Σ is diagonal). In this case, δ ought to be zero—and their simulations show that the estimates are almost unbiased. Instead, there is a problem with the standard deviation of $\hat{\delta}$. If ρ^* is high, then the returns will be autocorrelated.

Under the null hypothesis of $\delta = 0$, this autocorrelated is loaded onto the residuals. For that reason, the simulations use a Newey-West estimator of the covariance matrix (with an automatic choice of lag order). This should, ideally, solve the problem with the inference—but the simulations show that it doesn't: when Z_t^* is very autocorrelated (0.95 or higher) and reasonably important (so an R^2 from running (4.58), if we could, would be 0.05 or higher), then the 5% critical value (for a t-test of the hypothesis $\delta = 0$) would be 2.7 (to be compared with the nominal value of 1.96). Since the point estimates are almost unbiased, the interpretation is that the standard deviations are underestimated. In contrast, with low autocorrelation and/or low importance of Z_t^*, the standard deviations are much more in line with nominal values.

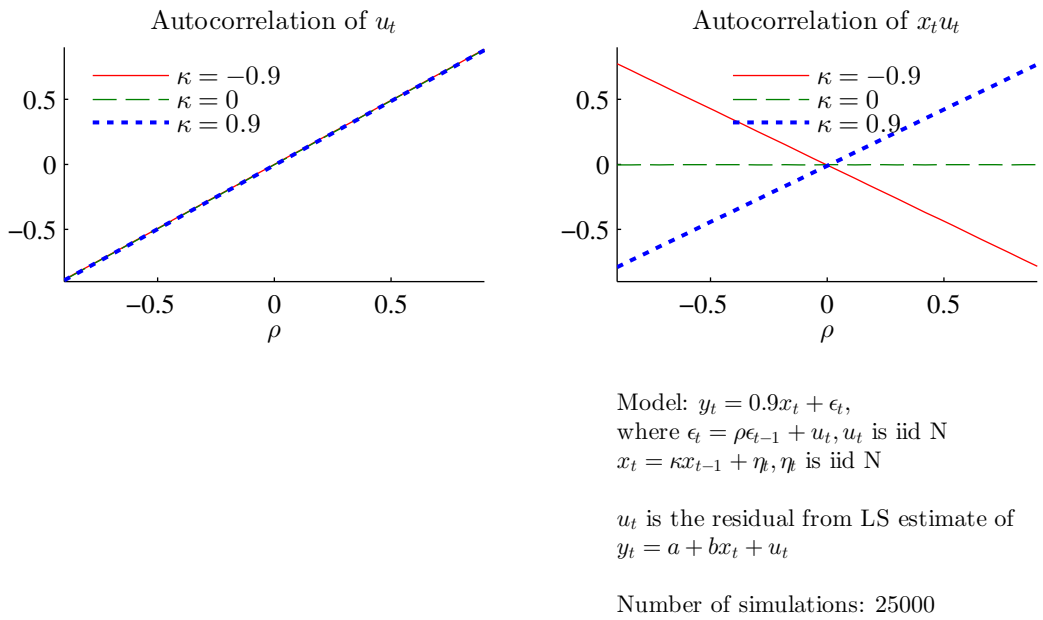


Figure 4.17: Autocorrelation of $x_t u_t$ when u_t has autocorrelation ρ

See *Figures 4.17–4.18* for an illustration. They show that we need a combination of an autocorrelated residuals and an autocorrelated regressor to create a problem for the usual LS formula for the standard deviation of a slope coefficient. When the autocorrelation is very high, even the Newey-West estimator is likely to underestimate the true uncertainty.

To study the interaction between spurious regressions and data mining, Ferson, Sarkissian, and Simin (2003) let Z_t be chosen from a vector of L possible predictors—which all are generated by a diagonal VAR(1) system as in (4.59) with uncorrelated errors. It is assumed that the researchers choose Z_t by running L regressions, and then picks the one with the highest R^2. When $\rho^* = 0.15$ and the researcher chooses between $L = 10$ predictors, the simulated 5% critical value is 3.5. Since this does not depend on the importance of Z_t^*, it is interpreted as a typical feature of "data mining," which is bad enough. When the autocorrelation is 0.95, then the importance of Z_t^* start to become important—"spurious regressions" interact with the data mining to create extremely high simulated critical values. A possible explanation is that the data mining exercise is likely to pick out the most autocorrelated predictor, and that a highly autocorrelated predictor exacerbates the spurious regression problem.

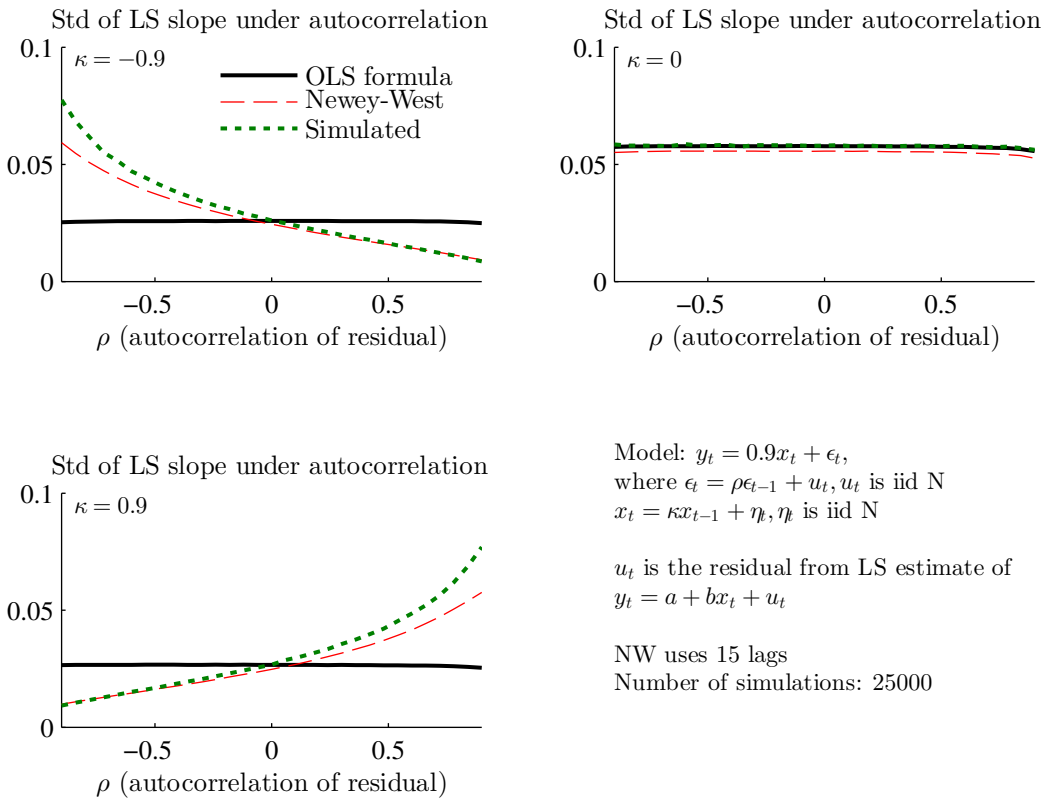


Figure 4.18: Standard error of OLS estimator, autocorrelated errors

4.8 Out-of-Sample Forecasting Performance

4.8.1 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"

$$R^2_{OS} = 1 - \sum_{t=s}^{T} (r_t - \hat{r}_t)^2 \,/\, \sum_{t=s}^{T} (r_t - \tilde{r}_t)^2 \,, \tag{4.60}$$

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.

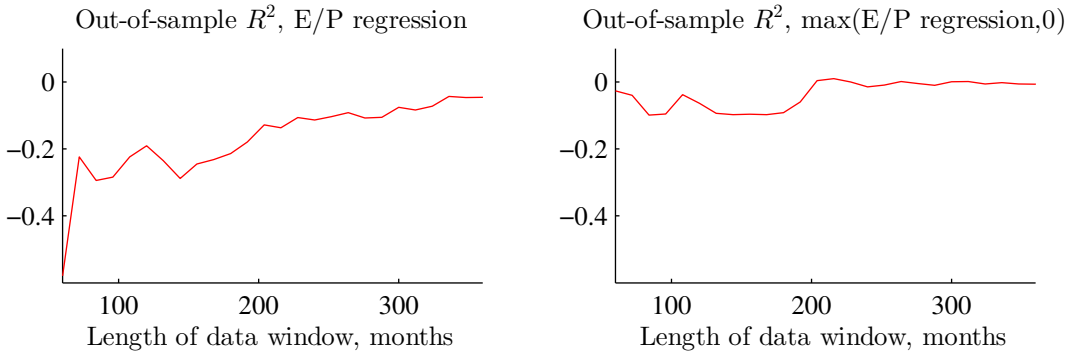
The evidence shows that the out-of-sample forecasting performance is very weak—as claimed by Goyal and Welch (2008).

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 4.19 for an illustration.

4.8.2 More Evidence on Out-of-Sample Forecasting Performance

Figures 4.20–4.24 illustrate the *out-of-sample performance on daily returns*. Figure 4.20 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 4.21 tries a simpler strategies: 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 4.20 are actually dominated by smaller momentum type changes (positive autocorrelation). However, always holding the S&P 500



US stock returns (1-year, in excess of riskfree) 1926:1-2011:12

Estimation is done on moving data window,
forecasts are made out of sample for: 1957:1-2011:12

Figure 4.19: Predictability of US stock returns, in-sample and out-of-sample

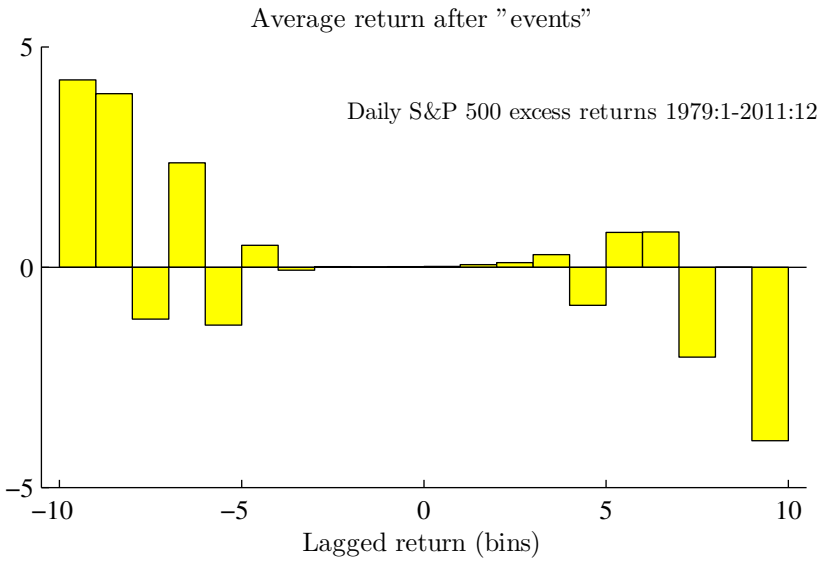


Figure 4.20: Short-run predictability of US stock returns, out-of-sample

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 4.22 shows that the momentum strategy

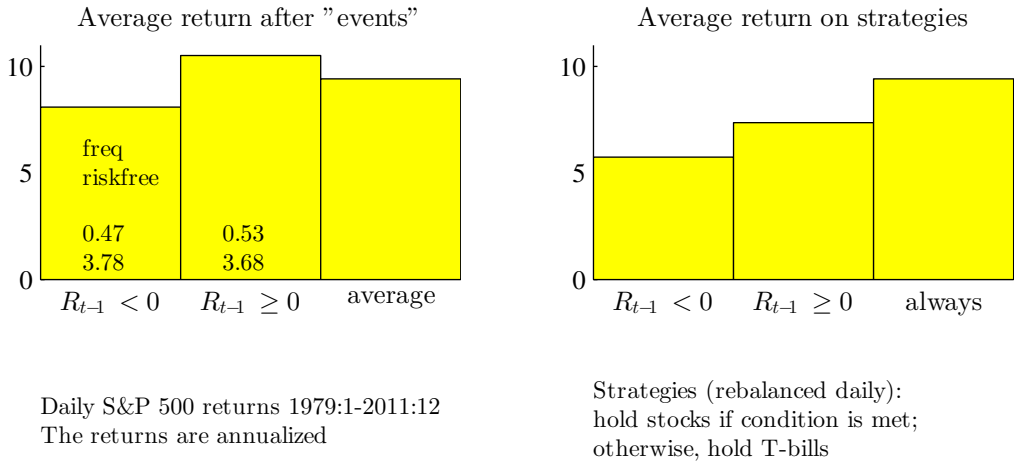


Figure 4.21: Short-run predictability of US stock returns, out-of-sample

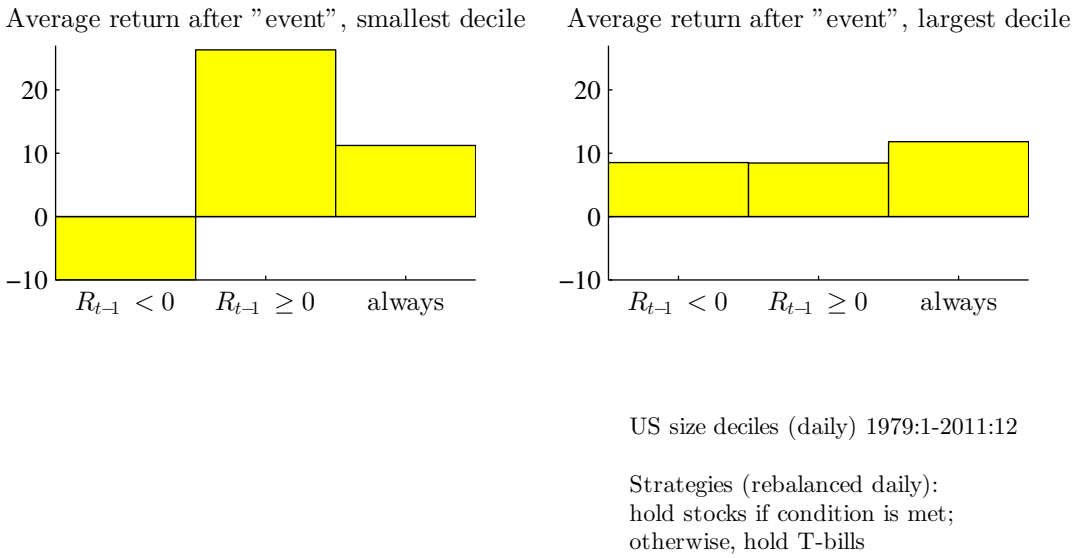
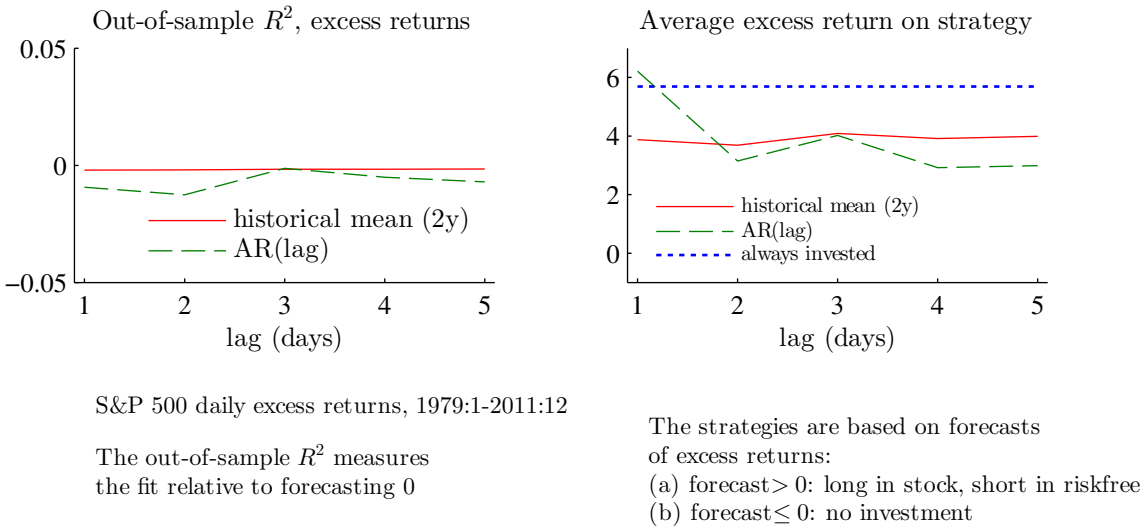


Figure 4.22: Short-run predictability of US stock returns, out-of-sample

works reasonably well on small stocks.

Figure 4.23 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 histor-



S&P 500 daily excess returns, 1979:1-2011:12

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 4.23: Short-run predictability of US stock returns, out-of-sample

ical 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 4.24 compares the results for small and large stocks—and illustrates that there is more predictability for small stocks.

Figures 4.25–4.27 illustrate the *out-of-sample performance on long-run returns*. Figure 4.25 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, 4.26 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 4.27 studies the out-of-sample R^2 for simple forecasting models, and also allows 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.

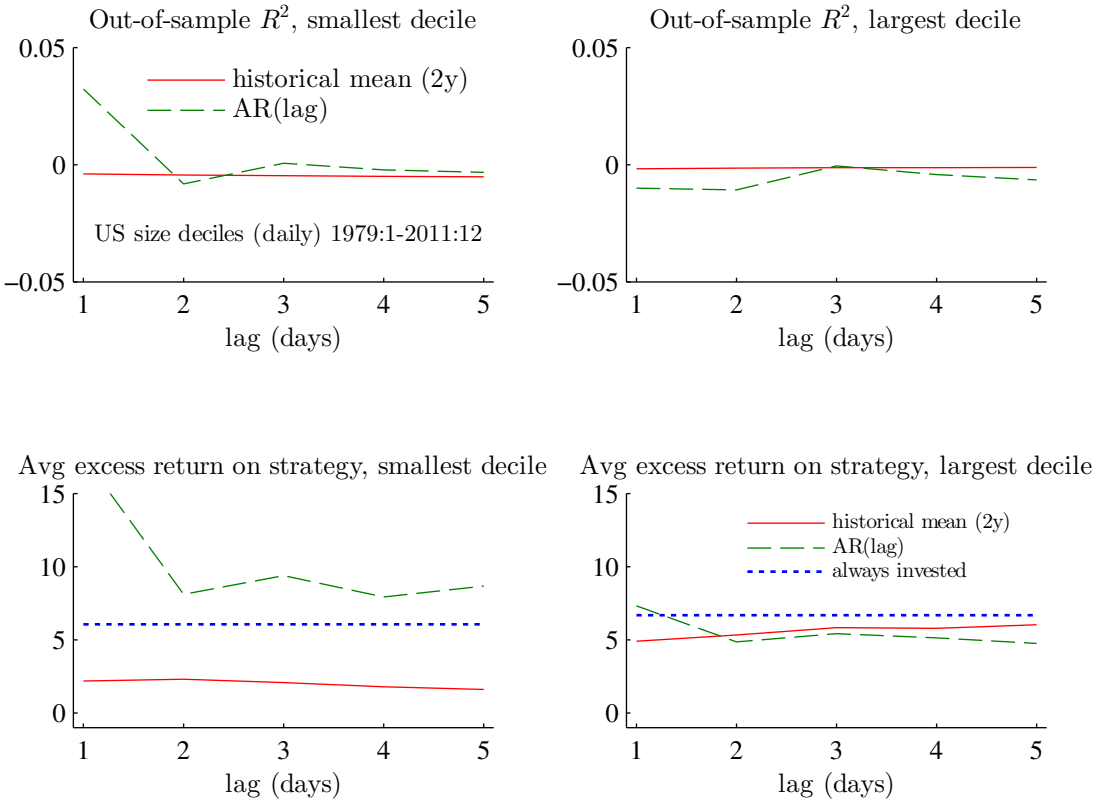


Figure 4.24: Short-run predictability of US stock returns, out-of-sample. See Figure 4.23 for details on the strategies.

4.8.3 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")

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 effi-

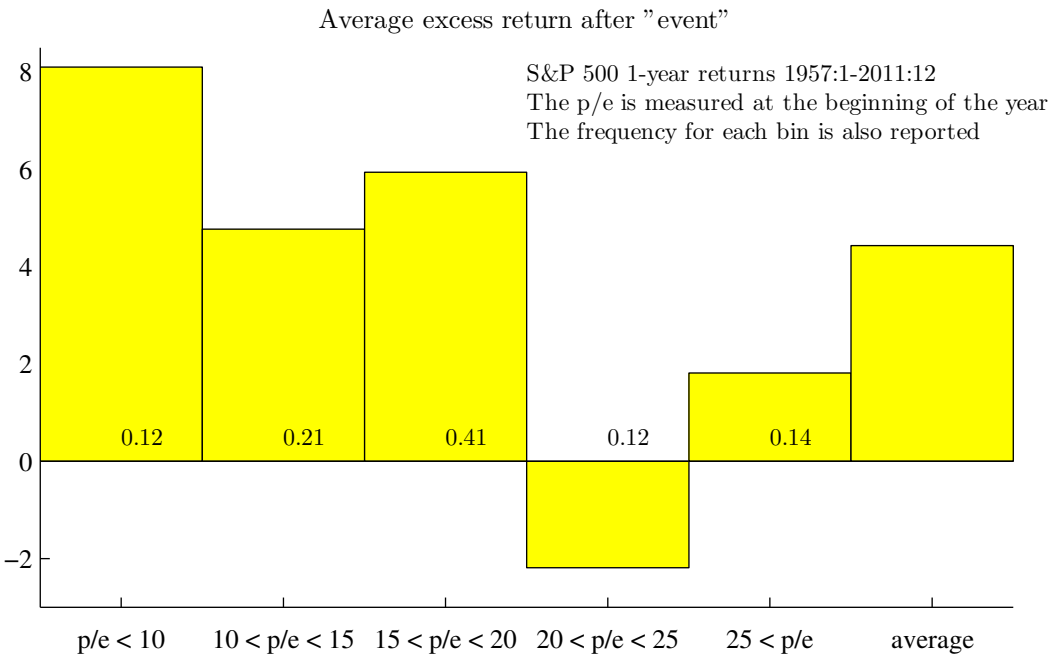


Figure 4.25: Long-run predictability of US stock returns, out-of-sample

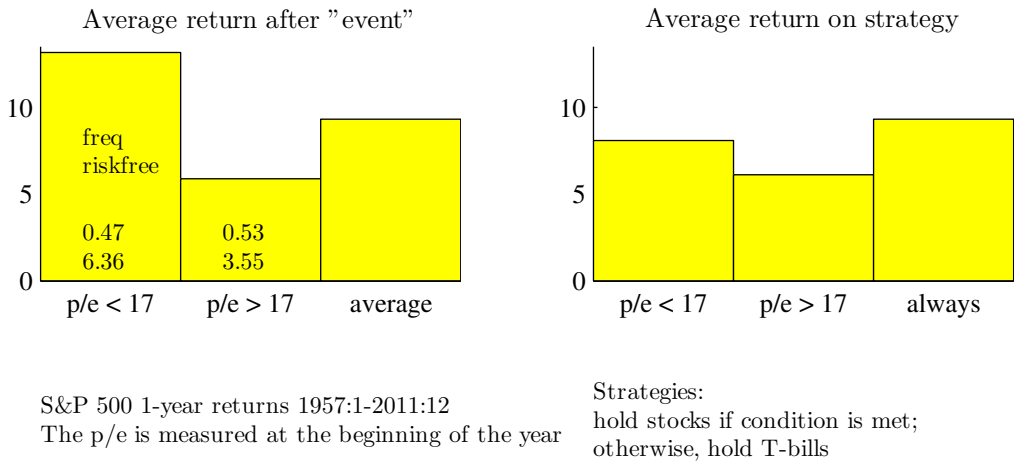
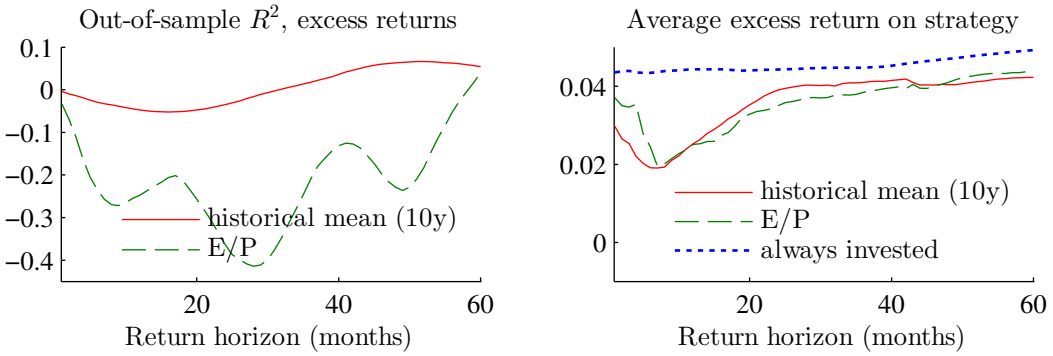


Figure 4.26: Long-run predictability of US stock returns, out-of-sample

cient: 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 is not very important) and that old patterns can tell us where the price will move in the near future. This is an



Monthly US stock returns in excess of riskfree rate
Estimation is done on moving data window,
forecasts are made out of sample for 1957:1-2011:12

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 4.27: Long-run predictability of US stock returns, out-of-sample

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 *filter rule* like "buy after an increase of $x\%$ and sell after a decrease of $y\%$" is clearly based on the perception that the current price movement will continue.

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

$$\begin{bmatrix} \text{buy in } t \text{ if} & MA_{t-1}(S) > MA_{t-1}(L)(1+b) \\ \text{sell in } t \text{ if} & MA_{t-1}(S) < MA_{t-1}(L)(1-b) \\ \text{no change} & \text{otherwise} \end{bmatrix}, \text{ where} \tag{4.61}$$

$$MA_{t-1}(S) = (p_{t-1} + \ldots + p_{t-S})/S.$$

The difference between the two moving averages is called an *oscillator* (or sometimes, moving average convergence divergence[2]). A version of the moving average oscillator is the *relative strength index*[3], which is the ratio of average price level on "up" days to the average price on "down" days—during the last z (14 perhaps) days.

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 window of length L. With a bandwidth b (perhaps 0.01), the rule for period t could be

$$\begin{bmatrix} \text{buy in } t \text{ if} & P_t > M_{t-1}(1+b) \\ \text{sell in } t \text{ if} & P_t < m_{t-1}(1-b) \\ \text{no change} & \text{otherwise} \end{bmatrix}, \text{ where} \tag{4.62}$$

$$M_{t-1} = \max(p_{t-1}, \ldots, p_{t-S})$$
$$m_{t-1} = \min(p_{t-1}, \ldots, p_{t-S}).$$

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

[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).

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 is high, 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.)

"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 (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 4.28–4.29* for an illustration.

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.

Some investors argue that markets show periods of mean reversion and then periods with trends—an that both can be exploited. Clearly, the concept of a 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.

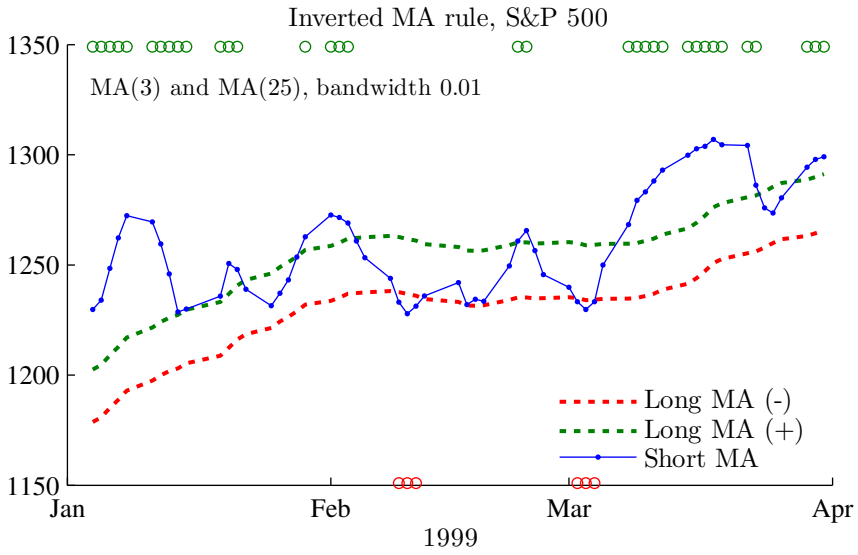


Figure 4.28: Examples of trading rules.

4.9 Security Analysts

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 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).

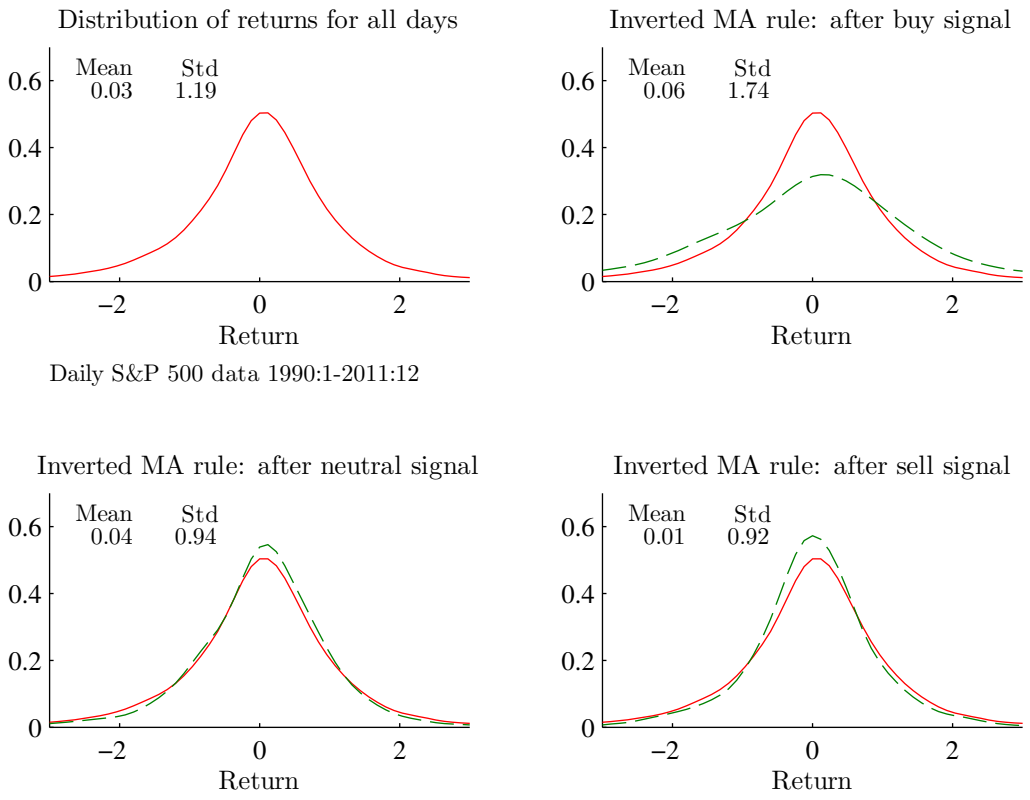


Figure 4.29: Examples of trading rules.

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.

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.

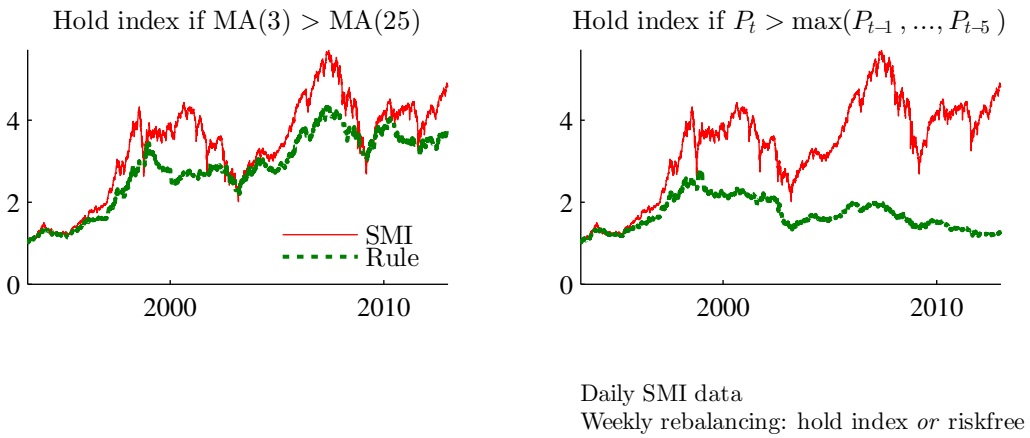


Figure 4.30: Examples of trading rules applied to SMI. The rule portfolios are rebalanced every Wednesday: if condition (see figure titles) is satisfied, then the index is held for the next week, otherwise a government bill is held. The figures plot the portfolio values.

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."

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).

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.

Bibliography

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.

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., and J. H. Cochrane, 1999, "By force of habit: a consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, 107, 205–251.

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.

Campbell, J. Y., and L. M. Viceira, 1999, "Consumption and portfolio decisions when expected returns are time varying," *Quarterly Journal of Economics*, 114, 433–495.

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.

Clark, T. E., and M. W. McCracken, 2001, "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105, 85–110.

Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.

Diebold, F. X., 2001, *Elements of forecasting*, South-Western, 2nd edn.

Diebold, F. X., and R. S. Mariano, 1995, "Comparing predcitve accuracy," *Journal of Business and Economic Statistics*, 13, 253–265.

Epstein, L. G., and S. E. Zin, 1991, "Substitution, risk aversion, and the temporal behavior of asset returns: an empirical analysis," *Journal of Political Economy*, 99, 263–286.

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.

Leitch, G., and J. E. Tanner, 1991, "Economic forecast evaluation: profit versus the conventional error measures," *American Economic Review*, 81, 580–590.

Lo, A. W., and A. C. MacKinlay, 1997, "Maximizing predictability in the stock and bond markets," *Macroeconomic Dynamics*, 1, 102–134.

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., 2006, "C-CAPM Refinements and the cross-section of returns," *Financial Markets and Portfolio Management*, 20, 49–73.

Söderlind, P., 2010, "Predicting stock price movements: regressions versus economists," *Applied Economics Letters*, 17, 869–874.

Stekler, H. O., 1991, "Macroeconomic forecast evaluation techniques," *International Journal of Forecasting*, 7, 375–384.

Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.

The Economist, 1993, "Frontiers of finance," pp. 5–20.

5 Predicting and Modelling Volatility

Sections denoted by a star (*) is not required reading.

Reference: Campbell, Lo, and MacKinlay (1997) 12.2; Taylor (2005) 8–11; Hamilton (1994) 21; Hentschel (1995); Franses and van Dijk (2000); Andersen, Bollerslev, Christoffersen, and Diebold (2005)

5.1 Heteroskedasticity

5.1.1 Descriptive Statistics of Heteroskedasticity (Realized Volatility)

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 *realized variances* from "rolling samples." For a zero-mean variable, u_t, this could mean

$$\sigma_t^2 = \frac{1}{q}\sum_{s=1}^{q} u_{t-s}^2 = (u_{t-1}^2 + u_{t-2}^2 + \ldots + u_{t-q}^2)/q, \tag{5.1}$$

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. Unfortunately, this method can produce quite abrupt changes in the estimate.

See Figures 5.1–5.3 for illustrations.

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 be $(1-\lambda)\lambda^s$ where $0 < \lambda < 1$, so

$$\sigma_t^2 = (1-\lambda)\sum_{s=1}^{\infty} \lambda^{s-1} u_{t-s}^2 = (1-\lambda)(u_{t-1}^2 + \lambda u_{t-2}^2 + \lambda^2 u_{t-3}^2 + \ldots), \tag{5.2}$$

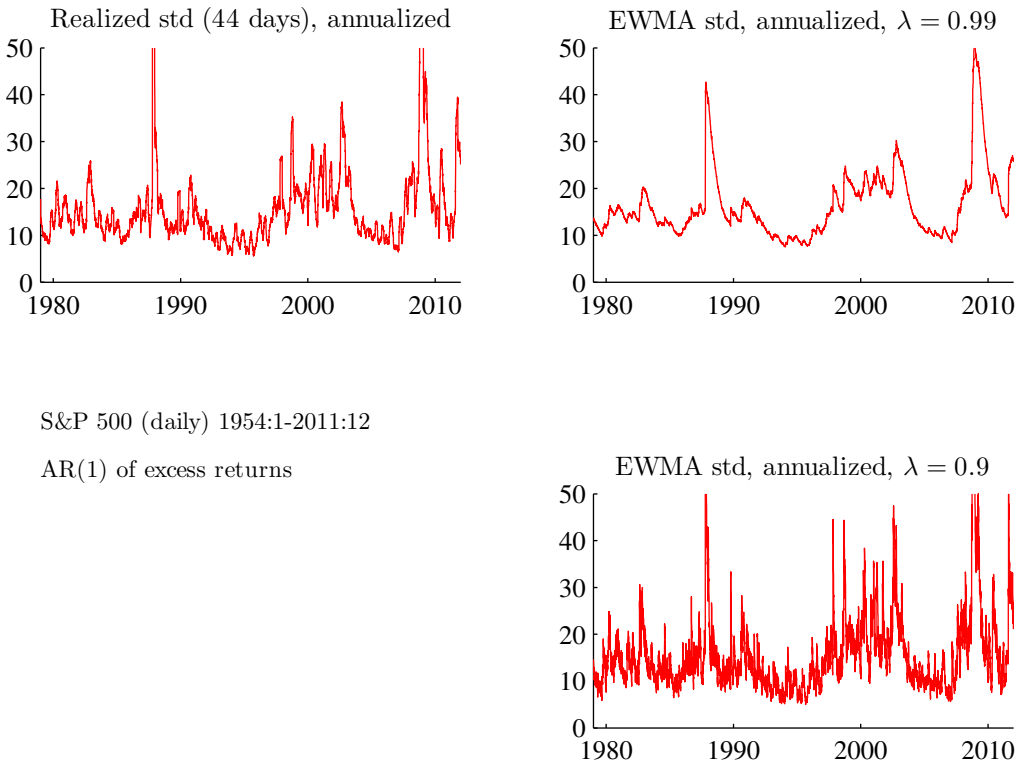


Figure 5.1: Standard deviation

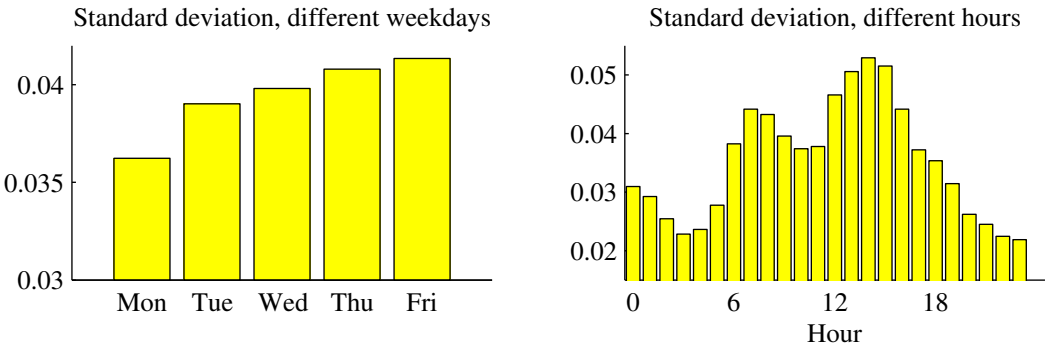


Figure 5.2: Standard deviation of EUR/USD exchange rate changes

which can also be calculated in a recursive fashion as

$$\sigma_t^2 = (1-\lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2. \tag{5.3}$$

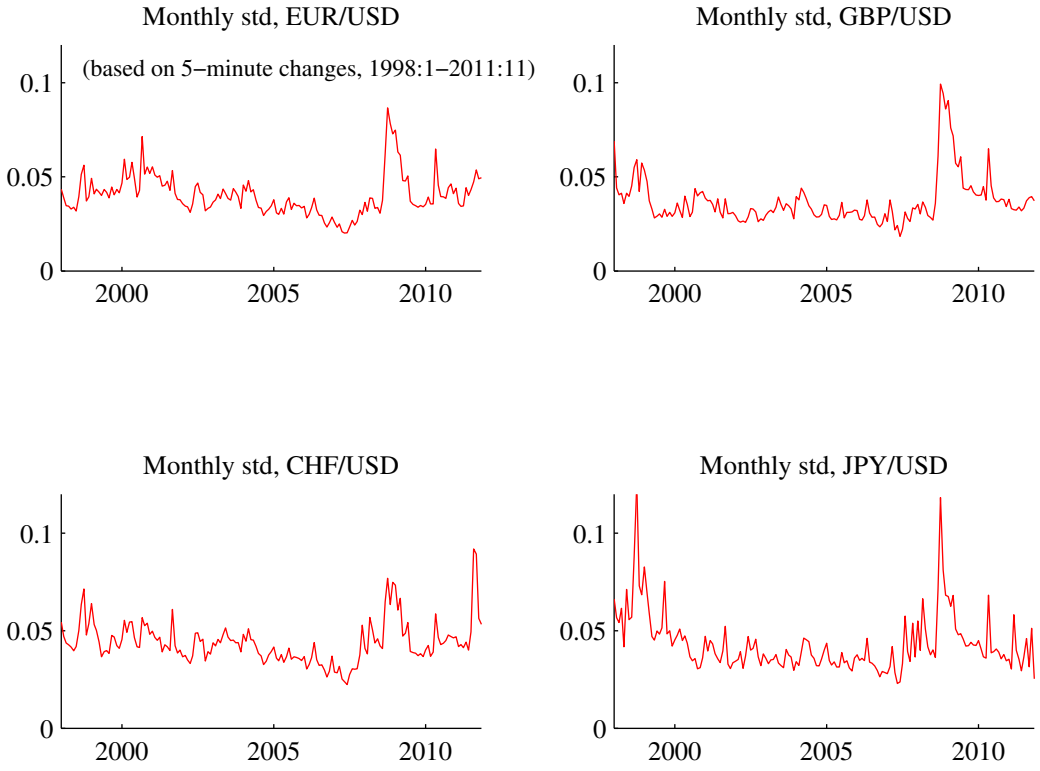


Figure 5.3: Standard deviation of exchange rate changes

The initial value (before the sample) could be assumed to be zero or (better) the unconditional variance in a historical sample. The EWMA is commonly used by practitioners. For instance, the RISK Metrics (formerly part of JP Morgan) 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 5.4 for an illustration of the weights.

Remark 5.1 *(VIX) Although VIX is based on option prices, it is calculated in a way that makes it (an estimate of) the risk-neutral expected variance until expiration, not the implied volatility, see Britten-Jones and Neuberger (2000) and Jiang and Tian (2005).*

See Figure 5.5 for an example.

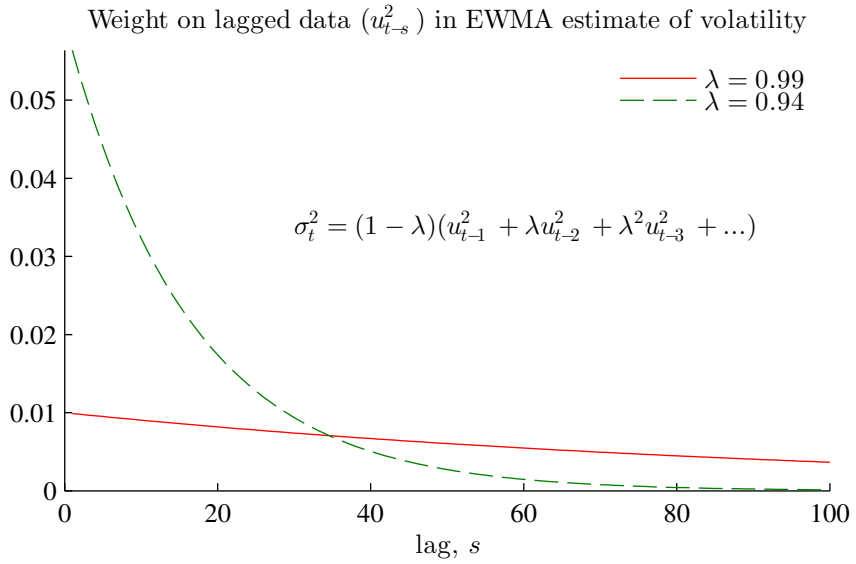


Figure 5.4: Weights on old data in the EMA approach to estimate volatility

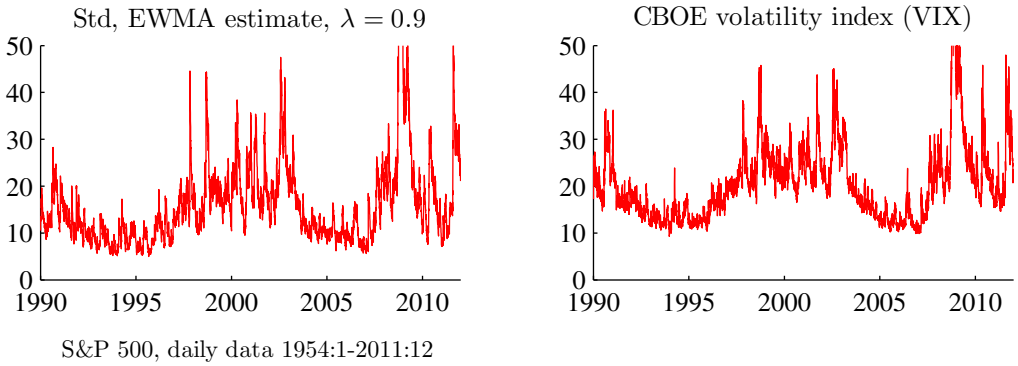


Figure 5.5: Different estimates of US equity market volatility

We can also estimate the realized covariance of two series (u_{it} and u_{jt}) by

$$\sigma_{ij,t} = \frac{1}{q}\sum_{s=1}^{q} u_{i,t-s}u_{j,t-s} = (u_{i,t-1}u_{j,t-1} + u_{i,t-2}u_{j,t-2} + \ldots + u_{i,t-q}u_{j,t-q})/q, \quad (5.4)$$

as well as the EWMA

$$\sigma_{ij,t} = (1-\lambda)u_{i,t-1}u_{j,t-1} + \lambda\sigma_{ij,t-1}. \quad (5.5)$$

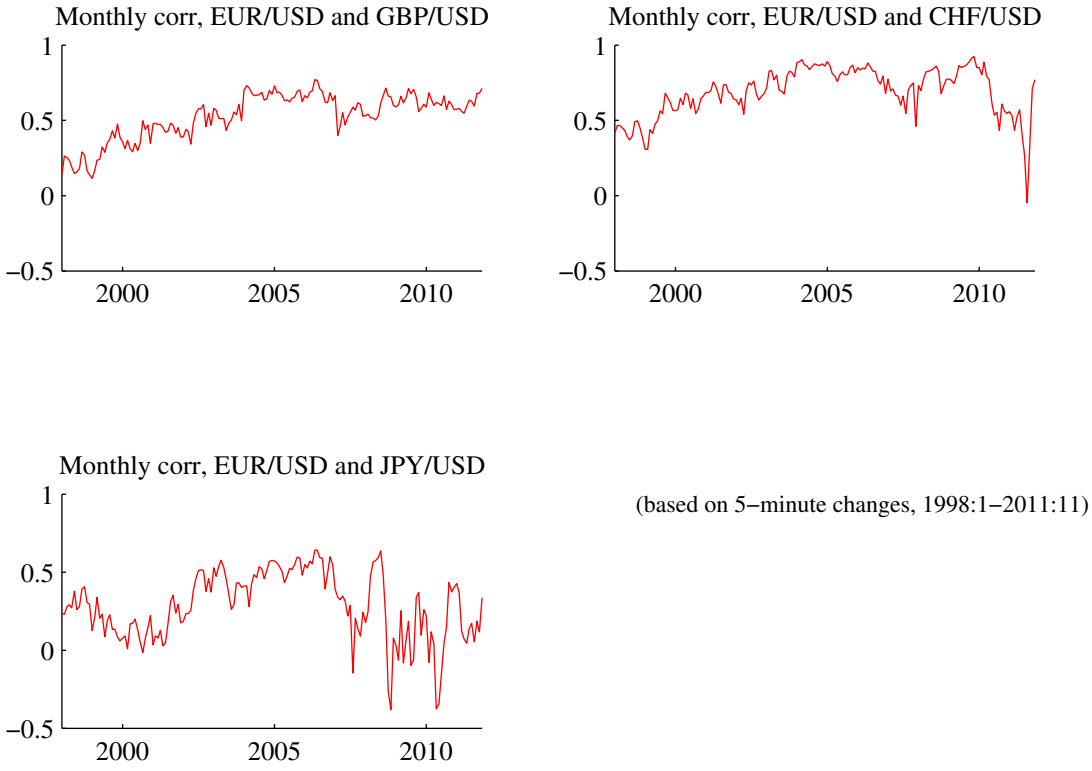


Figure 5.6: Correlation of exchange rate changes

By combining with the estimates of the variances, it is straightforward to estimate correlations.

See Figures 5.6–5.7 for illustrations.

5.1.2 Variance and Volatility Swaps

Instead of investing in straddles, it is also possible to invest in *variance swaps*. Such a contract has a zero price in inception (in t) and the payoff at expiration (in $t + m$) is

$$\text{Variance swap payoff}_{t+m} = \text{realized variance}_{t+m} - \text{variance swap rate}_t, \tag{5.6}$$

where the variance swap rate (also called the strike or forward price for) is agreed on at inception (t) and the realized volatility is just the sample variance for the swap period. Both rates are typically annualized, for instance, if data is daily and includes only trading

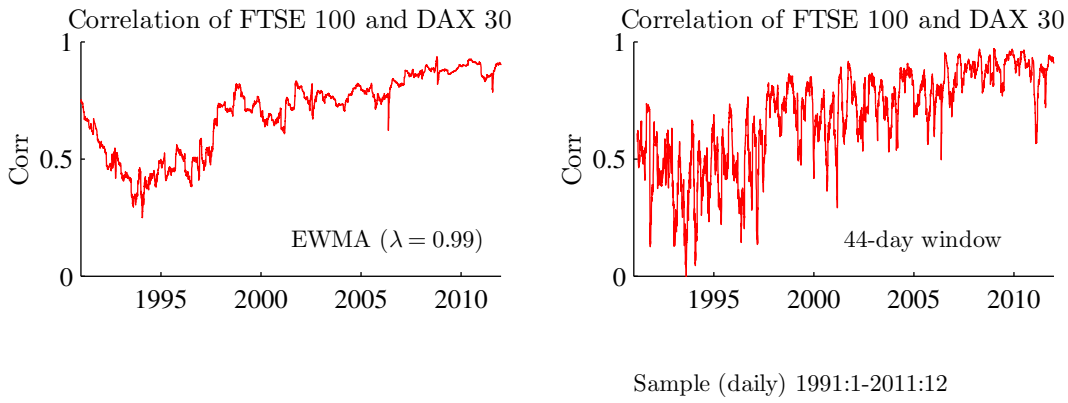


Figure 5.7: Time-varying correlations (EWMA and realized)

days, then the variance is multiplied by 252 or so (as a proxy for the number of trading days per year).

A *volatility swap* is similar, except that the payoff it is expressed as the difference between the standard deviations instead of the variances

$$\text{Volatility swap payoff}_{t+m} = \sqrt{\text{realized variance}_{t+m}} - \text{volatility swap rate}_t, \tag{5.7}$$

If we use daily data to calculate the realized variance from t until the expiration(RV_{t+m}), then

$$RV_{t+m} = \frac{252}{m}\sum_{s=1}^{m} R_{t+s}^2, \tag{5.8}$$

where R_{t+s} is the net return on day $t+s$. (This formula assumes that the mean return is zero—which is typically a good approximation for high frequency data. In some cases, the average is taken only over $m-1$ days.)

Notice that both variance and volatility swaps pays off if actual (realized) volatility between t and $t+m$ is higher than expected in t. In contrast, the futures on the VIX pays off when the expected volatility (in $t+m$) is higher than what was thought in t. In a way, we can think of the VIX futures as a futures on a volatility swap (between $t+m$ and a month later).

Since VIX^2 is a good approximation of variance swap rate for a 30-day contract, the return can be approximated as

$$\text{Return of a variance swap}_{t+m} = (RV_{t+m} - VIX_t^2)/VIX_t^2. \tag{5.9}$$

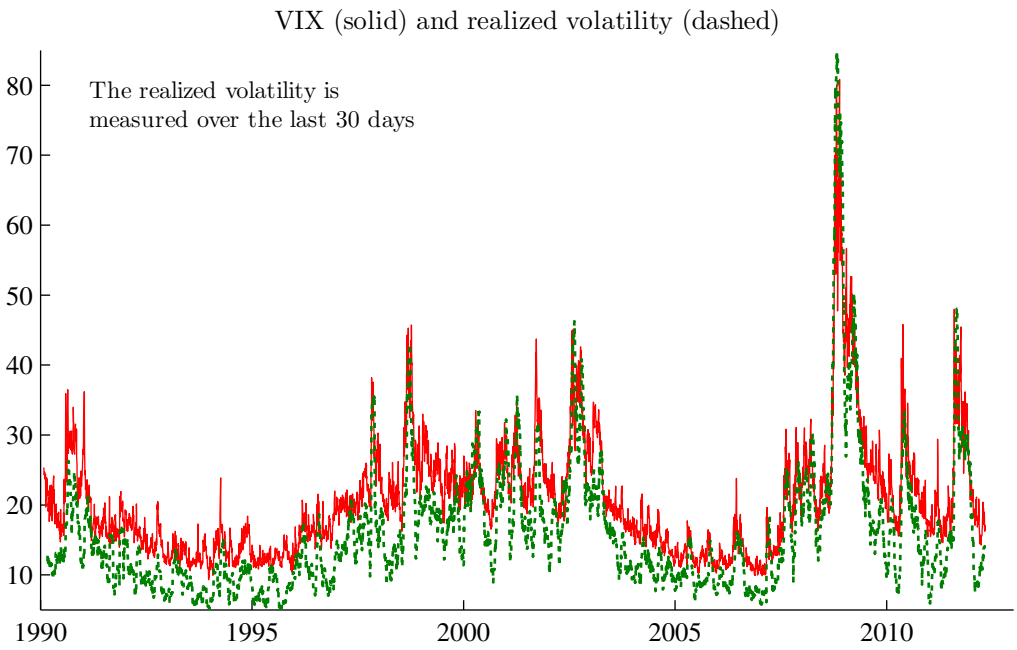


Figure 5.8: VIX and realized volatility (variance)

Figures 5.8 and 5.9 illustrate the properties for the VIX and realized volatility of the S&P 500. It is clear that the mean return of a variance swap (with expiration of 30 days) would have been negative on average. (Notice: variance swaps were not traded for the early part of the sample in the figure.) The excess return (over a riskfree rate) would, of course, have been even more negative. This suggests that selling variance swaps (which has been the speciality of some hedge funds) might be a good deal—except that it will incur some occasional really large losses (the return distribution has positive skewness). Presumably, buyers of the variance swaps think that this negative average return is a reasonable price to pay for the "hedging" properties of the contracts—although the data does not suggest a very strong negative correlation with S&P 500 returns.

5.1.3 Forecasting Realized Volatility

Implied volatility from options (iv) should contain information about future volatility—as is therefore often used as a predictor. It is unclear, however, if the iv is more informative than recent (actual) volatility, especially since they are so similar—see Figure 5.8.

Table 5.1 shows that the iv (here represented by VIX) is close to be an unbiased

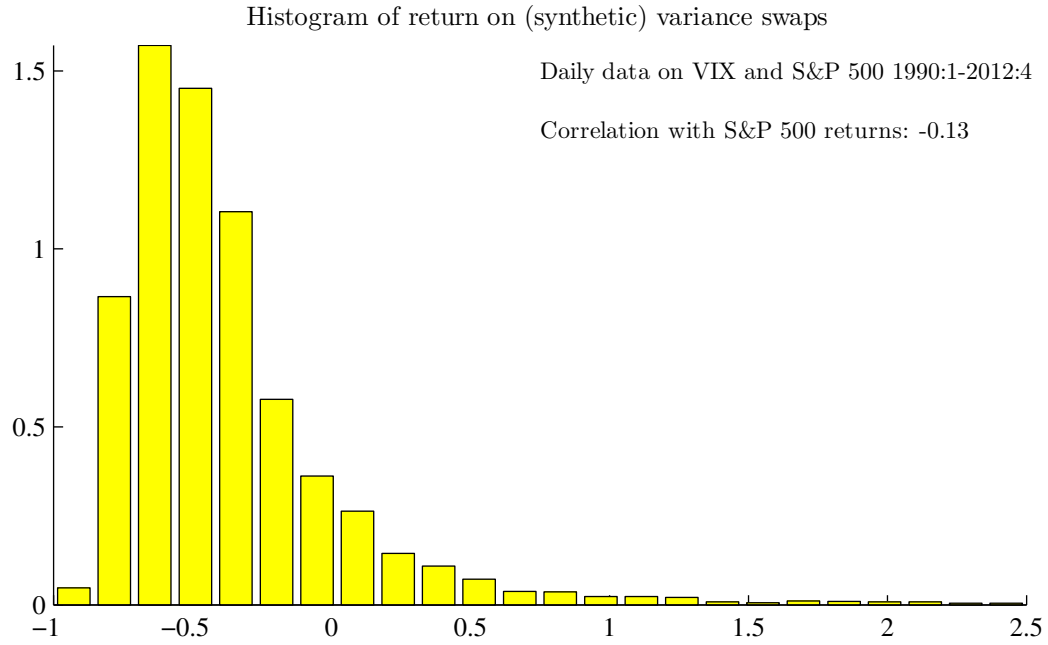


Figure 5.9: Distribution of return from investing in variance swaps

predictor of future realized volatility since the slope coefficient is close to one. However, the intercept is negative, which suggests that the iv overestimate future realized volatility. This is consistent with the presence of risk premia in the iv, but also with subjective beliefs (pdfs) that are far from looking like normal distributions. By using both iv and the recent realized volatility, the forecast powers seems to improve.

Remark 5.2 *(Restricting the predicted volatility to be positive) A linear regression (like those in Table 5.1) can produce negative volatility forecasts. An easy way to get around that is to specify the regression in terms on the log volatility.*

Remark 5.3 *(Restricting the predicted correlation to be between* -1 *and* 1*) The perhaps easiest way to do that is to specify the regression equation in terms of the Fisher transformation,* $z = 1/2 \ln[(1+\rho)/(1-\rho)]$*, where* ρ *is the correlation coefficient. The correlation coefficient can then be calculated by the inverse transformation* $\rho = [\exp(2z)-1]/[\exp(2z)+1]$.

	(1)	(2)	(3)
lagged RV	0.75		0.27
	(10.98)		(2.20)
lagged VIX		0.91	0.63
		(12.54)	(7.25)
constant	4.01	−2.64	−1.16
	(4.26)	(−2.05)	(−1.48)
R2	0.56	0.60	0.62
obs	5555.00	5575.00	5555.00

Table 5.1: Regression of 22-day realized S&P return volatility 1990:1-2012:4. 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.

	Corr(EUR,GBP)	Corr(EUR,CHF)	Corr(EUR,JPY)
lagged Corr(EUR,GBP)	0.91		
	(28.94)		
lagged Corr(EUR,CHF)		0.87	
		(11.97)	
lagged Corr(EUR,JPY)			0.81
			(16.84)
constant	0.05	0.09	0.05
	(2.97)	(1.76)	(2.83)
R2	0.85	0.76	0.66
obs	166.00	166.00	166.00

Table 5.2: Regression of monthly realized correlations 1998:1-2011:11. All exchange rates are against the USD. The monthly correlations are calculated from 5 minute data. Numbers in parentheses are t-stats, based on Newey-West with 1 lag.

5.1.4 Heteroskedastic Residuals in a Regression

Suppose we have a regression model

$$y_t = x_t'b + \varepsilon_t, \text{ where} \tag{5.10}$$
$$\mathrm{E}\,\varepsilon_t = 0 \text{ and } \mathrm{Cov}(x_{it}, \varepsilon_t) = 0.$$

	RV(EUR)	RV(GBP)	RV(CHF)	RV(JPY)
lagged RV(EUR)	0.62			
	(7.59)			
lagged RV(GBP)		0.73		
		(10.70)		
lagged RV(CHF)			0.33	
			(2.59)	
lagged RV(JPY)				0.56
				(5.12)
constant	0.12	0.07	0.29	0.20
	(3.40)	(2.51)	(3.99)	(2.97)
D(Tue)	0.04	0.02	0.07	0.00
	(2.91)	(1.55)	(2.15)	(0.11)
D(Wed)	0.06	0.06	0.04	0.06
	(4.15)	(3.97)	(1.53)	(1.92)
D(Thu)	0.07	0.06	0.09	0.08
	(4.86)	(3.24)	(3.90)	(1.83)
D(Fri)	0.08	0.04	0.09	0.06
	(3.54)	(2.04)	(5.19)	(1.67)
R2	0.39	0.53	0.11	0.31
obs	3629.00	3629.00	3629.00	3629.00

Table 5.3: 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.

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 heteroskedasticity, 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 5.11.

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 (5.10) with an ARCH structure of the residuals—and estimate the whole thing with maximum likelihood (MLE) is one way. As a by-product

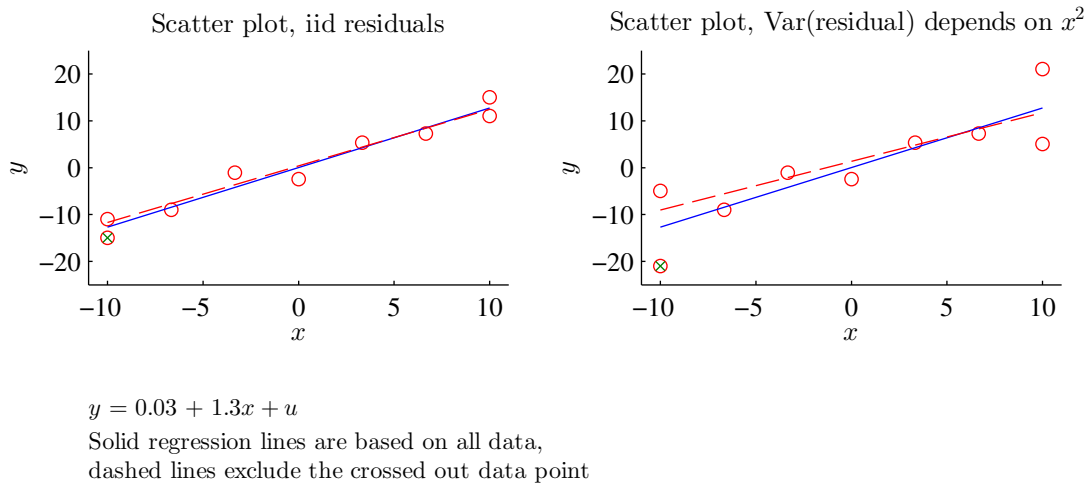


Figure 5.10: Effect of heteroskedasticity on uncertainty about regression line

we get the correct standard errors provided, of course, the assumed distribution is correct. Second, we could stick to OLS, but use another expression for the variance of the coefficients: a "heteroskedasticity consistent covariance matrix," among which "White's covariance matrix" is the most common.

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 (denoted w_t)—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.

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

$$\hat{\varepsilon}_t^2 = w_t'\gamma + v_t, \tag{5.11}$$

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

Example 5.4 *(White's test) If the regressors include* $(1, x_{1t}, x_{2t})$ *then* w_t *in (5.11) is the vector* $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$.

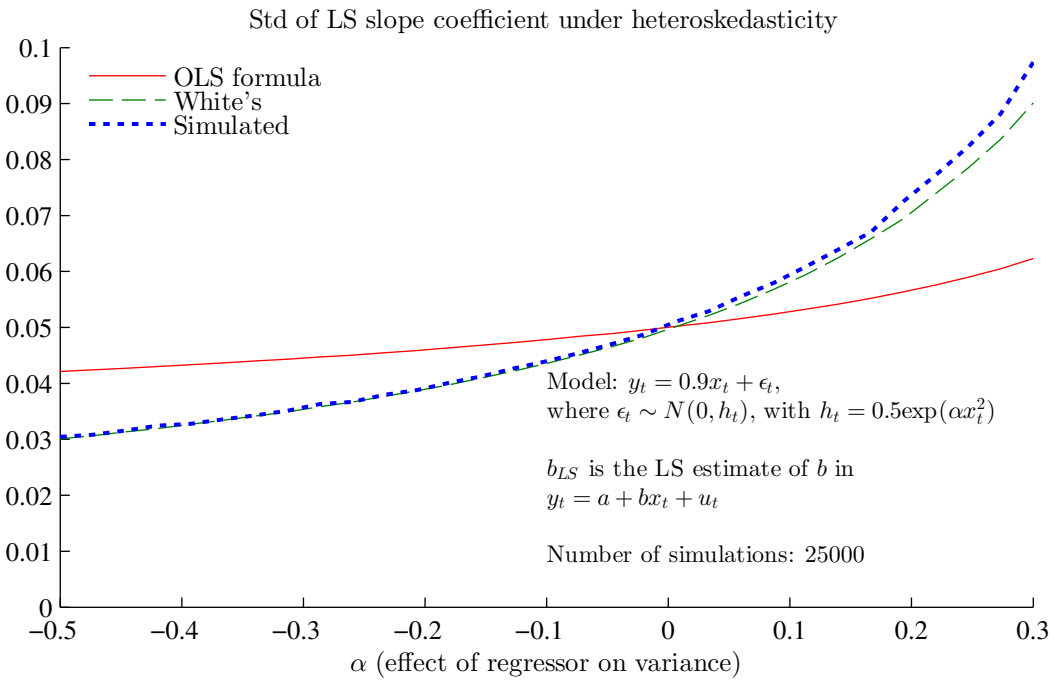


Figure 5.11: Variance of OLS estimator, heteroskedastic errors

5.1.5 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)

$$u_t^2 = \omega + a_1 u_{t-1}^2 + \ldots + a_q u_{t-q}^2 + v_t, \tag{5.12}$$

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 (5.10). However, in this case, it is not obvious that ARCH effects makes the standard expression for the LS covariance matrix invalid—this is tested by White's test as in (5.11).

It is straightforward to phrase Engle's test in terms of GMM moment conditions. We

simply use a first set of moment conditions to estimate the parameters of the regression model, and then test if the following additional (ARCH related) moment conditions are satisfied at those parameters

$$\mathrm{E}\left[\begin{array}{c} u_{t-1}^2 \\ \vdots \\ u_{t-q}^2 \end{array}\right](u_t^2 - a_0) = \mathbf{0}_{q\times 1}. \tag{5.13}$$

An alternative test (see Harvey (1989) 259–260), is to apply a Box-Ljung test on $\hat{u}_t^2$, to see if the squared fitted residuals are autocorrelated. We just have to adjust the degrees of freedom in the asymptotic chi-square distribution by subtracting the number of parameters estimated in the regression equation. These tests for ARCH effects will typically capture GARCH (see below) effects as well.

5.2 ARCH Models

Consider the regression model

$$\begin{aligned} y_t &= x_t'b + u_t, \text{ where} \\ \mathrm{E}\, u_t &= 0 \text{ and } \mathrm{Cov}(x_{it}, u_t) = 0. \end{aligned} \tag{5.14}$$

We will study different ways of modelling how the volatility of the residual is autocorrelated.

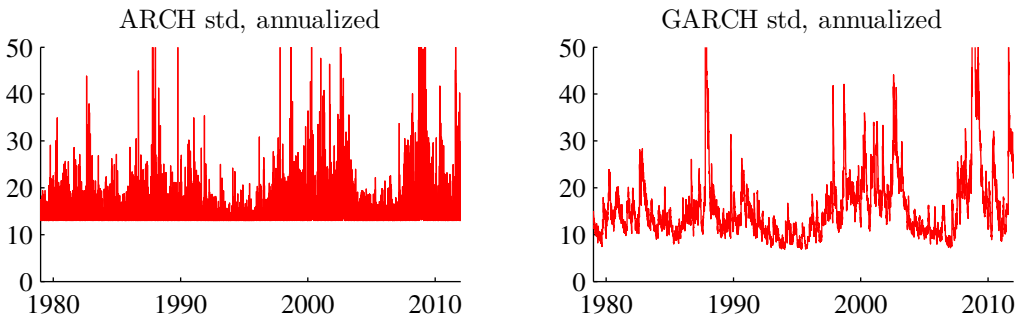
5.2.1 Properties of ARCH(1)

In the ARCH(1) model the residual in the regression equation (5.14) can be written

$$\begin{aligned} u_t &= v_t\sigma_t, \text{ with} \\ v_t &\sim \text{iid with } \mathrm{E}\, v_t = 0 \text{ and } \mathrm{Var}(v_t) = 1, \end{aligned} \tag{5.15}$$

and the conditional variance is generated by

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha u_{t-1}^2, \text{ with} \\ \omega &> 0 \text{ and } 0 \leq \alpha < 1. \end{aligned} \tag{5.16}$$



S&P 500 (daily) 1954:1-2011:12

AR(1) of excess returns
with ARCH(1) or GARCH(1,1) errors

AR(1) coef: 0.10
ARCH coef: 0.32
GARCH coefs: 0.08 0.91

Figure 5.12: ARCH and GARCH estimates

Notice that σ_t^2 is the conditional variance of u_t, and it is known already in $t-1$. (Warning: some authors use a different convention for the time subscripts.) We also assume that v_t is truly random, and hence independent of σ_t^2.

See Figure 5.12 for an illustration.

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. To see the latter, notice that the forecast (made in t) of volatility in $t+s$ is (since σ_{t+1}^2 is known in t)

$$\mathrm{E}_t\, \sigma_{t+s}^2 = \bar{\sigma}^2 + \alpha^{s-1}\left(\sigma_{t+1}^2 - \bar{\sigma}^2\right), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1-\alpha}, \tag{5.17}$$

where $\bar{\sigma}^2$ is the unconditional variance. The forecast of the variance is just like in an AR(1) process. A value of $\alpha < 1$ is needed to make the difference equation stable.

The conditional variance of u_{t+s} is clearly equal to the expected value of σ_{t+s}^2

$$\mathrm{Var}_t(u_{t+s}) = \mathrm{E}_t\, \sigma_{t+s}^2. \tag{5.18}$$

Proof. (of (5.17)–(5.18)) 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 (5.17). Further periods follow the same pattern.

To prove (5.18), notice that $\mathrm{Var}_t(u_{t+s}) = \mathrm{E}_t\, v_{t+s}^2\sigma_{t+s}^2 = \mathrm{E}_t\, v_{t+s}^2\, \mathrm{E}_t\, \sigma_{t+s}^2$ since v_{t+s} and σ_{t+s} are independent. In addition, $\mathrm{E}_t\, v_{t+s}^2 = 1$, which proves (5.18). ■

If we assume that v_t is iid $N(0,1)$, then the distribution of u_{t+1}, conditional on the information in t, is $N(0, \sigma_{t+1}^2)$, where σ_{t+1} is known already in t. Therefore, the one-step ahead distribution is normal—which can be used for estimating the model with MLE. However, the distribution of u_{t+2} (still conditional on the information in t) is more complicated. Notice that

$$u_{t+2} = v_{t+2}\sigma_{t+2} = v_{t+2}\sqrt{\omega + \alpha v_{t+1}^2\sigma_{t+1}^2}, \tag{5.19}$$

which is a nonlinear function of v_{t+2} and v_{t+1}, both of which are standard normal. This makes u_{t+2} have a non-normal distribution. In fact, it will have fatter tails than a normal distribution with the same variance (excess kurtosis). This spills over to the unconditional distribution which has the following kurtosis

$$\frac{\mathrm{E}\, u_t^4}{(\mathrm{E}\, u_t^2)^2} = \begin{cases} 3\frac{1-\alpha^2}{1-3\alpha^2} > 3 & \text{if denominator is positive} \\ \infty & \text{otherwise.} \end{cases} \tag{5.20}$$

As a comparison, the kurtosis of a normal distribution is 3. This means that we can expected u_t to have fat tails, but that the standardized residuals u_t/σ_t perhaps look more normally distributed. See Figure 5.14 for an illustration (although based on a GARCH model).

Example 5.5 *(Kurtosis) With $\alpha = 1/3$, the kurtosis is 4, at $\alpha = 0.5$ it is 9 and at $\alpha = 0.6$ it is infinite.*

Proof. (of (5.20)) Since v_t and σ_t are independent, we have $\mathrm{E}(u_t^2) = \mathrm{E}(v_t^2\sigma_t^2) = \mathrm{E}\,\sigma_t^2$ and $\mathrm{E}(u_t^4) = \mathrm{E}(v_t^4\sigma_t^4) = \mathrm{E}(\sigma_t^4)\,\mathrm{E}(v_t^4) = \mathrm{E}(\sigma_t^4)3$, where the last equality follows from $\mathrm{E}(v_t^4) = 3$ for a standard normal variable. To find $\mathrm{E}(\sigma_t^4)$, square (5.16) and take

expectations (and use $\mathrm{E}\,\sigma_t^2 = \omega/(1-\alpha)$)

$$\begin{aligned}\mathrm{E}\,\sigma_t^4 &= \omega^2 + \alpha^2\,\mathrm{E}\,u_{t-1}^4 + 2\omega\alpha\,\mathrm{E}\,u_{t-1}^2 \\ &= \omega^2 + \alpha^2\,\mathrm{E}(\sigma_t^4)3 + 2\omega^2\alpha/(1-\alpha), \text{ so} \\ \mathrm{E}\,\sigma_t^4 &= \frac{1+\alpha}{1-3\alpha^2}\frac{\omega^2}{(1-\alpha)}.\end{aligned}$$

Multiplying by 3 and dividing by $(\mathrm{E}\,u_t^2)^2 = \omega^2/(1-\alpha)^2$ gives (5.20). ■

5.2.2 Estimation of the ARCH(1) Model

Suppose we want to estimate the ARCH model—perhaps because we are interested in the heteroskedasticity or because we want a more efficient estimator of the regression equation than LS. We therefore want to estimate the full model (5.14)–(5.16) by ML or GMM.

The most common way to estimate the model is to assume that v_t is 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

$$\ln \mathcal{L} = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln\sigma_t^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{u_t^2}{\sigma_t^2}, \text{ if} \quad (5.21)$$

v_t is iid $N(0,1)$.

By plugging in (5.14) for u_t and (5.16) for σ_t^2, the likelihood function is written in terms of the data and model parameters. The likelihood function is then maximized with respect to the parameters. Note that we need a starting value of $\sigma_1^2 = \omega + \alpha u_0^2$. The most convenient (and common) way is to maximize the likelihood function conditional on a y_0 and x_0. 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. The optimization should preferably impose the constraints in (5.16). The MLE is consistent.

Remark 5.6 *(Likelihood function of $x_t \sim N(\mu,\sigma^2)$) The pdf of an $x_t \sim N(\mu,\sigma^2)$ is*

$$pdf(x_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x_t-\mu)^2}{\sigma^2}\right),$$

so the log-likelihood is

$$\ln \mathcal{L}_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\frac{(x_t-\mu)^2}{\sigma^2}.$$

If x_t and x_s are independent (uncorrelated if normally distributed), then the joint pdf is the product of the marginal pdfs—and the joint log-likelihood is the sum of the two likelihoods.

Remark 5.7 *(Coding the ARCH(1) ML estimation) A straightforward way of coding the estimation problem (5.14)–(5.16) and (5.21) is as follows.*
First, guess values of the parameters b (a vector), and ω, and α. The guess of b can be taken from an LS estimation of (5.14), and the guess of ω and α from an LS estimation of $\hat{u}_t^2 = \omega + \alpha\hat{u}_{t-1}^2 + \varepsilon_t$ where $\hat{u}_t$ are the fitted residuals from the LS estimation of (5.14).
Second, loop over the sample (first $t = 1$, then $t = 2$, etc.) and calculate $\hat{u}_t$ from (5.14) and σ_t^2 from (5.16). Plug in these numbers in (5.21) 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 5.8 *(Imposing parameter constraints on ARCH(1)) To impose the restrictions in (5.16) when the previous remark is implemented, iterate over values of $(b, \tilde{\omega}, \tilde{\alpha})$ and let $\omega = \tilde{\omega}^2$ and $\alpha = \exp(\tilde{a})/[1+\exp(\tilde{a})]$.*

It is often found that the fitted normalized residuals, $\hat{u}_t/\sigma_t$, still have too fat tails compared with $N(0,1)$. 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).

Another possibility is to estimate the model by GMM using, for instance, the following moment conditions

$$\mathrm{E}\begin{bmatrix} x_t u_t \\ u_t^2 - \sigma_t^2 \\ u_{t-1}^2(u_t^2 - \sigma_t^2) \end{bmatrix} = \mathbf{0}_{(k+2)\times 1}, \tag{5.22}$$

where u_t and σ_t^2 are given by (5.14) and (5.16).

It is straightforward to add more lags to (5.16). For instance, an ARCH(p) would be

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \ldots + \alpha_p u_{t-p}^2. \tag{5.23}$$

We then have to add more moment conditions to (5.22), but 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 $\Sigma_{j=1}^{p} \alpha_j \leq 1$.

5.3 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

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2, \text{ with} \\ \omega &> 0; \alpha, \beta \geq 0; \text{ and } \alpha + \beta < 1, \end{aligned} \tag{5.24}$$

combined with (5.14) and (5.15).

See Figure 5.12 for an illustration.

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. To see the latter, notice that we in period t can forecast the future conditional variance (σ_{t+s}^2) as (since σ_{t+1}^2 is known in t)

$$\mathrm{E}_t\, \sigma_{t+s}^2 = \bar{\sigma}^2 + (\alpha + \beta)^{s-1} \left(\sigma_{t+1}^2 - \bar{\sigma}^2 \right), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1 - \alpha - \beta}, \tag{5.25}$$

where $\bar{\sigma}^2$ is the unconditional variance. This has the same form as in the ARCH(1) model (5.17), but where the sum of α and β is like an AR(1) parameter. The restriction $\alpha + \beta < 1$ must hold for this difference equation to be stable.

As for the ARCH model, the conditional variance of u_{t+s} is clearly equal to the expected value of σ_{t+s}^2

$$\mathrm{Var}_t(u_{t+s}) = \mathrm{E}_t\, \sigma_{t+s}^2. \tag{5.26}$$

Assuming that u_t has no autocorrelation, it follows directly from (5.25) that the ex-

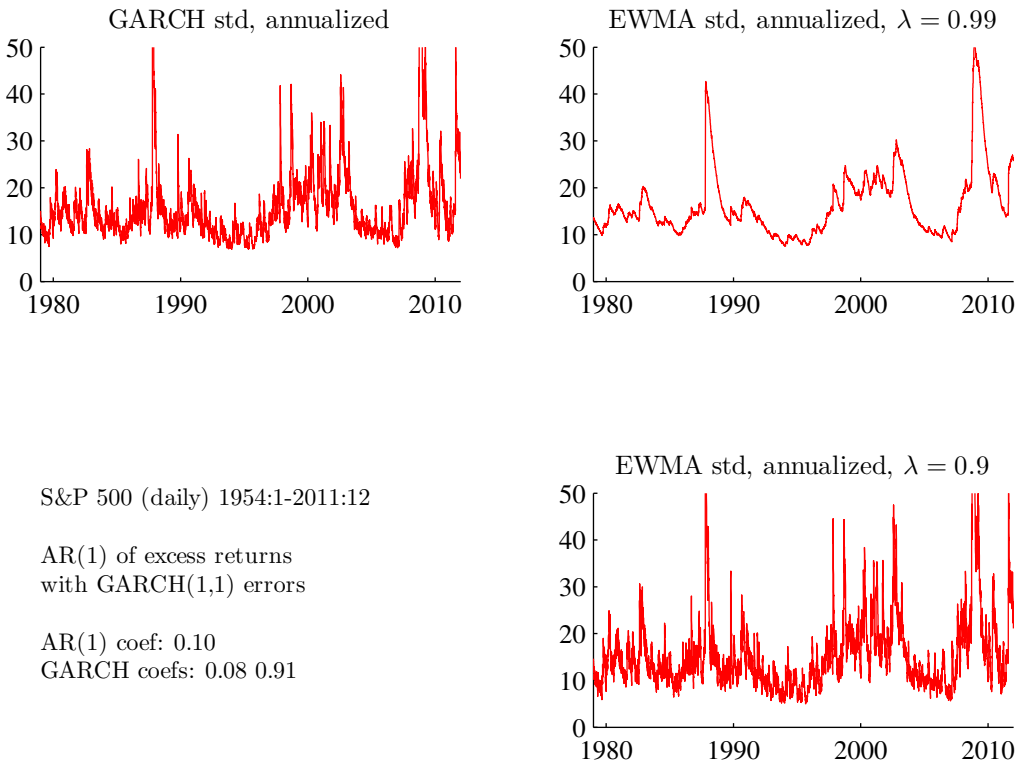


Figure 5.13: Conditional standard deviation, estimated by GARCH(1,1) model

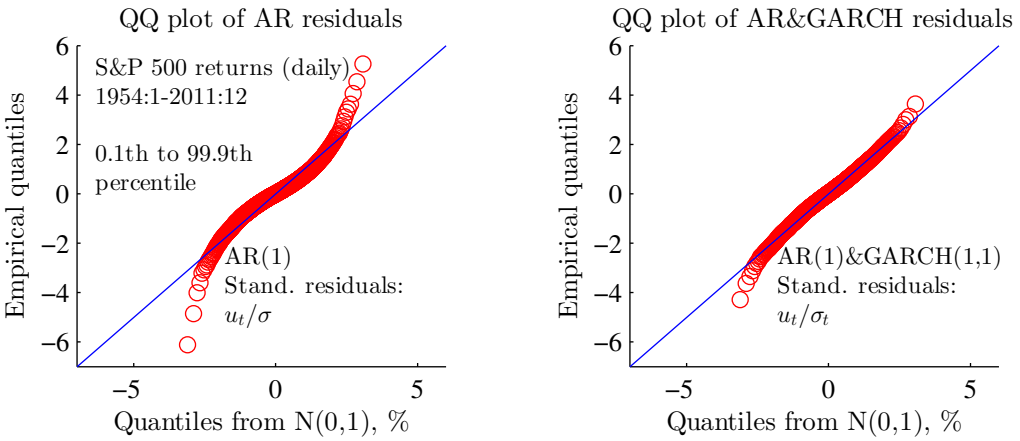
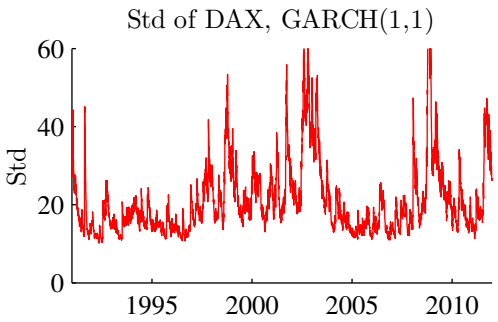


Figure 5.14: QQ-plot of residuals



Daily DAX returns 1991:1-2011:12

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	0.086	0.011
β_1	0.898	0.011

Figure 5.15: Results for a univariate GARCH model

pected variance of a longer time period ($u_{t+1} + u_{t+2} + \ldots + u_{t+K}$) is

$$\begin{aligned} \text{Var}_t(\textstyle\sum_{s=1}^{K} u_{t+s}) &= \text{E}_t \textstyle\sum_{s=1}^{K} \sigma_{t+s}^2 = K\bar{\sigma}^2 + \textstyle\sum_{s=1}^{K} (\alpha+\beta)^{s-1} \left(\sigma_{t+1}^2 - \bar{\sigma}^2\right) \\ &= K\bar{\sigma}^2 + \frac{1-(\alpha+\beta)^K}{1-(\alpha+\beta)} \left(\sigma_{t+1}^2 - \bar{\sigma}^2\right). \end{aligned} \tag{5.27}$$

This is useful for portfolio choice and asset pricing when the horizon is longer than one period (day, perhaps).

See Figures 5.13–5.14 for illustrations.

Proof. (of (5.25)–(5.27)) Notice that $\text{E}_t\, \sigma_{t+2}^2 = \omega + \alpha\, \text{E}_t\, v_{t+1}^2\, \text{E}_t\, \sigma_{t+1}^2 + \beta\sigma_{t+1}^2$ since v_t is independent of σ_t. Morover, $\text{E}_t\, v_{t+1}^2 = 1$ and $\text{E}_t\, \sigma_{t+1}^2 = \sigma_{t+1}^2$ (known in t). Combine to get $\text{E}_t\, \sigma_{t+2}^2 = \omega + (\alpha+\beta)\sigma_{t+1}^2$. Similarly, $\text{E}_t\, \sigma_{t+3}^2 = \omega + (\alpha+\beta)\, \text{E}_t\, \sigma_{t+2}^2$. Substitute for $\text{E}_t\, \sigma_{t+2}^2$ to get $\text{E}_t\, \sigma_{t+3}^2 = \omega + (\alpha+\beta)[\omega + (\alpha+\beta)\sigma_{t+1}^2]$, which can be written as (5.25). Further periods follow the same pattern.

To prove (5.27), use (5.25) and notice that $\sum_{s=1}^{K} (\alpha+\beta)^{s-1} = \left[1-(\alpha+\beta)^K\right] / \left[1-(\alpha+\beta)\right]$.

■

Remark 5.9 *(EWMA) The GARCH(1,1) has many similarities with the exponential moving average estimator of volatility*

$$\sigma_t^2 = (1-\lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2.$$

This methods is commonly used by practitioners. For instance, the RISK Metrics uses this method with $\lambda = 0.94$. *Clearly,* λ *plays the same type of role as* β *in (5.24) and* $1-\lambda$ *as* α. *The main differences are that the exponential moving average does not have*

a constant and volatility is non-stationary (the coefficients sum to unity). See Figure 5.13 for a comparison.

The kurtosis of the process is

$$\frac{\mathrm{E}\, u_t^4}{(\mathrm{E}\, u_t^2)^2} = \begin{cases} 3\frac{1-(\alpha+\beta)^2}{1-2\alpha^2-(\alpha+\beta)} > 3 & \text{if denominator is positive} \\ \infty & \text{otherwise.} \end{cases} \tag{5.28}$$

Proof. (of (5.28)) Since v_t and σ_t are independent, we have $\mathrm{E}(u_t^2) = \mathrm{E}(v_t^2\sigma_t^2) = \mathrm{E}\,\sigma_t^2$ and $\mathrm{E}(u_t^4) = \mathrm{E}(v_t^4\sigma_t^4) = \mathrm{E}(\sigma_t^4)\,\mathrm{E}(v_t^4) = \mathrm{E}(\sigma_t^4)3$, where the last equality follows from $\mathrm{E}(v_t^4) = 3$ for a standard normal variable. We also have $\mathrm{E}(u_t^2\sigma_t^2) = \mathrm{E}\,\sigma_t^4$

$$\begin{aligned} \mathrm{E}\,\sigma_t^4 &= \mathrm{E}(\omega + \alpha u_{t-1}^2 + \beta\sigma_{t-1}^2)^2 \\ &= \omega^2 + \alpha^2\,\mathrm{E}\,u_{t-1}^4 + \beta^2\,\mathrm{E}\,\sigma_{t-1}^4 + 2\omega\alpha\,\mathrm{E}\,u_{t-1}^2 + 2\omega\beta\,\mathrm{E}\,\sigma_{t-1}^2 + 2\alpha\beta\,\mathrm{E}(u_{t-1}^2\sigma_{t-1}^2) \\ &= \omega^2 + \alpha^2\,\mathrm{E}(\sigma_t^4)3 + \beta^2\,\mathrm{E}\,\sigma_t^4 + 2\omega\alpha\,\mathrm{E}\,\sigma_t^2 + 2\omega\beta\,\mathrm{E}\,\sigma_t^2 + 2\alpha\beta\,\mathrm{E}\,\sigma_t^4 \\ &= \frac{\omega^2 + 2\omega(\alpha+\beta)\,\mathrm{E}\,\sigma_t^2}{1 - 2\alpha^2 - (\alpha+\beta^2)^2}. \end{aligned}$$

Use $\mathrm{E}\,\sigma_t^2 = \omega/(1-a-\beta)$, multiply by 3 and divide by $(\mathrm{E}\,u_t^2)^2 = \omega^2/(1-\alpha-\beta)^2$ gives (5.28). ■

The GARCH(1,1) corresponds to an ARCH(∞) with geometrically declining weights, which is seen by solving (5.24) recursively by substituting for σ_{t-1}^2 (and then σ_{t-2}^2, σ_{t-3}^2, ...)

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha\sum_{j=0}^{\infty}\beta^j u_{t-1-j}^2. \tag{5.29}$$

This suggests that a GARCH(1,1) might be a reasonable approximation of a high-order ARCH.

Proof. (of (5.29)) Substitute for σ_{t-1}^2 in (5.24), and then for σ_{t-2}^2, etc

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha u_{t-1}^2 + \beta\overbrace{(\omega + \alpha u_{t-2}^2 + \beta\sigma_{t-2}^2)}^{\sigma_{t-1}^2} \\ &= \omega\,(1+\beta) + \alpha u_{t-1}^2 + \beta\alpha u_{t-2}^2 + \beta^2\sigma_{t-2}^2 \\ &= \vdots \end{aligned}$$

and we get (5.29). ■

To estimate the model consisting of (5.14), (5.15) and (5.24) we can still use the likelihood function (5.21) and do a MLE. We typically create the starting value of u_0^2 as in the ARCH model (use y_0 and x_0 to create u_0), but this time we also need a starting value of σ_0^2. It is often recommended that we use $\sigma_0^2 = \text{Var}(\hat{u}_t)$, where $\hat{u}_t$ are the residuals from a LS estimation of (5.14). It is also possible to assume another distribution than $N(0,1)$.

Remark 5.10 *(Imposing parameter constraints on GARCH(1,1)) To impose the restrictions in (5.24), iterate over values of* $(b, \tilde{\omega}, \tilde{\alpha}, \tilde{\beta})$ *and let* $\omega = \tilde{\omega}^2$, $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$, *and* $\beta = \exp(\tilde{\beta})/[1 + \exp(\tilde{\alpha}) + \exp(\tilde{\beta})]$.

To estimate the GARCH(1,1) with GMM, we can, for instance, use the following moment conditions (where σ_t^2 is given by (5.24))

$$\text{E}\begin{bmatrix} x_t u_t \\ u_t^2 - \sigma_t^2 \\ u_{t-1}^2(u_t^2 - \sigma_t^2) \\ u_{t-2}^2(u_t^2 - \sigma_t^2) \end{bmatrix} = \mathbf{0}_{(k+3)\times 1}, \text{ where } u_t = y_t - x_t' b. \tag{5.30}$$

Remark 5.11 *(Value at Risk) The value at risk (as fraction of the investment) at the* α *level (say,* $\alpha = 0.95$*) is* $VaR_\alpha = -\text{cdf}^{-1}(1-\alpha)$, *where* $\text{cdf}^{-1}()$ *is the inverse of the cdf—so* $\text{cdf}^{-1}(1-\alpha)$ *is the* $1-\alpha$ *quantile of the return distribution. See Figure 5.16 for an illustration. When the return has an* $N(\mu, \sigma^2)$ *distribution, then* $VaR_{95\%} = -(\mu - 1.64\sigma)$. *See Figures 5.17–5.19 for an example of time-varying VaR, based on a GARCH model.*

5.4 Non-Linear Extensions

A very large number of extensions of the basic GARCH model have been suggested. Estimation is straightforward since MLE is done as for any other GARCH model—just the specification of the variance equation differs.

An asymmetric GARCH (Glosten, Jagannathan, and Runkle (1993)) can be constructed as

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \delta(u_{t-1} > 0) u_{t-1}^2, \text{ where} \tag{5.31}$$

$$\delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases}$$

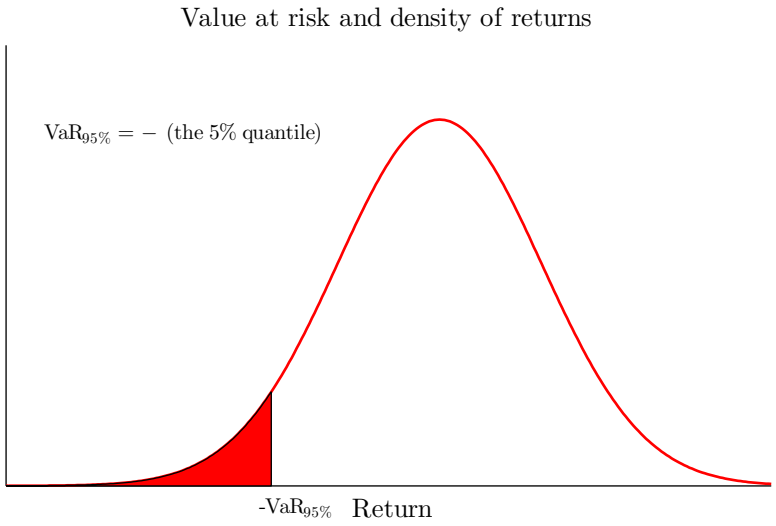
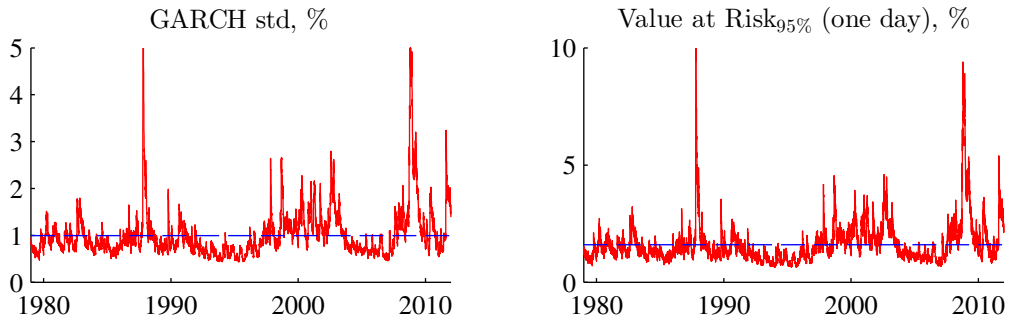


Figure 5.16: Value at risk



S&P 500, daily data 1954:1-2011:12

The horizontal lines are from the unconditional distribution

The VaR is based on N()

Figure 5.17: Conditional volatility and VaR

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

$$\ln \sigma_t^2 = \omega + \alpha \frac{|u_{t-1}|}{\sigma_{t-1}} + \beta \ln \sigma_{t-1}^2 + \gamma \frac{u_{t-1}}{\sigma_{t-1}}. \tag{5.32}$$

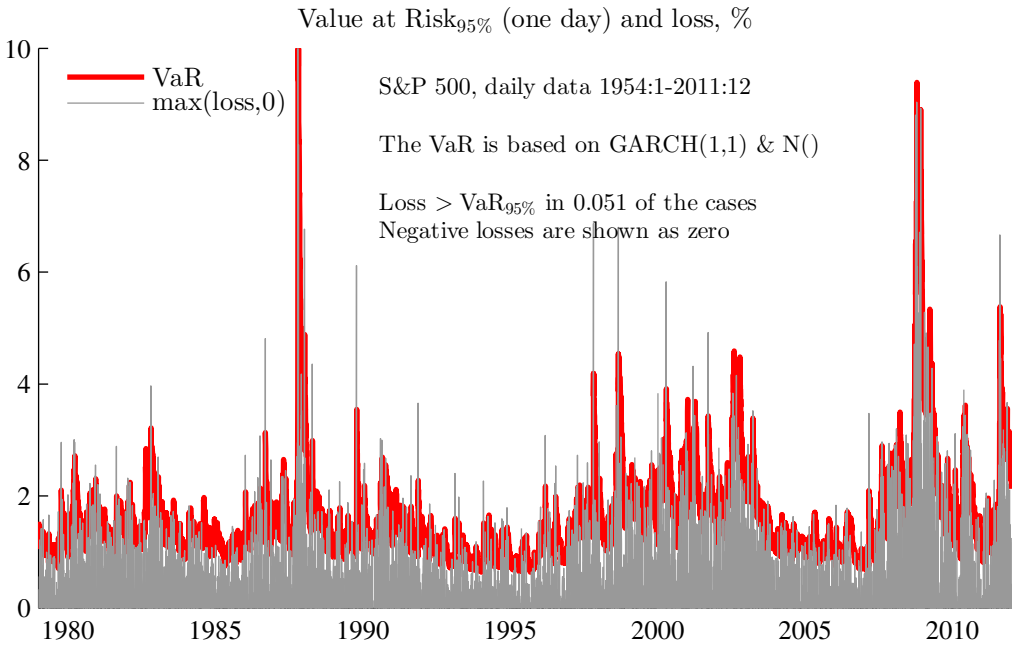


Figure 5.18: Backtesting VaR from a GARCH model, assuming normally distributed shocks

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 $|u_{t-1} - b|^{3/2}$.

5.5 GARCH Models with Exogenous Variables

We could easily extend the GARCH(1,1) model by adding exogenous variables x_{t-1}, for instance, VIX

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma x_{t-1}, \tag{5.33}$$

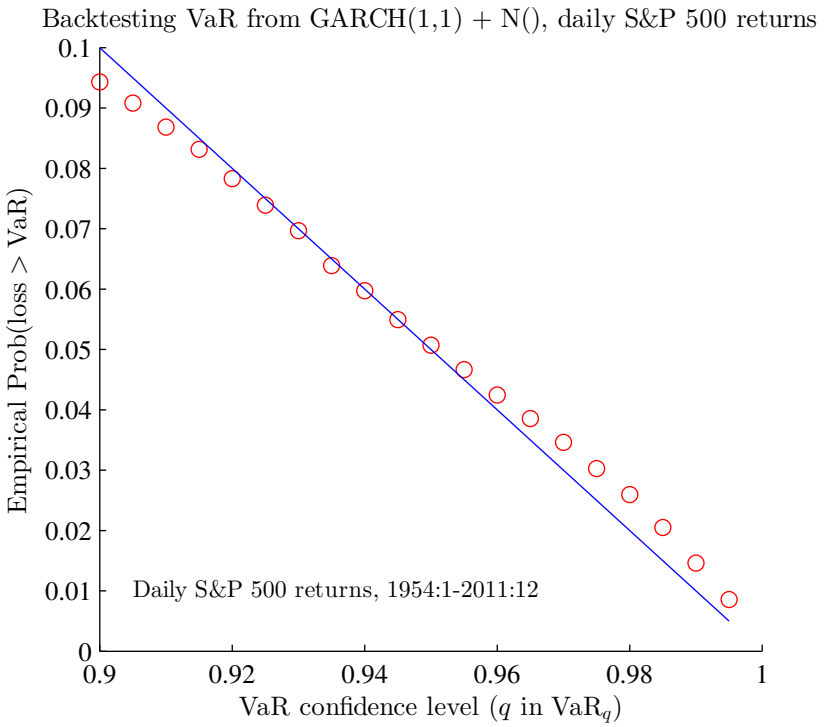


Figure 5.19: Backtesting VaR from a GARCH model, assuming normally distributed shocks

where care must be taken to guarantee that $\sigma_t^2 > 0$. One possibility is to make sure that $x_t > 0$ and then restrict γ to be non-negative. Alternatively, we could use an EGARCH formulation like

$$\ln \sigma_t^2 = \omega + \alpha \frac{|u_{t-1}|}{\sigma_{t-1}} + \beta \ln \sigma_{t-1}^2 + \gamma x_{t-1}. \tag{5.34}$$

These models can be estimated with maximum likelihood.

5.6 Stochastic Volatility Models

A *stochastic volatility model* differs from GARCH models by making the volatility truly stochastic. Recall that in a GARCH model, the volatility in period t (σ_t) is know already in $t-1$. This is not the case in a stochastic volatility model where the log volatility follows

an ARMA process. The simplest case is the AR(1) formulation

$$\begin{aligned} \ln \sigma_t^2 &= \omega + \beta \ln \sigma_{t-1}^2 + \theta \eta_t, \\ \text{with } \eta_t &\sim iidN(0,1), \end{aligned} \tag{5.35}$$

combined with (5.14) and (5.15).

The estimation of a stochastic volatility model is complicated—and the basic reason is that it is very difficult to construct the likelihood function. So far, the most practical way to do MLE is by simulations.

Instead, stochastic volatility models are often estimated by quasi-MLE. For the model (5.15) and (5.35), this could be done as follows: square (5.15) and take logs to get

$$\ln u_t^2 = \mathrm{E} \ln v_t^2 + \ln \sigma_t^2 + (\ln v_t^2 - \mathrm{E} \ln v_t^2). \tag{5.36}$$

We could use this as the measurement equation in a Kalman filter (pretending that $\ln v_t^2 - \mathrm{E} \ln v_t^2$ is normally distributed), and (5.35) as the state equation. (The Kalman filter is a convenient way to calculate the likelihood function.) In essence, this is an AR(1) model with "noisy observations."

If $\ln v_t^2$ is normally distributed , then this will give MLE, otherwise just a quasi-MLE. For instance, if v_t is $iidN(0,1)$ (see Ruiz (1994)) then we have approximately $\mathrm{E} \ln v_t^2 \approx -1.27$ and $\mathrm{Var}(\ln v_t^2) = \pi^2/2$ (with $\pi = 3.14...$) so we could write the measurement equation as

$$\begin{aligned} \ln u_t^2 &= -1.27 + \ln \sigma_t^2 + w_t, \text{ with} \\ w_t &\sim N(0, \pi^2/2). \end{aligned} \tag{5.37}$$

In this case, only the state equation contains parameters that we need to estimate: ω, β, θ.

See Figure 5.20 for an example.

5.7 (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.

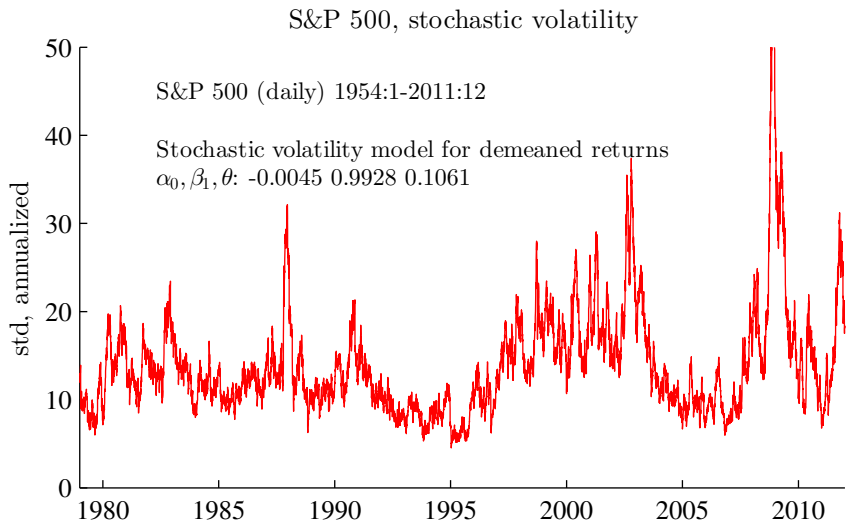


Figure 5.20: Conditional standard deviation, stochastic volatility model

Example 5.12 *(Mean-variance portfolio choice) A mean variance investor solves*

$$\max_\alpha \mathrm{E}\, R_p - \sigma_p^2 k/2,$$
$$\text{subject to } R_p = \alpha R_m + (1-\alpha) R_f,$$

where R_m is the return on the risky asset (the market index) and R_f is the riskfree return. The solution is

$$\alpha = \frac{1}{k}\frac{\mathrm{E}(R_m - R_f)}{\sigma_m^2}.$$

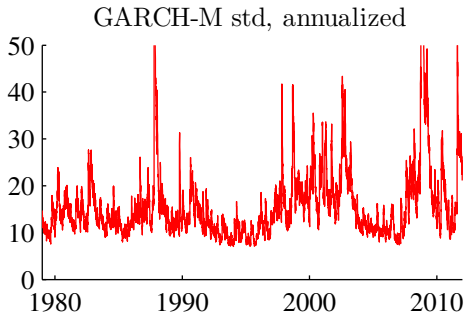
In equilibrium, this weight is one (since the net supply of bonds is zero), so we get

$$\mathrm{E}(R_m - R_f) = k\sigma_m^2,$$

which says that the expected excess return is increasing in both the market volatility and risk aversion (k).

We modify the "mean equation" (5.14) to include the conditional variance σ_t^2 or the standard deviation σ_t (taken from any of the models for heteroskedasticity) as a regressor

$$y_t = x_t'b + \varphi\sigma_t^2 + u_t, \ \mathrm{E}(u_t|x_t, \sigma_t) = 0. \tag{5.38}$$



S&P 500 (daily) 1954:1-2011:12

AR(1) + GARCH-M of excess returns
with GARCH(1,1) errors

AR(1) coef and coef on σ_t: 0.10 0.07
GARCH coefs: 0.08 0.91

Figure 5.21: GARCH-M example

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 (5.21) to do MLE.

It can also be noted (see Gourieroux and Jasiak (2001) 11.3) that a slightly modified GARCH-M model is the discrete time sampling version of a continuous time stochastic volatility model (where the mean is affected by one Wiener process and the variance by another).

See Figure 5.21 for an example.

Remark 5.13 *(Coding of (G)ARCH-M) We can use the same approach as in Remark 5.7, except that we use (5.38) instead of (5.14) to calculate the residuals (and that we obviously also need a guess of φ).*

5.8 Multivariate (G)ARCH

5.8.1 Different Multivariate Models

This section gives a brief summary of some multivariate models of heteroskedasticity. Let the model (5.14) be a multivariate model where y_t and u_t are $n \times 1$ vectors. We define the conditional (on the information set in $t-1$) covariance matrix of u_t as

$$\Sigma_t = \mathrm{E}_{t-1}\, u_t u_t'. \tag{5.39}$$

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

$$\text{vech}(\Sigma_t) = C + A\text{vech}(u_{t-1}u'_{t-1}) + B\text{vech}(\Sigma_{t-1}). \tag{5.40}$$

This typically gives too many parameters to handle—and makes it difficult to impose sufficient restrictions to make Σ_t is positive definite (compare the restrictions of positive coefficients in (5.24)).

Example 5.14 *(vech formulation, $n = 2$) For instance, with $n = 2$ we have*

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = C + A \begin{bmatrix} u^2_{1,t-1} \\ u_{1,t-1}u_{2,t-1} \\ u^2_{2,t-1} \end{bmatrix} + B \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix},$$

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.

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. 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.

Example 5.15 *(Diagonal model, $n = 2$) With $n = 2$ we have*

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} u^2_{1,t-1} \\ u_{1,t-1}u_{2,t-1} \\ u^2_{2,t-1} \end{bmatrix} + \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix},$$

which gives $3 + 3 + 3 = 9$ parameters (in C, A, and B, respectively).

The BEKK Model

The *BEKK model* makes Σ_t positive definite by specifying a quadratic form

$$\Sigma_t = C + A'u_{t-1}u'_{t-1}A + B'\Sigma_{t-1}B, \tag{5.41}$$

where C is symmetric and A and B are $n \times n$ matrices. Notice that this equation is specified in terms of Σ_t, not vech(Σ_t). Recall that a quadratic form positive definite, provided the matrices are of full rank.

Example 5.16 *(BEKK model, $n = 2$) With $n = 2$ we have*

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}' \begin{bmatrix} u_{1,t-1}^2 & u_{1,t-1}u_{2,t-1} \\ u_{1,t-1}u_{2,t-1} & u_{2,t-1}^2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}' \begin{bmatrix} \sigma_{11,t-1} & \sigma_{12,t-1} \\ \sigma_{12,t-1} & \sigma_{22,t-1} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

which gives $3 + 4 + 4 = 11$ parameters (in C, A, and B, respectively).

The Constant Correlation Model

The *constant correlation model* assumes that every variance follows a univariate GARCH process and that the conditional correlations are constant. 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. The price is, of course, the assumption of no movements in the correlations.

Example 5.17 *(Constant correlation model, $n = 2$) With $n = 2$ the covariance matrix is*

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix}$$

and each of σ_{11t} and σ_{22t} follows a GARCH process. Assuming a GARCH(1,1) as in (5.24) gives 7 parameters (2×3 GARCH parameters and one correlation), which is convenient.

Remark 5.18 *(Imposing parameter constraints on a correlation) To impose the restriction that $-1 < \rho < 1$, iterate over $\tilde{\rho}$ and let $\rho = 1 - 2/[1 + \exp(\tilde{\rho})]$.*

Remark 5.19 *(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 $u_{2t}/\sqrt{\sigma_{22,t}}$.*

The Dynamic Correlation Model

The *dynamic correlation model* (see Engle (2002) and Engle and Sheppard (2001)) allows the correlation to change over time. In short, the model assumes that each conditional variance follows a univariate GARCH process and the conditional correlation matrix is (essentially) allowed to follow a univariate GARCH equation.

The conditional covariance matrix is (by definition)

$$\Sigma_t = D_t R_t D_t, \text{ with } D_t = \text{diag}(\sqrt{\sigma_{ii,t}}), \tag{5.42}$$

and R_t is the conditional correlation matrix (discussed below).

Remark 5.20 *(diag(a_i) notation) diag(a_i) denotes the $n \times n$ matrix with elements $a_1, a_2, \ldots a_n$ along the main diagonal and zeros elsewhere. For instance, if $n = 2$, then*

$$diag(a_i) = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}.$$

The conditional correlation matrix R_t is allowed to change like in a univariate GARCH model, but with a transformation that guarantees that it is actually a valid correlation matrix. First, let v_t be the vector of standardized residuals and let $\bar{Q}$ be the unconditional correlation matrix of v_t. For instance, if assume a GARCH(1,1) structure for the correlation matrix, then we have

$$Q_t = (1 - \alpha - \beta)\bar{Q} + \alpha v_{t-1} v_{t-1}' + \beta Q_{t-1}, \text{ with } v_{i,t} = u_{i,t}/\sqrt{\sigma_{ii,t}}, \tag{5.43}$$

where α and β are two *scalars* and $\bar{Q}$ is the unconditional covariance matrix of the normalized residuals (v_t). To guarantee that the conditional correlation matrix is indeed a correlation matrix, Q_t is treated as if it where a covariance matrix and R_t is simply the implied correlation matrix. That is,

$$R_t = \text{diag}\left(\sqrt{q_{ii,t}}\right)^{-1} Q_t \text{diag}\left(\sqrt{q_{ii,t}}\right)^{-1}. \tag{5.44}$$

The basic idea of this model is to estimate a conditional correlation matrix as in (5.44) and then scale up with conditional variances (from univariate GARCH models) to get a conditional covariance matrix as in (5.42).

See Figures 5.22–5.23 for illustrations—which also suggest that the correlation is

close to what an EWMA method delivers. The DCC model is used in a study of asset pricing in, for instance, Duffee (2005).

Example 5.21 *(Dynamic correlation model, $n = 2$) With $n = 2$ the covariance matrix Σ_t is*

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12,t} \\ \rho_{12,t} & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_{11,t}} & 0 \\ 0 & \sqrt{\sigma_{22,t}} \end{bmatrix},$$

and each of σ_{11t} and σ_{22t} follows a GARCH process. To estimate the dynamic correlations, we first calculate (where α and β are two scalars)

$$\begin{bmatrix} q_{11,t} & q_{12,t} \\ q_{12,t} & q_{22,t} \end{bmatrix} = (1-\alpha-\beta) \begin{bmatrix} 1 & \bar{q}_{12} \\ \bar{q}_{12} & 1 \end{bmatrix} + \alpha \begin{bmatrix} v_{1,t-1} \\ v_{2,t-1} \end{bmatrix} \begin{bmatrix} v_{1,t-1} \\ v_{2,t-1} \end{bmatrix}' + \beta \begin{bmatrix} q_{11,t-1} & q_{12,t-1} \\ q_{12,t-1} & q_{22,t-1} \end{bmatrix},$$

where $v_{i,t-1} = u_{i,t-1}/\sqrt{\sigma_{ii,t-1}}$ and $\bar{q}_{ij}$ is the unconditional correlation of $v_{i,t}$ and $v_{j,t}$ and we get the conditional correlations by

$$\begin{bmatrix} 1 & \rho_{12,t} \\ \rho_{12,t} & 1 \end{bmatrix} = \begin{bmatrix} 1 & q_{12,t}/\sqrt{q_{11,t}q_{22,t}} \\ q_{12,t}/\sqrt{q_{11,t}q_{22,t}} & 1 \end{bmatrix}.$$

Assuming a GARCH(1,1) as in (5.24) gives 9 parameters (2×3 GARCH parameters, $(\bar{q}_{12}, \alpha, \beta)$).

To see what DCC generates, consider the correlation coefficient from a bivariate model

$$\begin{aligned} \rho_{12,t} &= \frac{q_{12,t}}{\sqrt{q_{11,t}}\sqrt{q_{22,t}}}, \text{ where} \\ q_{12,t} &= (1-\alpha-\beta)\bar{q}_{12} + \alpha v_{1,t-1}v_{2,t-1} + \beta q_{12,t-1} \\ q_{11,t} &= (1-\alpha-\beta) + \alpha v_{1,t-1}v_{1,t-1} + \beta q_{11,t-1} \\ q_{22,t} &= (1-\alpha-\beta) + \alpha v_{2,t-1}v_{2,t-1} + \beta q_{22,t-1}. \end{aligned} \tag{5.45}$$

This is a complicated expression, but the the numerator is the main driver: $q_{11,t}$ and $q_{22,t}$ are variances of normalized variables—so they should not be too far from unity. Therefore, $q_{12,t}$ is close to being the correlation itself. The equation for $q_{12,t}$ shows that it has a GARCH structure: it depends on $v_{1,t-1}v_{2,t-1}$ and $q_{12,t-1}$. Provided α and β are large numbers, we can expect the correlation to be strongly autocorrelated.

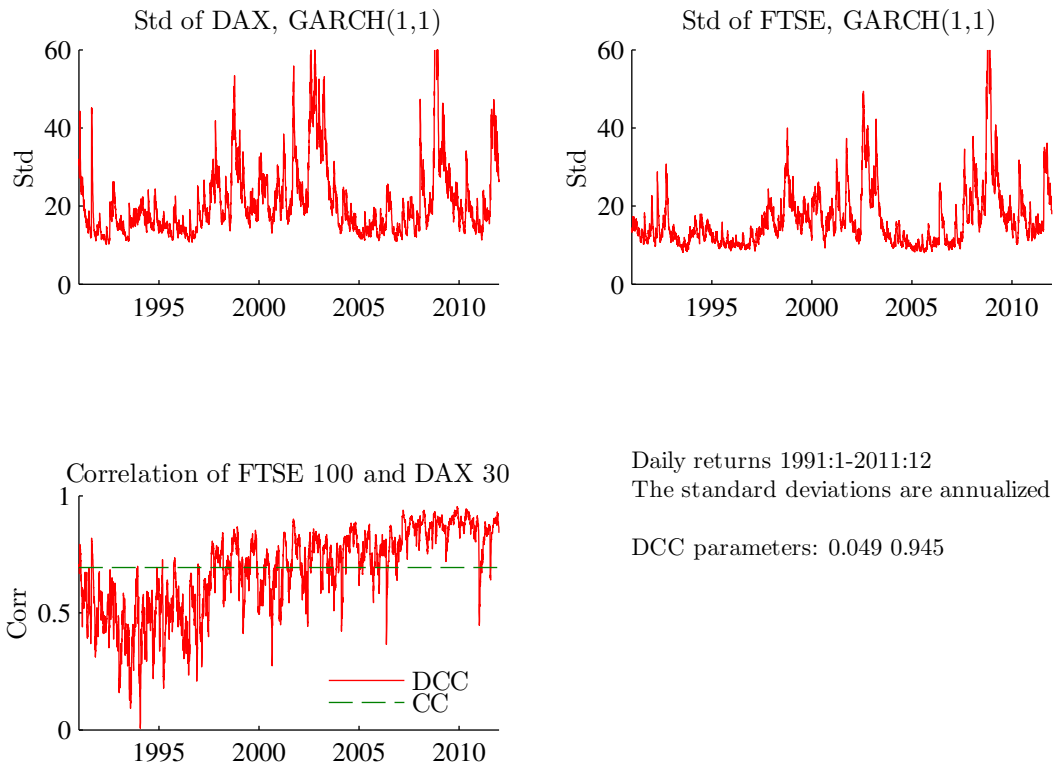


Figure 5.22: Results for multivariate GARCH models

5.8.2 Estimation of a Multivariate Model

In principle, it is straightforward to specify the likelihood function of the model and then maximize it with respect to the model parameters. For instance, if u_t is iid $N(0, \Sigma_t)$, then the log likelihood function is

$$\ln \mathcal{L} = -\frac{Tn}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln|\Sigma_t| - \frac{1}{2}\sum_{t=1}^{T}u_t'\Sigma_t^{-1}u_t. \tag{5.46}$$

In practice, the optimization problem can be difficult since there are typically many parameters. At least, good starting values are required.

Remark 5.22 *(Starting values of a constant correlation GARCH(1,1) model) Estimate GARCH(1,1) models for each variable separately, then estimate the correlation matrix on the standardized residuals.*

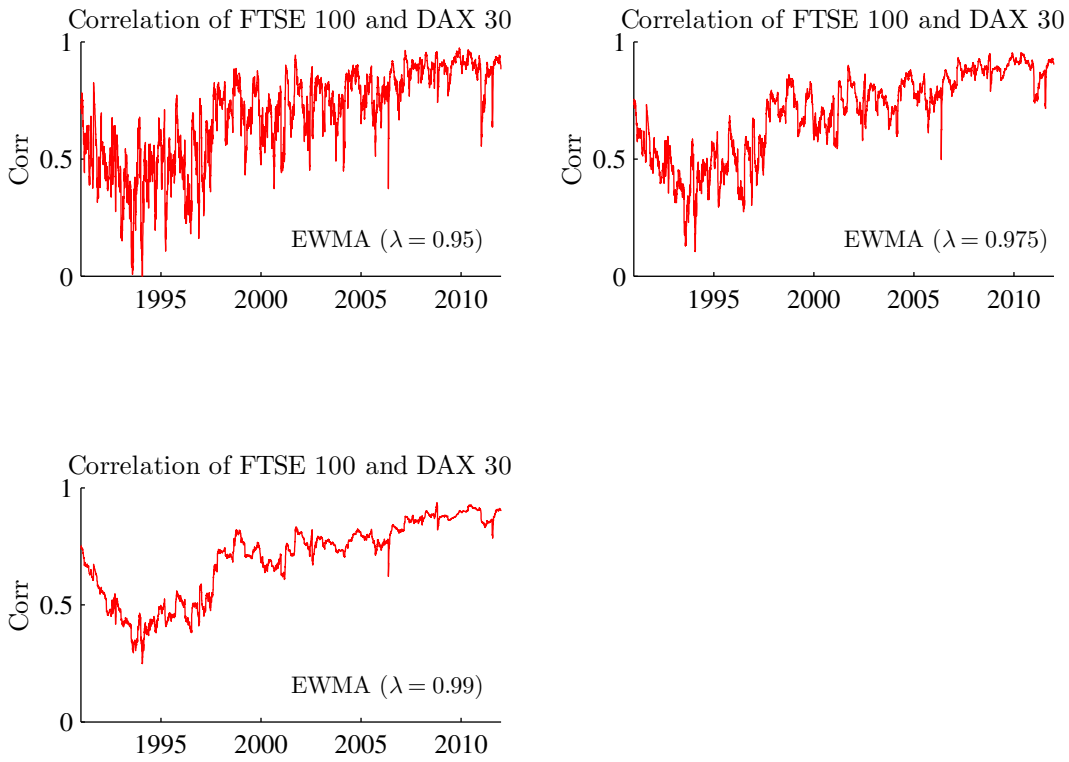


Figure 5.23: Time-varying correlations (different EWMA estimates)

Remark 5.23 *(Estimation of the dynamic correlation model) Engle and Sheppard (2001) suggest estimating the dynamic correlation matrix by two-step procedure. First, estimate the univariate GARCH processes. Second, use the standardized residuals to estimate the dynamic correlations by maximizing the likelihood function (5.46 if we assume normally distributed errors) with respect to the parameters α and β. In this second stage, both the parameters for the univariate GARCH process and the unconditional covariance matrix $\bar{Q}$ are kept constant.*

5.9 "A Closed-Form GARCH Option Valuation Model" by Heston and Nandi

References: Heston and Nandi (2000) (HN); Duan (1995)

This paper derives an option price formula for an asset that follows a GARCH process. This is applied to S&P 500 index options, and it is found that the model works well

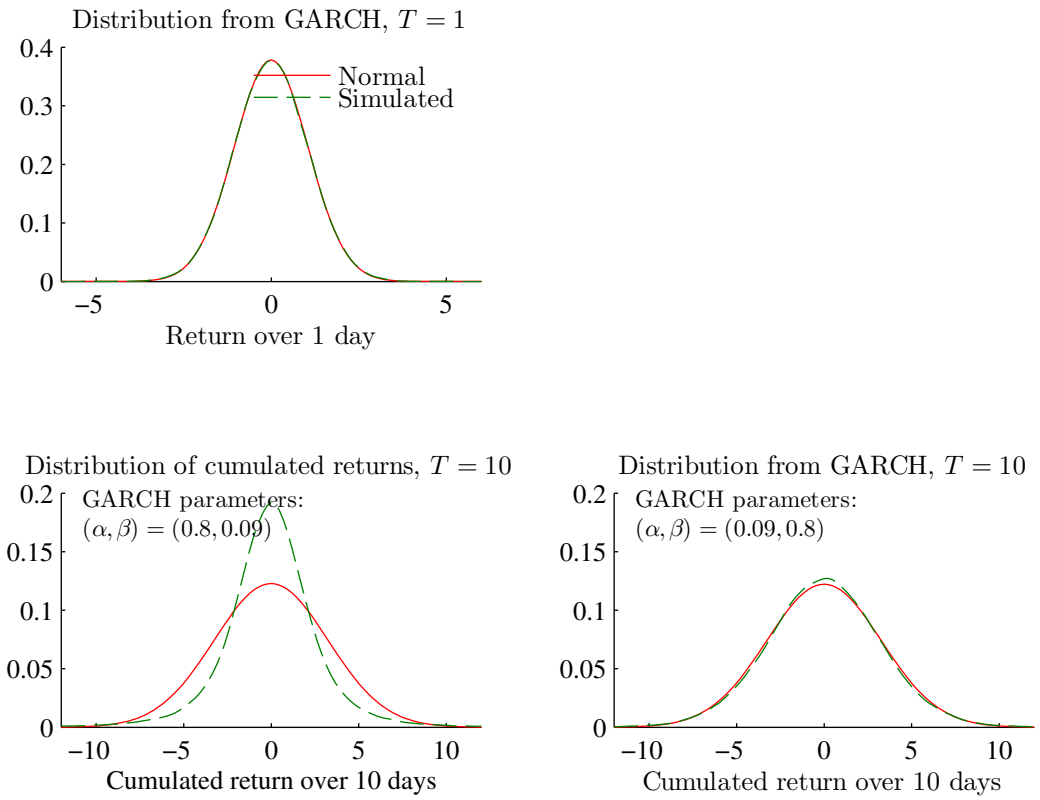


Figure 5.24: Comparison of normal and simulated distribution of m-period returns

compared to a Black-Scholes formula.

5.9.1 Background: GARCH vs Normality

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 5.24 for a comparison of the actual distribution of the log asset price at different horizons when the returns are generated by a GARCH model—and a normal distribution with the same mean and variance. It is clear that the normal distribution is a good approximation unless the horizon is short and the ARCH component ($\alpha_1 u_{t-1}^2$) dominates the GARCH component ($\beta_1 \sigma_{t-1}^2$).

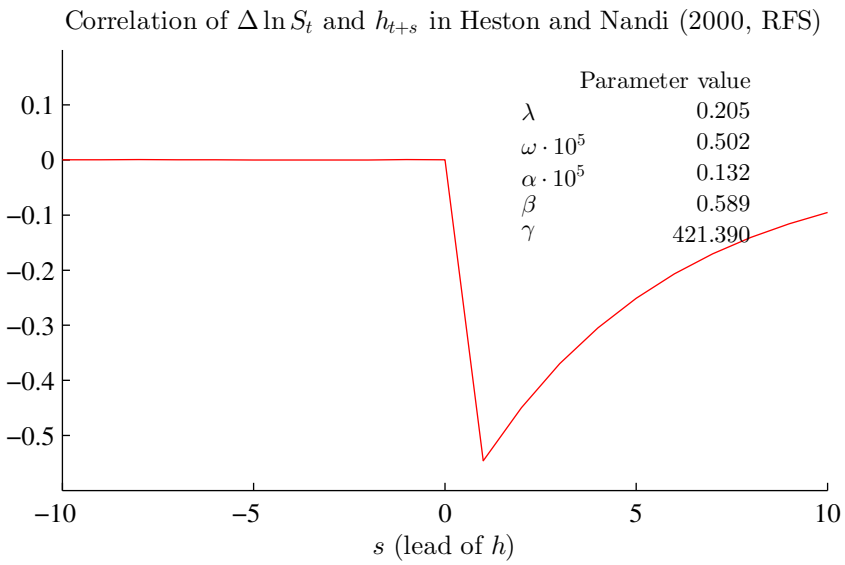


Figure 5.25: Simulated correlations of $\Delta \ln S_t$ and h_{t+s}

5.9.2 Option Price Formula: Part 1

Over the period from t to $t+\Delta$ the change of log asset price minus a riskfree rate (including dividends/accumulated interest), that is, the continuously compounded excess return, follows a kind of GARCH(1,1)-M process

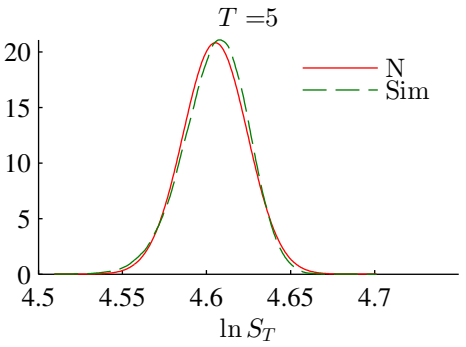
$$\ln S_t - \ln S_{t-\Delta} - r = \lambda h_t + \sqrt{h_t} z_t, \text{ where } z_t \text{ is iid } N(0,1) \tag{5.47}$$

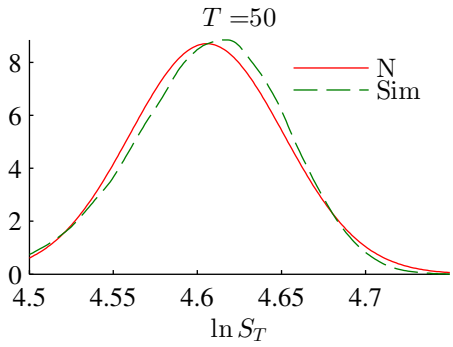
$$h_t = \omega + \alpha_1 (z_{t-\Delta} - \gamma_1 \sqrt{h_{t-\Delta}})^2 + \beta_1 h_{t-\Delta}. \tag{5.48}$$

The conditional variance would be a standard GARCH(1,1) process if $\gamma_1 = 0$. The additional term makes the response of h_t to an innovation symmetric around $\gamma_i \sqrt{h_{t-i\Delta}}$ instead of around zero. (HN also treat the case when the process is of higher order.)

If $\gamma_1 > 0$ then the return, $\ln S_t - \ln S_{t-\Delta}$, is negatively correlated with subsequent volatility $h_{t+\Delta}$—as often observed in data. To see this, note that the effect on the return of z_t is linear, but that a negative z_t drives up the conditional variance $h_{t+\Delta} = \omega + \alpha_1 (z_t - \gamma_1 \sqrt{h_t})^2 + \beta_1 h_t$ more than a positive z_t (if $\gamma_1 > 0$). The effect on the correlations is illustrated in Figure 5.25.

The process (5.47)–(5.48) does of course mean that the conditional (as of $t - \Delta$) distribution of the log asset price $\ln S_t$ is normally distributed. This is not enough to price





Heston-Nandi model, $\ln(S_0) = \ln(100) \approx 4.605$

Parameter	value
λ	0.205
$\omega \cdot 10^5$	0.502
$\alpha \cdot 10^5$	0.132
β	0.589
γ	421.390

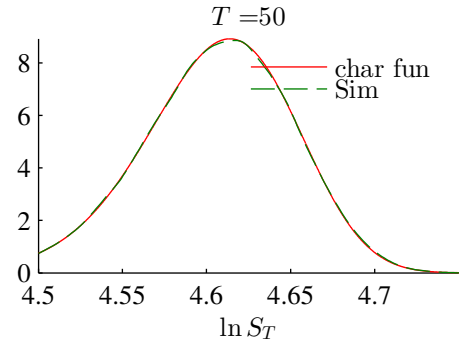


Figure 5.26: Distribution (physical) of $\ln S_T$ in the Heston-Nandi model

options on this asset, since we cannot use a dynamic hedging approach to establish a no-arbitrage price since there are (by the very nature of the discrete model) jumps in the price of the underlying asset. Recall that the price on a call option with strike price K is

$$C_{t-\Delta} = \mathrm{E}_{t-\Delta}\left\{M_t \max\left[S_t - K, 0\right]\right\}. \tag{5.49}$$

Alternatively, we can write

$$C_{t-\Delta} = e^{-r\Delta}\,\mathrm{E}^*_{t-\Delta}\left\{\max\left[S_t - K, 0\right]\right\}, \tag{5.50}$$

where $\mathrm{E}^*_{t-\Delta}$ is the expectations operator for the risk neutral distribution. See, for instance, Huang and Litzenberger (1988).

For parameter estimates on a more recent sample, see Table 5.4. These estimates suggests that λ has the wrong sign (high volatility predicts low future returns) and the

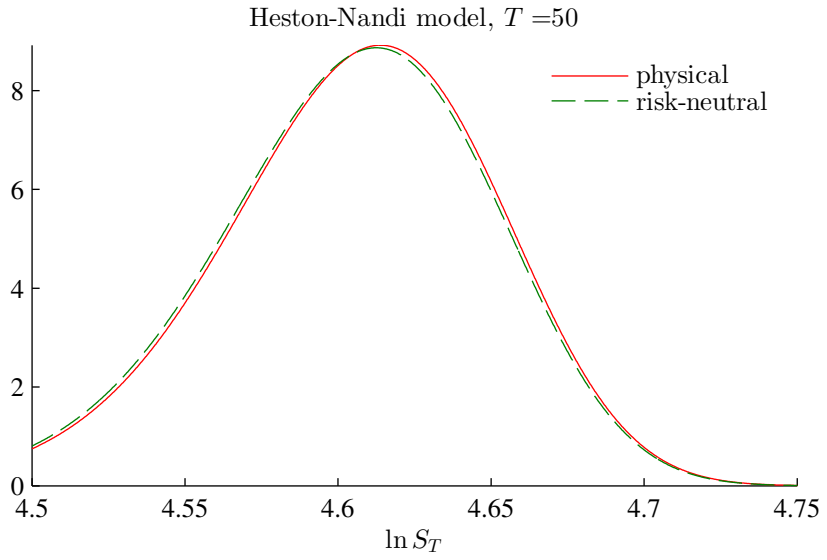


Figure 5.27: Physical and riskneutral distribution of lnS_T in the Heston-Nandi model

persistence of volatility is much higher than in HN (β is much higher).

λ	-2.5
ω	1.22e-006
α	0.00259
β	0.903
γ	6.06

Table 5.4: Estimate of the Heston-Nandi model on daily S&P500 excess returns, in %. Sample: 1990:1-2011:5

5.9.3 Option Price Formula: Part 2

HN assume that the risk neutral distribution of $\ln S_t$ (conditional on the information in $t-\Delta$) is normal, that is

Assumption: the price in $t-\Delta$ of a call option expiring in t follows BS.

This is the same as assuming that $\ln S_t$ and $\ln M_t$ have a bivariate normal distribution (conditional on the information in $t-\Delta$)—since this is what it takes to motivates the BS

model. This type of assumption was first used in a GARCH model by Duan (1995), who effectively assumed that $\ln M_t$ was iid normally distributed (this assumption is probably implicit in HN).

HN show that the risk neutral process must then be as in (5.47)–(5.48), but with γ_1 replaced by $\gamma_1^* = \gamma_1 + \lambda + 1/2$ and λ replaced by $-1/2$ (not in γ_1^*, of course). This means that they use the assumption about the conditional (as of $t - \Delta$) distribution of S_t to build up a conditional (as of $t - \Delta$) risk neutral distribution of S_T for any $T > t$. This risk neutral distribution can be calculated by clever tricks (as in HN) or by Monte Carlo simulations.

Once we have a risk neutral process it is (in principle, at least) straightforward to derive any option price (for any time to expiry). For a European call option with strike price K and expiry at date T, the result is

$$C_t(S_t, r, K, T) = e^{-r\Delta} \mathrm{E}_t^* \max[S_T - K, 0] \tag{5.51}$$

$$= S_t P_1 - e^{-r\Delta} K P_2, \tag{5.52}$$

where P_1 and P_2 are two risk neutral probabilities (implied by the risk neutral version of (5.47)–(5.48), see above). It can be shown that P_2 is the risk neutral probability that $S_T > K$, and that P_1 is the delta, $\partial C_t(S_t, r, K, T)/\partial S_t$ (just like in the Black-Scholes model). In practice, HN calculate these probabilities by first finding the risk neutral characteristic function of S_T, $f(\phi) = \mathrm{E}_t^* \exp(i\phi \ln S_T)$, where $i^2 = -1$, and then inverting to get the probabilities.

Remark 5.24 *(Characteristic function and the pdf) The characteristic function of a random variable x is*

$$f(\phi) = \mathrm{E} \exp(i\phi x)$$

$$= \int_x \exp(i\phi x) pdf(x) dx,$$

where pdf(x) is the pdf. This is a Fourier transform of the pdf (if x is a continuous random variable). For instance, the cf of a $N(\mu, \sigma^2)$ distribution is $\exp(i\phi\mu - \phi^2\sigma^2/2)$. *The pdf can therefore be recovered by the inverse Fourier transform as*

$$\mathrm{pdf}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-i\phi x) f(\phi) d\phi.$$

In practice, we typically use a fast (discrete) Fourier transform to perform this calculation, since there are very quick computer algorithms for doing that (see the appendix).

Remark 5.25 *(Characteristic function of* $\ln S_T$ *in the HN model) First, define*

$$A_t = A_{t+1} + i\phi r + B_{t+1}\omega - \frac{1}{2}\ln(1 - 2\alpha_1 B_{t+1})$$
$$B_t = i\phi(\lambda + \gamma_1) - \frac{1}{2}\gamma_1^2 + \beta_1 B_{t+1} + \frac{1}{2}\frac{(i\phi - \gamma_1)^2}{1 - \alpha_1 B_{t+1}},$$

which can be calculated recursively backwards ((A_T, B_T*), then (*A_{T-1}, B_{T-1}*), and so forth until (*A_0, B_0*)) starting from* $A_T = 0$ *and* $B_T = 0$*, where* T *is the investment horizon (time to expiration of the option contract). Notice that* i *is the imaginary number such that* $i^2 = -1$*. Second, the characteristics function for the horizon* T *is*

$$f(\phi) = S_0^{i\phi} \exp(A_0 + B_0 h_1).$$

Clearly, A_0 *and* B_0 *need to be recalculated for each value of* ϕ*.*

Remark 5.26 *(Characteristic function in the iid case) In the special case when* α_1*,* γ_1 *and* β_1 *are all zero, then process (5.47)–(5.48) has constant variance. Then, the recursions give*

$$A_0 = Ti\phi r + (T-1)\omega\left(i\phi\lambda - \frac{1}{2}\phi^2\right)$$
$$B_0 = i\phi\lambda - \frac{1}{2}\phi^2.$$

We can then write the characteristic function as

$$\begin{aligned} f(\phi) &= \exp(i\phi \ln S_0 + A_0 + B_0\omega) \\ &= \exp\left[i\phi\left[\ln S_0 + T(r + \omega\lambda)\right] - \phi^2 T\omega/2\right], \end{aligned}$$

which is the characteristic function of a normally distributed variable with mean $\ln S_0 + T(r + \omega\lambda)$ *and variance* $T\omega$*.*

5.9.4 Application to S&P 500 Index Option

Returns on the index are calculated by using official index plus dividends. The riskfree rate is taken to be a synthetic T-bill rate created by interpolating different bills to match

the maturity of the option. Weekly data for 1992–1994 are used (created by using lots of intraday quotes for all Wednesdays).

HN estimate the "GARCH(1,1)-M" process (5.47)–(5.48) with ML on daily data on the S&P500 index returns. It is found that the β_i parameter is large, α_i is small, and that $\gamma_1 > 0$ (as expected). The latter seems to be important for the estimated h_t series (see Figures 1 and 2).

Instead of using the "GARCH(1,1)-M" process estimated from the S&P500 index returns, all the model parameters are subsequently estimated from option prices. Recall that the probabilities P_1 and P_2 in (5.52) depend (nonlinearly) on the parameters of the risk neutral version of (5.47)–(5.48). The model parameters can therefore be estimated by minimizing the sum (across option price observation) squared pricing errors.

In one of several different estimations, HN estimate the model on option data for the first half 1992 and then evaluate the model by comparing implied and actual option prices for the second half of 1992. These implied option prices use the model parameters estimated on data for the first half of the year and an estimate of h_t calculated using these parameters and the latest S&P 500 index returns. The performance of this model is compared with a Black-Scholes model (among other models), where the implied volatility in week $t-1$ is used to price options in period t. This exercise is repeated for 1993 and 1994.

It is found that the GARCH model outperforms (in terms of MSE) the B-S model. In particular, it seems as if the GARCH model gives much smaller errors for deep out-of-the-money options (see Figures 2 and 3). HN argue that this is due to two aspects of the model: the time-profile of volatility (somewhat persistent, but mean-reverting) and the negative correlation of returns and volatility.

5.10 "Fundamental Values and Asset Returns in Global Equity Markets," by Bansal and Lundblad

Reference: Bansal and Lundblad (2002) (BL)

This paper studies how stock indices for five major markets are related to news about future cash flows (dividends and/or earnings). It uses monthly data on France, Germany, Japan, UK, US, and a world market index for the period 1973–1998.

BL argue that their present value model (stock price equals the present value of future

cash flows) can account for observed volatility of equity returns and the cross-correlation across markets. This is an interesting result since most earlier present value models have generated too small movements in returns—and also too small correlations across markets. The crucial features of the model are a predictable long-run component in cash flows and time-varying systematic risk.

5.10.1 Basic Model

It is assumed that the individual stock markets can be described by CAPM

$$R^e_{it} = \beta_i R^e_{mt} + \varepsilon_{it}, \tag{5.53}$$

where R^e_{mt} is the world market index. As in CAPM, the market return is proportional to its volatility—here modelled as a GARCH(1,1) process. We there fore have a GARCH-M ("-in-Mean") process

$$R^e_{mt} = \lambda\sigma^2_{mt} + \varepsilon_{mt},\ \mathrm{E}_{t-1}\,\varepsilon_{mt} = 0 \text{ and } \mathrm{Var}_{t-1}(\varepsilon_{mt}) = \sigma^2_{mt}, \tag{5.54}$$

$$\sigma^2_{mt} = \zeta + \gamma\varepsilon^2_{m,t-1} + \delta\sigma^2_{m,t-1}. \tag{5.55}$$

(Warning: BL uses a different timing/subscript convention for the GARCH model.)

5.10.2 The Price-Dividend Ratio

A gross return

$$R_{i,t+1} = \frac{D_{i,t+1} + P_{i,t+1}}{P_{it}}, \tag{5.56}$$

can be approximated in terms of logs (lower case letters)

$$r_{i,t+1} \approx \rho_i \underbrace{(p_{i,t+1} - d_{i,t+1})}_{z_{i,t+1}} - \underbrace{(p_{it} - d_{it})}_{z_{it}} + \underbrace{(d_{i,t+1} - d_{it})}_{g_{i,t+1}}, \tag{5.57}$$

where ρ_i is the average dividend-price ratio for asset i.

Take expectations as of t and solve recursively forward to get the log price/dividend ratio as a function of expected future dividend growth rates (g_i) and returns (r_i)

$$p_{it} - d_{it} = z_{it} \approx \sum_{s=0}^{\infty} \rho_i^s\, \mathrm{E}_t\,(g_{i,t+s+1} - r_{i,t+s+1})\,. \tag{5.58}$$

To calculate the right hand side of (5.58), notice the following things. First, the dividend growth ("cash flow dynamics") is modelled as an ARMA(1,1)—see below for details. Second, the riskfree rate (r_{ft}) is assumed to follow an AR(1). Third, the expected return equals the riskfree rate plus the expected excess return—which follows (5.53)–(5.55).

Since all these three processes are modelled as univariate first-order time-series processes, the solution is

$$p_{it} - d_{it} = z_{it} = A_{i,0} + A_{i,1} g_{it} + A_{i,2} \sigma^2_{m,t+1} + A_{i,3} r_{ft}. \tag{5.59}$$

(BL use an expected dividend growth instead of the actual but that is just a matter of convenience, and has another timing convention for the volatility.) This solution can be thought of as the "fundamental" (log) price-dividend ratio. The main theme of the paper is to study how well this fundamental log price-dividend ratio can explain the actual values.

The model is estimated by GMM (as a system), but most of the moment conditions are conventional. In practice, this means that *(i)* the betas and the AR(1) for the riskfree rate are estimated by OLS; *(ii)* the GARCH-M by MLE; *(iii)* the ARMA(1,1) process by moment conditions that require the innovations to be orthogonal to the current levels; and *(iv)* moment conditions for changes in $p_{it} - d_{it} = z_{it}$ define3d in (5.59). This is the "overidentified" part of the model.

5.10.3 A Benchmark Case with No Predictability

As a benchmark for comparison, consider the case when the right hand side in (5.58) equals a constant. This would happen when the growth rate of cash flows is unpredictable, the riskfree rate is constant, and the market risk premium is too (which here requires that the conditional variance of the market return is constant). In this case, the price-dividend ratio is constant, so the log return equals the cash flow growth plus a constant.

This benchmark case would not be very successful in matching the observed volatility and correlation (across markets) of returns: cash flow growth seems to be a lot less volatile than returns and also a lot less correlated across markets.

What if we allowed for predictability of cash flow growth, but still kept the assumptions of constant real interest rate and market risk premium? Large movements in predictable cash flow growth could then generate large movements in returns, but hardly the

correlation across markets.

However, large movements in the market risk premium would contribute to both. It is clear that both mechanisms are needed to get a correlation between zero and one. It can also be noted that the returns will be more correlated during volatile periods—since this drives up the market risk premium which is a common component in all returns.

5.10.4 Cash Flow Dynamics

The growth rate of cash flow, g_{it}, is modelled as an ARMA(1,1). The estimation results show that the AR parameter is around 0.95 and that the MA parameter is around -0.85. This means that the growth rate is almost an iid process with very low autocorrelation—but only almost. Since the MA parameter is not negative enough to make the sum of the AR and MA parameters zero, a positive shock to the growth rate will have a long-lived effect (even if small). See Figure 5.28.

Remark 5.27 *(ARMA(1,1)) An ARMA(1,1) model is*

$$y_t = ay_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}, \text{ where } \varepsilon_t \text{ is white noise.}$$

The model can be written on MA form as

$$y_t = \varepsilon_t + \sum_{s=1}^{\infty} a^{s-1}(a+\theta)\varepsilon_{t-s}.$$

The autocorrelations are

$$\rho_1 = \frac{(1+a\theta)(a+\theta)}{1+\theta^2+2a\theta}, \text{ and } \rho_s = a\rho_{s-1} \text{ for } s = 2,3,\ldots$$

and the conditional expectations are

$$\mathrm{E}_t\, y_{t+s} = a^{s-1}(ay_t + \theta\varepsilon_t),\ s = 1,2,\ldots$$

5.10.5 Results

1. The hypothesis that the CAPM regressions have zero intercepts (for all five country indices) cannot be rejected.

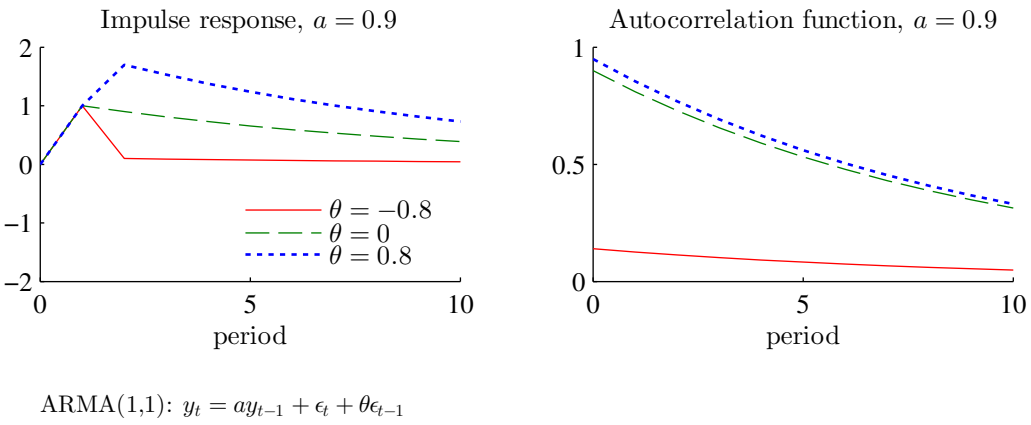


Figure 5.28: Impulse response and autcorrelation functions of ARMA(1,1)

2. Most of the parameters are precisely estimated, except λ (the risk aversion).

3. Market volatility is very persistent.

4. Cash flow has a small, but very persistent effect of news.

5. The overidentifying restrictions are rejected , but the model still seems able to account for quite a bit of the data: the volatility and correlation (across countries) of the fundamental price-dividend ratios are quite similar to those in the data. Note that the cross correlations are driven by the common movements in the riskfree rate and the world market risk premia (driven by σ_{mt}^2).

A Using an FFT to Calculate the PDF from the Characteristic Function

A.1 Characteristic Function

The characteristic function $h(x)$ of a random variable x is

$$\begin{aligned} h(\phi) &= \mathrm{E}\exp(i\phi x) \\ &= \int_{-\infty}^{\infty} \exp(i\phi x) f(x) dx, \end{aligned} \tag{A.1}$$

where $f(x)$ is the pdf. This is a Fourier transform of the pdf (if x is a continuous random variable). For instance, the cf of a $N(\mu, \sigma^2)$ distribution is $\exp(i\phi\mu - \phi^2\sigma^2/2)$. The pdf can therefore be recovered by the inverse Fourier transform as

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-i\phi x)h(\phi)d\phi. \tag{A.2}$$

In practice, we typically use a fast (discrete) Fourier transform to perform this calculation, since there are very quick computer algorithms for doing that.

A.2 FFT in Matlab

The *fft in Matlab* is

$$Q_k = \sum_{j=1}^{N} q_j e^{-\frac{2\pi i}{N}(j-1)(k-1)} \tag{A.3}$$

and the *ifft* is

$$q_j = \frac{1}{N}\sum_{k=1}^{N} Q_k e^{\frac{2\pi i}{N}(j-1)(k-1)}. \tag{A.4}$$

A.3 Invert the Characteristic Function

Approximate the characteristic function (A.1) as the integral over $[x_{\min}, x_{\max}]$ (assuming the pdf is zero outside)

$$h(\phi) = \int_{x_{\min}}^{x_{\max}} e^{i\phi x} f(x)dx. \tag{A.5}$$

Approximate this by a Riemann sum

$$h(\phi) \approx \sum_{k=1}^{N} e^{i\phi x_k} f(x_k)\Delta x. \tag{A.6}$$

Split up $[x_{\min}, x_{\max}]$ into N intervals of equal size, so the step (and interval width) is

$$\Delta x = \frac{x_{\max} - x_{\min}}{N}. \tag{A.7}$$

The mid point of the kth interval is

$$x_k = x_{\min} + (k - 1/2)\Delta x, \tag{A.8}$$

which means that $x_1 = x_{\min} + \Delta x/2$, $x_2 = x_{\min} + 1.5\Delta x$ and that $x_N = x_{\max} - \Delta x/2$.

Example A.1 *With* $(x_{\min}, x_{\max}) = (1, 7)$ *and* $N = 3$*, then* $\Delta x = (7-1)/3 = 2$*. The* x_j *values are*

$$\begin{bmatrix} \underline{k} & \underline{x_k = x_{\min} + (k-1/2)\Delta x} \\ 1 & 1 + 1/2 \times 2 = 2 \\ 2 & 1 + 3/2 \times 2 = 4 \\ 3 & 1 + 5/2 \times 2 = 6. \end{bmatrix}$$

This gives the Riemann sum

$$h_j \approx \sum_{k=1}^{N} e^{i\phi[x_{\min}+(k-1/2)\Delta x]} f_k \Delta x, \tag{A.9}$$

where $h_j = h(\phi_j)$ and $f_k = f(x_k)$.

We want

$$\phi_j = b + \frac{2\pi}{N}\frac{j-1}{\Delta x}, \tag{A.10}$$

so we can control the central location of ϕ. Use that in the Riemann sum

$$h_j \approx \sum_{k=1}^{N} e^{i[x_{\min}+(k-1/2)\Delta x]\frac{2\pi}{N}\frac{j-1}{\Delta x}} e^{i[x_{\min}+(k-1/2)\Delta x]b} f_k \Delta x, \tag{A.11}$$

and multiply both sides by $\exp\left[-i(x_{\min} + 1/2\Delta x)\frac{2\pi}{N}\frac{j-1}{\Delta x}\right]/N$ to get

$$\underbrace{e^{-i(x_{\min}+1/2\Delta x)\frac{2\pi}{N}\frac{j-1}{\Delta x}}\frac{1}{N}h_j}_{q_j} \approx \frac{1}{N}\sum_{k=1}^{N} e^{\frac{2\pi i}{N}(j-1)(k-1)} \underbrace{e^{i[x_{\min}+(k-1/2)\Delta x]b} f_k \Delta x}_{Q_k}, \tag{A.12}$$

which has the same for as the ifft (A.4). We should therefore be able to calculate Q_k by applying the fft (A.3) on q_j. We can then recover the density function as

$$f_k = e^{-i[x_{\min}+(k-1/2)\Delta x]b} Q_k/\Delta x. \tag{A.13}$$

Bibliography

Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold, 2005, "Volatility forecasting," Working Paper 11188, NBER.

Bansal, R., and C. Lundblad, 2002, "Market efficiency, fundamental values, and the size of the risk premium in global equity markets," *Journal of Econometrics*, 109, 195–237.

Britten-Jones, M., and A. Neuberger, 2000, "Option prices, implied price processes, and stochastic volatility," *Journal of Finance*, 55, 839–866.

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

Duan, J., 1995, "The GARCH option pricing model," *Mathematical Finance*, 5, 13–32.

Duffee, G. R., 2005, "Time variation in the covariance between stock returns and consumption growth," *Journal of Finance*, 60, 1673–1712.

Engle, R. F., 2002, "Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business and Economic Statistics*, 20, 339–351.

Engle, R. F., and K. Sheppard, 2001, "Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH," Discussion Paper 2001-15, University of California, San Diego.

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.

Gourieroux, C., and J. Jasiak, 2001, *Financial econometrics: problems, models, and methods*, Princeton University Press.

Hamilton, J. D., 1994, *Time series analysis*, Princeton University Press, Princeton.

Harvey, A. C., 1989, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.

Hentschel, L., 1995, "All in the family: nesting symmetric and asymmetric GARCH models," *Journal of Financial Economics*, 39, 71–104.

Heston, S. L., and S. Nandi, 2000, "A closed-form GARCH option valuation model," *Review of Financial Studies*, 13, 585–625.

Huang, C.-F., and R. H. Litzenberger, 1988, *Foundations for financial economics*, Elsevier Science Publishing, New York.

Jiang, G. J., and Y. S. Tian, 2005, "The model-free implied volatility and its information content," *Review of Financial Studies*, 18, 1305–1342.

Nelson, D. B., 1991, "Conditional heteroskedasticity in asset returns," *Econometrica*, 59, 347–370.

Ruiz, E., 1994, "Quasi-maximum likelihood estimation of stochastic volatility models," *Journal of Econometrics*, 63, 289–306.

Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.

6 Factor Models

Sections denoted by a star (*) is not required reading.

6.1 CAPM Tests: Overview

Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 5

Let $R_{it}^e = R_{it} - R_{ft}$ be the excess return on asset i in excess over the riskfree asset, and let $f_t = R_{mt} - R_{ft}$ be the excess return on the market portfolio. CAPM with a riskfree return says that $\alpha_i = 0$ in

$$\begin{aligned} R_{it}^e &= \alpha + \beta f_t + \varepsilon_{it}, \text{ where} \\ \mathrm{E}\,\varepsilon_{it} &= 0 \text{ and } \mathrm{Cov}(f_t, \varepsilon_{it}) = 0. \end{aligned} \tag{6.1}$$

The economic importance of a non-zero intercept (α) is that the tangency portfolio changes if the test asset is added to the investment opportunity set. See Figure 6.1 for an illustration.

The basic test of CAPM is to estimate (6.1) on a single asset and then test if the intercept is zero. This can easily be extended to several assets, where we test if all the intercepts are zero.

Notice that the test of CAPM can be given two interpretations. If we assume that R_{mt} is the correct benchmark, then it is a test of whether asset R_{it} is "correctly" priced (this is the approach in mutual fund evaluations). Alternatively, if we assume that R_{it} is correctly priced, then it is a test of the mean-variance efficiency of R_{mt} (compare the Roll critique).

6.2 Testing CAPM: Traditional LS Approach

6.2.1 CAPM with One Asset: Traditional LS Approach

If the residuals in the CAPM regression are iid, then the traditional LS approach is just fine: estimate (6.1) and form a t-test of the null hypothesis that the intercept is zero. If the disturbance is iid normally distributed, then this approach is the ML approach.

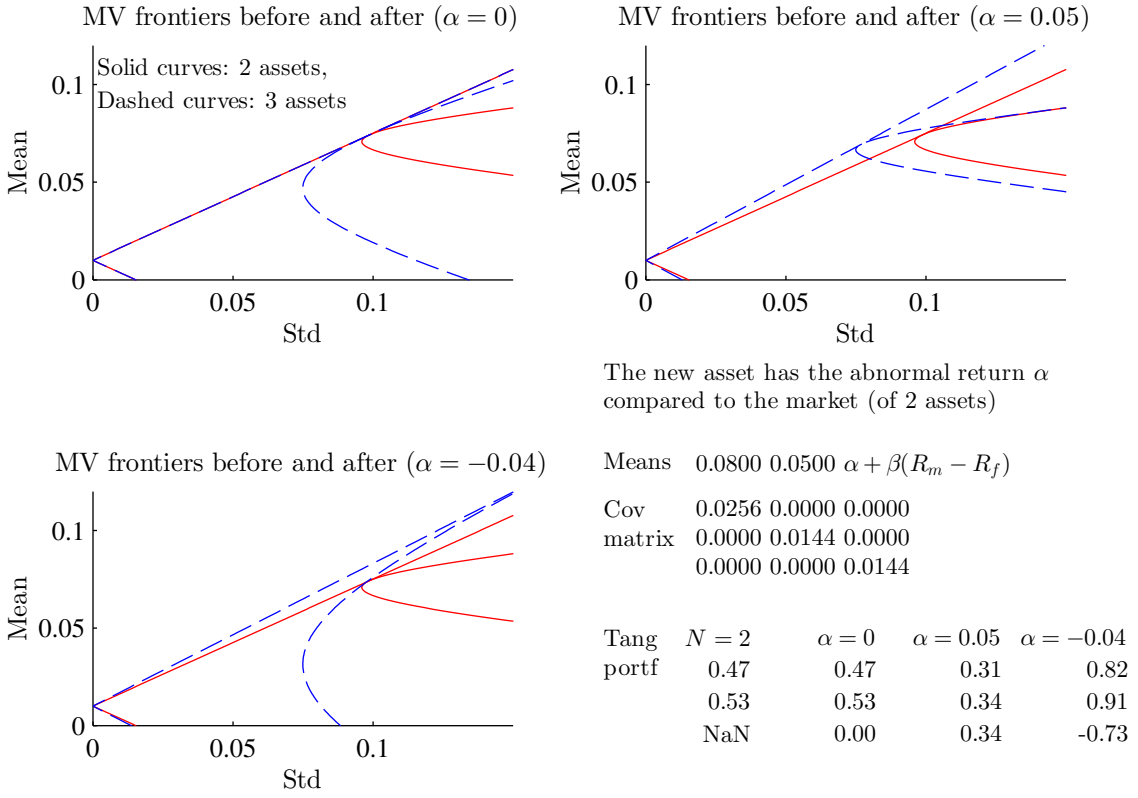


Figure 6.1: MV frontiers with 2 and 3 assets

The variance of the estimated intercept in the CAPM regression (6.1) is

$$\text{Var}(\hat{\alpha} - \alpha_0) = \left[1 + \frac{(\text{E}\, f_t)^2}{\text{Var}\,(f_t)}\right] \text{Var}(\varepsilon_{it})/T \tag{6.2}$$

$$= (1 + SR_f^2)\, \text{Var}(\varepsilon_{it})/T, \tag{6.3}$$

where SR_f^2 is the squared Sharpe ratio of the market portfolio (recall: f_t is the excess return on market portfolio). We see that the uncertainty about the intercept is high when the disturbance 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 gives a large uncertainty about what the risk-adjusted return should be.

Proof. (of (6.2)) Consider the regression equation $y_t = x_t' b_0 + u_t$. With iid errors that are independent of all regressors (also across observations), the LS estimator, $\hat{b}_{Ls}$, is

asymptotically distributed as

$$\sqrt{T}(\hat{b}_{Ls} - b_0) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Sigma_{xx}^{-1}), \text{ where } \sigma^2 = \mathrm{E}\, u_t^2 \text{ and } \Sigma_{xx} = \mathrm{E}\, \Sigma_{t=1}^{T} x_t x_t'/T.$$

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

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

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

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

The t-test of the hypothesis that $\alpha_0 = 0$ is then

$$\frac{\hat{\alpha}}{\mathrm{Std}(\hat{\alpha})} = \frac{\hat{\alpha}}{\sqrt{(1 + SR_f^2)\, \mathrm{Var}(\varepsilon_{it})/T}} \xrightarrow{d} N(0, 1) \text{ under } H_0\text{: } \alpha_0 = 0. \tag{6.4}$$

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).

Remark 6.1 *(Quadratic forms of normally distributed random variables) If the $n \times 1$ vector $X \sim N(0, \Sigma)$, then $Y = X'\Sigma^{-1}X \sim \chi_n^2$. Therefore, if the n scalar random variables X_i, $i = 1, ..., n$, are uncorrelated and have the distributions $N(0, \sigma_i^2)$, $i = 1, ..., n$, then $Y = \Sigma_{i=1}^{n} X_i^2/\sigma_i^2 \sim \chi_n^2$.*

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

$$\frac{\hat{\alpha}^2}{\mathrm{Var}(\hat{\alpha})} = \frac{\hat{\alpha}^2}{(1 + SR_f^2)\, \mathrm{Var}(\varepsilon_{it})/T} \xrightarrow{d} \chi_1^2 \text{ under } H_0\text{: } \alpha_0 = 0. \tag{6.5}$$

The chi-square test is equivalent to the t-test when we are testing only one restriction, but it has the advantage that it also allows us to test several restrictions at the same time. Both the t-test and the chi–square tests are Wald tests (estimate unrestricted model and then test the restrictions).

It is quite straightforward to use the properties of minimum-variance frontiers (see Gibbons, Ross, and Shanken (1989), and MacKinlay (1995)) to show that the test statistic

in (6.5) can be written

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} = \frac{(\widehat{SR}_c)^2 - (\widehat{SR}_f)^2}{[1 + (\widehat{SR}_f)^2]/T}, \tag{6.6}$$

where SR_f is the Sharpe ratio of the market portfolio and SR_c is the Sharpe ratio of the tangency portfolio when investment in both the market return and asset i is possible. (Recall that the tangency portfolio is the portfolio with the highest possible Sharpe ratio.) If the market portfolio has the same (squared) Sharpe ratio as the tangency portfolio of the mean-variance frontier of R_{it} and R_{mt} (so the market portfolio is mean-variance efficient also when we take R_{it} into account) then the test statistic, $\hat{\alpha}_i^2/\text{Var}(\hat{\alpha}_i)$, is zero—and CAPM is not rejected.

Proof. (of (6.6)) From the CAPM regression (6.1) we have

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

Suppose we use this information to construct a mean-variance frontier for both R_{it} and R_{mt}, and we find the tangency portfolio, with excess return R_{ct}^e. It is straightforward to show that the square of the Sharpe ratio of the tangency portfolio is $\mu^{e\prime}\Sigma^{-1}\mu^e$, 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

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

which we can write as

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

Use the notation $f_t = R_{mt} - R_{ft}$ and combine this with (6.3) and to get (6.6). ■

It is also possible to construct small sample test (that do not rely on any asymptotic results), which may be a better approximation of the correct distribution in real-life samples—provided the strong assumptions are (almost) satisfied. The most straightforward modification is to transform (6.5) into an $F_{1,T-1}$-test. This is the same as using a t-test in (6.4) since it is only one restriction that is tested (recall that if $Z \sim t_n$, then $Z^2 \sim F(1,n)$).

An alternative testing approach is to use an LR or LM approach: restrict the intercept

in the CAPM regression to be zero and estimate the model with ML (assuming that the errors are normally distributed). For instance, for an LR test, the likelihood value (when $\alpha = 0$) is then compared to the likelihood value without restrictions.

A common finding 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). 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.

6.2.2 CAPM with Several Assets: Traditional LS Approach

Suppose we have n test assets. Stack the expressions (6.1) for $i = 1, \ldots, n$ as

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} f_t + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}, \text{ where} \tag{6.7}$$
$$\mathrm{E}\,\varepsilon_{it} = 0 \text{ and } \mathrm{Cov}(f_t, \varepsilon_{it}) = 0.$$

This is a system of seemingly unrelated regressions (SUR)—with the same regressor (see, for instance, Greene (2003) 14). In this case, the efficient estimator (GLS) is LS on each equation separately. Moreover, the covariance matrix of the coefficients is particularly simple.

Under the null hypothesis of zero intercepts and iid residuals (although possibly correlated across regressions), the LS estimate of the intercept has the following asymptotic distribution

$$\sqrt{T}\hat{\alpha} \to^d N\left[\mathbf{0}_{n\times 1}, \Sigma(1 + SR^2)\right], \text{ where} \tag{6.8}$$
$$\Sigma = \begin{bmatrix} \sigma_{11} & \ldots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \ldots & \hat{\sigma}_{nn} \end{bmatrix} \text{ with } \sigma_{ij} = \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt}).$$

In practice, we use the sample moments for the covariance matrix, $\sigma_{ij} = \sum_{t=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{jt}/T$. This result is well known, but a simple proof is found in Appendix A.

To test the null hypothesis that all intercepts are zero, we then use the test statistic

$$T\hat{\alpha}'(1 + SR^2)^{-1}\Sigma^{-1}\hat{\alpha} \sim \chi^2_n, \text{ where } SR^2 = [\mathrm{E}\, f/\,\mathrm{Std}(f)]^2. \tag{6.9}$$

6.2.3 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)

$$\bar{R}^e_{jt} = \frac{1}{N_j}\sum_{i \in \text{Group}\, j} R^e_{it}, \tag{6.10}$$

where N_j is the number of individuals in group j.

Then, we run a factor model

$$\bar{R}^e_{jt} = x_t'\beta_j + v_{jt}, \text{ for } j = 1, 2, \ldots, M \tag{6.11}$$

where x_t typically includes a constant and various return factors (for instance, excess returns on equity and bonds). By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the "alpha") is higher for the Mth group than for the for first group.

Example 6.2 *(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

$$R_{it}^e = x_t'\beta_i + \varepsilon_{it}, \text{ for } i = 1, 2, \ldots, N \tag{6.12}$$

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

$$\hat{\beta}_{pi} = z_i'c_p + w_{pi}. \tag{6.13}$$

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 Testing CAPM: GMM

6.3.1 CAPM with Several Assets: GMM and a Wald Test

To test n assets at the same time when the errors are non-iid we make use of the GMM framework. A special case is when the residuals are iid. The results in this section will then coincide with those in Section 6.2.

Write the n regressions in (6.7) on vector form as

$$\begin{aligned} R_t^e &= \alpha + \beta f_t + \varepsilon_t, \text{ where} \\ \mathrm{E}\,\varepsilon_t &= \mathbf{0}_{n\times 1} \text{ and } \mathrm{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{1\times n}, \end{aligned} \tag{6.14}$$

where α and β are $n \times 1$ vectors. Clearly, setting $n = 1$ gives the case of a single test asset.

The $2n$ GMM moment conditions are that, at the true values of α and β,

$$\mathrm{E}\, g_t(\alpha, \beta) = \mathbf{0}_{2n\times 1}, \text{ where} \tag{6.15}$$

$$g_t(\alpha, \beta) = \begin{bmatrix} \varepsilon_t \\ f_t\varepsilon_t \end{bmatrix} = \begin{bmatrix} R_t^e - \alpha - \beta f_t \\ f_t\left(R_t^e - \alpha - \beta f_t\right) \end{bmatrix}. \tag{6.16}$$

There are as many parameters as moment conditions, so the GMM estimator picks values of α and β such that the sample analogues of (6.15) are satisfied exactly

$$\bar{g}(\hat{\alpha}, \hat{\beta}) = \frac{1}{T}\sum_{t=1}^{T} g_t(\hat{\alpha}, \hat{\beta}) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} R_t^e - \hat{\alpha} - \hat{\beta} f_t \\ f_t(R_t^e - \hat{\alpha} - \hat{\beta} f_t) \end{bmatrix} = \mathbf{0}_{2n\times 1}, \tag{6.17}$$

which gives the LS estimator. For the inference, we allow for the possibility of non-iid errors, but if the errors are actually iid, then we (asymptotically) get the same results as in Section 6.2.

With point estimates and their sampling distribution it is straightforward to set up a Wald test for the hypothesis that all elements in α are zero

$$\hat{\alpha}' \operatorname{Var}(\hat{\alpha})^{-1}\hat{\alpha} \xrightarrow{d} \chi_n^2. \tag{6.18}$$

Remark 6.3 *(Easy coding of the GMM Problem (6.17)) Estimate by LS, equation by equation. Then, plug in the fitted residuals in (6.16) to generate time series of the moments (will be important for the tests).*

Remark 6.4 *(Distribution of GMM) Let the parameter vector in the moment condition have the true value b_0. Define*

$$S_0 = \operatorname{Cov}\left[\sqrt{T}\,\bar{g}\,(b_0)\right] \text{ and } D_0 = \operatorname{plim}\frac{\partial \bar{g}(b_0)}{\partial b'}.$$

When the estimator solves $\min \bar{g}\,(b)'\, S_0^{-1}\bar{g}\,(b)$ *or when the model is exactly identified, the distribution of the GMM estimator is*

$$\sqrt{T}(\hat{b} - b_0) \xrightarrow{d} N\left(\mathbf{0}_{k\times 1}, V\right), \text{ where } V = \left(D_0' S_0^{-1} D_0\right)^{-1} = D_0^{-1} S_0 (D_0^{-1})'.$$

Details on the Wald Test*

Note that, with a linear model, the Jacobian of the moment conditions does not involve the parameters that we want to estimate. This means that we do not have to worry about

evaluating the Jacobian at the true parameter values. The probability limit of the Jacobian is simply the expected value, which can written as

$$\begin{aligned}\operatorname{plim}\frac{\partial \bar{g}_t(\alpha,\beta)}{\partial[\alpha,\beta]} = D_0 &= -\mathrm{E}\begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} \otimes I_n \\ &= -\mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right) \otimes I_n, \end{aligned} \tag{6.19}$$

where $\otimes$ is the Kronecker product. (The last expression applies also to the case of several factors.) Notice that we order the parameters as a column vector with the alphas first and the betas second. It might be useful to notice that in this case

$$D_0^{-1} = -\mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right)^{-1} \otimes I_n, \tag{6.20}$$

since $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (if conformable).

Remark 6.5 *(Kronecker product) If A and B are matrices, then*

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

Example 6.6 *(Two test assets) With assets 1 and 2, the parameter vector is $b = [\alpha_1, \alpha_2, \beta_1, \beta_2]'$. Write out (6.15) as*

$$\begin{bmatrix} \bar{g}_1(\alpha,\beta) \\ \bar{g}_2(\alpha,\beta) \\ \bar{g}_3(\alpha,\beta) \\ \bar{g}_4(\alpha,\beta) \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \end{bmatrix},$$

where $\bar{g}_1(\alpha,\beta)$ denotes the sample average of the first moment condition. The Jacobian

is

$$\frac{\partial \bar{g}(\alpha,\beta)}{\partial[\alpha_1,\alpha_2,\beta_1,\beta_2]'} = \begin{bmatrix} \partial\bar{g}_1/\partial\alpha_1 & \partial\bar{g}_1/\partial\alpha_2 & \partial\bar{g}_1/\partial\beta_1 & \partial\bar{g}_1/\partial\beta_2 \\ \partial\bar{g}_2/\partial\alpha_1 & \partial\bar{g}_2/\partial\alpha_2 & \partial\bar{g}_2/\partial\beta_1 & \partial\bar{g}_2/\partial\beta_2 \\ \partial\bar{g}_3/\partial\alpha_1 & \partial\bar{g}_3/\partial\alpha_2 & \partial\bar{g}_3/\partial\beta_1 & \partial\bar{g}_3/\partial\beta_2 \\ \partial\bar{g}_4/\partial\alpha_1 & \partial\bar{g}_4/\partial\alpha_2 & \partial\bar{g}_4/\partial\beta_1 & \partial\bar{g}_4/\partial\beta_2 \end{bmatrix}$$

$$= -\frac{1}{T}\sum\nolimits_{t=1}^{T} \begin{bmatrix} 1 & 0 & f_t & 0 \\ 0 & 1 & 0 & f_t \\ f_t & 0 & f_t^2 & 0 \\ 0 & f_t & 0 & f_t^2 \end{bmatrix} = -\frac{1}{T}\sum\nolimits_{t=1}^{T} \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix}' \right) \otimes I_2.$$

The asymptotic covariance matrix of $\sqrt{T}$ times the sample moment conditions, evaluated at the true parameter values, that is at the true disturbances, is defined as

$$S_0 = \text{Cov}\left(\frac{\sqrt{T}}{T} \sum_{t=1}^{T} g_t(\alpha,\beta) \right) = \sum_{s=-\infty}^{\infty} R(s), \text{ where} \tag{6.21}$$

$$R(s) = \text{E}\, g_t(\alpha,\beta) g_{t-s}(\alpha,\beta)'. \tag{6.22}$$

With n assets, we can write (6.22) in terms of the $n \times 1$ vector ε_t as

$$\begin{aligned} R(s) &= \text{E}\, g_t(\alpha,\beta) g_{t-s}(\alpha,\beta)' \\ &= \text{E} \begin{bmatrix} \varepsilon_t \\ f_t\varepsilon_t \end{bmatrix} \begin{bmatrix} \varepsilon_{t-s} \\ f_{t-s}\varepsilon_{t-s} \end{bmatrix}' \\ &= \text{E} \left[\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes \varepsilon_t \right) \left(\begin{bmatrix} 1 \\ f_{t-s} \end{bmatrix} \otimes \varepsilon_{t-s} \right)' \right]. \end{aligned} \tag{6.23}$$

(The last expression applies also to the case of several factors.)

The Newey-West estimator is often a good estimator of S_0, but the performance of the test improved, by imposing (correct, of course) restrictions on the $R(s)$ matrices.

From Remark 6.4, we can write the covariance matrix of the $2n \times 1$ vector of parameters (n parameters in α and another n in β) as

$$\text{Cov}\left(\sqrt{T} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \right) = D_0^{-1} S_0 (D_0^{-1})'. \tag{6.24}$$

Example 6.7 *(Special case 1: f_t is independent of ε_{t-s}, errors are iid, and $n = 1$) With*

these assumptions $R(s) = \mathbf{0}_{2\times 2}$ *if* $s \neq 0$, *and* $S_0 = \begin{bmatrix} 1 & \mathrm{E}\, f_t \\ \mathrm{E}\, f_t & \mathrm{E}\, f_t^2 \end{bmatrix} \mathrm{Var}(\varepsilon_{it})$. *Combining with (6.19) gives*

$$\mathrm{Cov}\left(\sqrt{T}\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}\right) = \begin{bmatrix} 1 & \mathrm{E}\, f_t \\ \mathrm{E}\, f_t & \mathrm{E}\, f_t^2 \end{bmatrix}^{-1} \mathrm{Var}(\varepsilon_{it}),$$

which is the same expression as $\sigma^2 \Sigma_{xx}^{-1}$ *in (6.2), which assumed iid errors.*

Example 6.8 *(Special case 2: as in Special case 1, but* $n \geq 1$*) With these assumptions* $R(s) = \mathbf{0}_{2n\times 2n}$ *if* $s \neq 0$, *and* $S_0 = \begin{bmatrix} 1 & \mathrm{E}\, f_t \\ \mathrm{E}\, f_t & \mathrm{E}\, f_t^2 \end{bmatrix} \otimes \mathrm{E}\, \varepsilon_t \varepsilon_t'$. *Combining with (6.19) gives*

$$\mathrm{Cov}\left(\sqrt{T}\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}\right) = \begin{bmatrix} 1 & \mathrm{E}\, f_t \\ \mathrm{E}\, f_t & \mathrm{E}\, f_t^2 \end{bmatrix}^{-1} \otimes \left(\mathrm{E}\, \varepsilon_t \varepsilon_t'\right).$$

This follows from the facts that $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ *and* $(A \otimes B)(C \otimes D) = AC \otimes BD$ *(if conformable). This is the same as in the SURE case.*

6.3.2 CAPM and Several Assets: GMM and an LM Test

We could also construct an "LM test" instead by imposing $\alpha = \mathbf{0}$ in the moment conditions (6.15) and (6.17). The moment conditions are then

$$\mathrm{E}\, g(\beta) = \mathrm{E}\begin{bmatrix} R_t^e - \beta f_t \\ f_t(R_t^e - \beta f_t) \end{bmatrix} = \mathbf{0}_{2n\times 1}. \tag{6.25}$$

Since there are $q = 2n$ moment conditions, but only n parameters (the β vector), this model is overidentified.

We could either use a weighting matrix in the GMM loss function or combine the moment conditions so the model becomes exactly identified.

With a weighting matrix, the estimator solves

$$\min_b \bar{g}(b)' W \bar{g}(b), \tag{6.26}$$

where $\bar{g}(b)$ is the sample average of the moments (evaluated at some parameter vector b), and W is a positive definite (and symmetric) weighting matrix. Once we have estimated

the model, we can test the n overidentifying restrictions that all $q = 2n$ moment conditions are satisfied at the estimated n parameters $\hat{\beta}$. If not, the restriction (null hypothesis) that $\alpha = \mathbf{0}_{n\times 1}$ is rejected. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used.

Alternatively, to combine the moment conditions so the model becomes exactly identified, premultiply by a matrix A to get

$$A_{n\times 2n}\,\mathrm{E}\,g(\beta) = \mathbf{0}_{n\times 1}. \tag{6.27}$$

The model is then tested by testing if all $2n$ moment conditions in (6.25) are satisfied at this vector of estimates of the betas. This is the GMM analogue to a classical LM test. Once again, the test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used.

Details on how to compute the estimates effectively are given in Appendix B.1.

For instance, to effectively use only the last n moment conditions in the estimation, we specify

$$A\,\mathrm{E}\,g(\beta) = \begin{bmatrix} 0_{n\times n} & I_n \end{bmatrix} \mathrm{E}\begin{bmatrix} R_t^e - \beta f_t \\ f_t(R_t^e - \beta f_t) \end{bmatrix} = \mathbf{0}_{n\times 1}. \tag{6.28}$$

This clearly gives the classical LS estimator without an intercept

$$\hat{\beta} = \frac{\sum_{t=1}^{T} f_t R_t^e/T}{\sum_{t=1}^{T} f_t^2/T}. \tag{6.29}$$

Example 6.9 *(Combining moment conditions, CAPM on two assets) With two assets we can combine the four moment conditions into only two by*

$$A\,\mathrm{E}\,g_t(\beta_1,\beta_2) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathrm{E}\begin{bmatrix} R_{1t}^e - \beta_1 f_t \\ R_{2t}^e - \beta_2 f_t \\ f_t(R_{1t}^e - \beta_1 f_t) \\ f_t(R_{2t}^e - \beta_2 f_t) \end{bmatrix} = \mathbf{0}_{2\times 1}.$$

Remark 6.10 *(Test of overidentifying assumption in GMM) When the GMM estimator solves the quadratic loss function $\bar{g}(\beta)'S_0^{-1}\bar{g}(\beta)$ (or is exactly identified), then the J test statistic is*

$$T\bar{g}(\hat{\beta})'S_0^{-1}\bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k},$$

where q is the number of moment conditions and k is the number of parameters.

Remark 6.11 *(Distribution of GMM, more general results) When GMM solves* $\min_b \bar{g}(b)'W\bar{g}(b)$ *or* $A\bar{g}(\hat{\beta}) = \mathbf{0}_{k\times 1}$*, the distribution of the GMM estimator and the test of overidentifying assumptions are different than in Remarks 6.4 and 6.10.*

6.3.3 Size and Power of the CAPM Tests

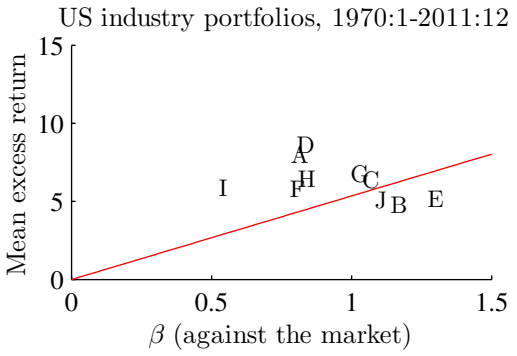
The size (using asymptotic critical values) and power in small samples is often found to be disappointing. Typically, these tests tend to reject a true null hypothesis too often (see Campbell, Lo, and MacKinlay (1997) Table 5.1) and the power to reject a false null hypothesis is often fairly low. These features are especially pronounced when the sample is small and the number of assets, n, is high. One useful rule of thumb is that a *saturation ratio* (the number of observations per parameter) below 10 (or so) is likely to give poor performance of the test. In the test here we have nT observations, $2n$ parameters in α and β, and $n(n+1)/2$ unique parameters in S_0, so the saturation ratio is $T/(2+(n+1)/2)$. For instance, with $T = 60$ and $n = 10$ or at $T = 100$ and $n = 20$, we have a saturation ratio of 8, which is very low (compare Table 5.1 in CLM).

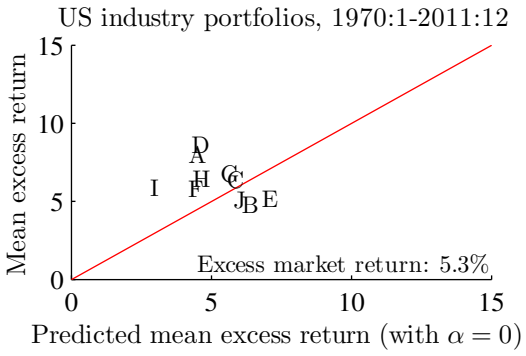
One possible way of dealing with the wrong size of the test is to use critical values from simulations of the small sample distributions (Monte Carlo simulations or bootstrap simulations).

6.3.4 Choice of Portfolios

This type of test is typically done on portfolios of assets, rather than on the individual assets themselves. There are several econometric and economic reasons for this. The econometric techniques we apply need the returns to be (reasonably) stationary in the sense that they have approximately the same means and covariance (with other returns) throughout the sample (individual assets, especially stocks, can change character as the company moves into another business). It might be more plausible that size or industry portfolios are stationary in this sense. Individual portfolios are typically very volatile, which makes it hard to obtain precise estimate and to be able to reject anything.

It sometimes makes economic sense to sort the assets according to a characteristic (size or perhaps book/market)—and then test if the model is true for these portfolios. Rejection of the CAPM for such portfolios may have an interest in itself.

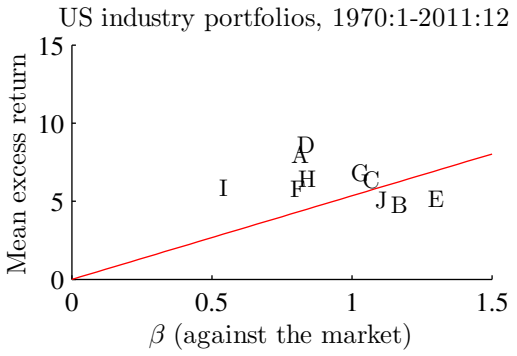




	alpha	pval	StdErr
all	NaN	0.02	NaN
A (NoDur)	3.79	0.01	8.86
B (Durbl)	-1.33	0.51	13.50
C (Manuf)	0.84	0.40	6.31
D (Enrgy)	4.30	0.06	14.62
E (HiTec)	-1.64	0.38	12.08
F (Telcm)	1.65	0.35	11.28
G (Shops)	1.46	0.34	9.84
H (Hlth)	2.10	0.24	11.63
I (Utils)	3.03	0.10	11.66
J (Other)	-0.70	0.53	7.15

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

Figure 6.2: CAPM, US industry portfolios



	alpha	t LS	t NW	t boot
all	NaN	NaN	NaN	NaN
A (NoDur)	3.79	2.76	2.75	2.74
B (Durbl)	-1.33	-0.64	-0.65	-0.64
C (Manuf)	0.84	0.85	0.84	0.84
D (Enrgy)	4.30	1.90	1.91	1.94
E (HiTec)	-1.64	-0.88	-0.88	-0.87
F (Telcm)	1.65	0.94	0.94	0.95
G (Shops)	1.46	0.95	0.96	0.95
H (Hlth)	2.10	1.17	1.19	1.18
I (Utils)	3.03	1.68	1.65	1.63
J (Other)	-0.70	-0.63	-0.62	-0.62

NW uses 1 lag
The bootstrap samples pairs of (y_t, x_t)
3000 simulations

Figure 6.3: CAPM, US industry portfolios, different t-stats

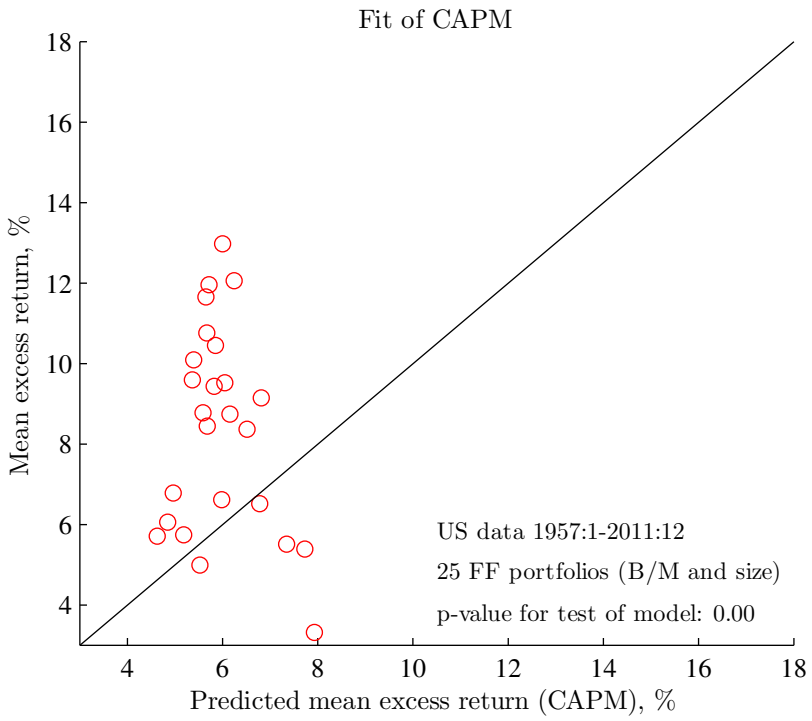


Figure 6.4: CAPM, FF portfolios

6.3.5 Empirical Evidence

See Campbell, Lo, and MacKinlay (1997) 6.5 (Table 6.1 in particular) and Cochrane (2005) 20.2.

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 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 quantile *and* the jth BE/ME quantile. This is illustrated in Table 6.1.

Tables 6.2–6.3 summarize some basic properties of these portfolios.

Fama and French 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 $\mathrm{E}\, R^e_{it} = \beta_i \mathrm{E}\, R^e_{mt}$). However, there is little relation between $\mathrm{E}\, R^e_{it}$ and β_i (see Figure 6.4). This

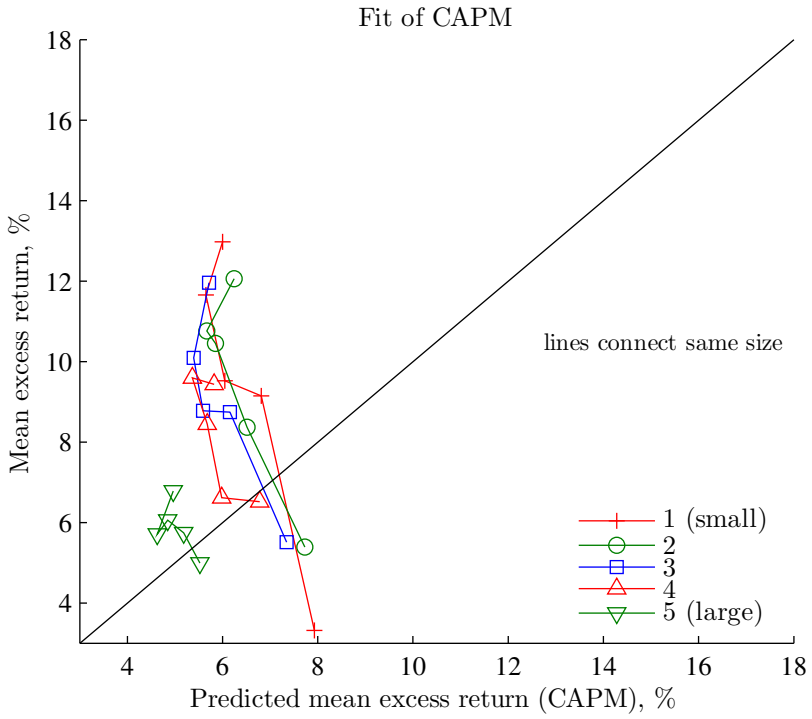


Figure 6.5: CAPM, FF portfolios

	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 6.1: Numbering of the FF indices in the figures.

lack of relation (a cloud in the $\beta_i \times \mathrm{E}\, R^e_{it}$ space) is due to the combination of two features of the data. First, *within a size quantile* there is a negative relation (across BE/ME quantiles) between $\mathrm{E}\, R^e_{it}$ and β_i—in stark contrast to CAPM (see Figure 6.5). Second, *within a BE/ME quantile*, there is a positive relation (across size quantiles) between $\mathrm{E}\, R^e_{it}$ and β_i—as predicted by CAPM (see Figure 6.6).

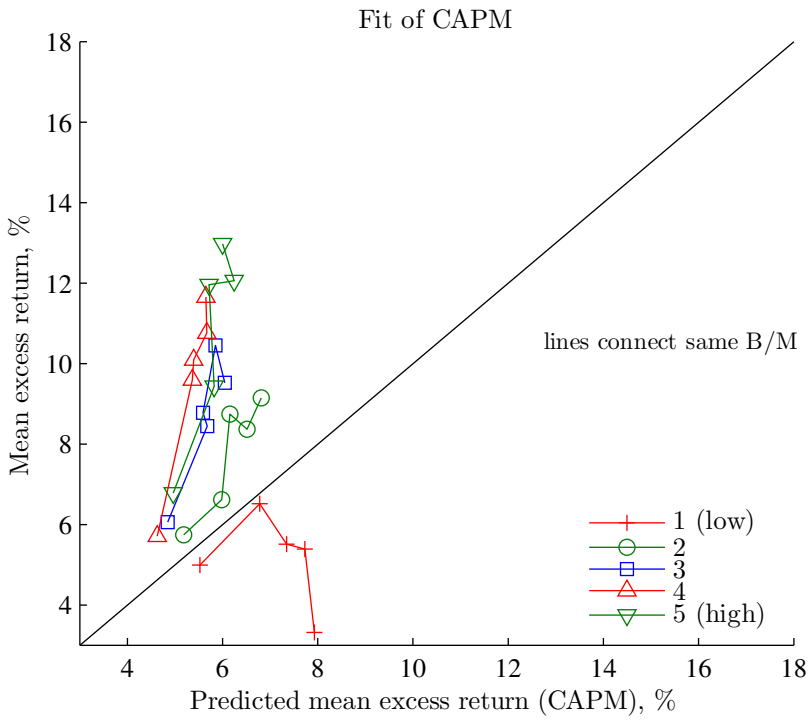


Figure 6.6: CAPM, FF portfolios

	Book value/Market value				
	1	2	3	4	5
Size 1	3.3	9.1	9.5	11.7	13.0
2	5.4	8.4	10.4	10.8	12.1
3	5.5	8.7	8.8	10.1	12.0
4	6.5	6.6	8.4	9.6	9.4
5	5.0	5.7	6.1	5.7	6.8

Table 6.2: Mean excess returns (annualised %), US data 1957:1–2011:12. 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).

6.4 Testing Multi-Factor Models (Factors are Excess Returns)

Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 6.2.1

	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	1.1	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 6.3: Beta against the market portfolio, US data 1957:1–2011:12. 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).

6.4.1 A Multi-Factor Model

When the K factors, f_t, are excess returns, the null hypothesis typically says that $\alpha_i = 0$ in

$$\begin{aligned} R_{it}^e &= \alpha_i + \beta_i' f_t + \varepsilon_{it}, \text{ where} \\ \mathrm{E}\,\varepsilon_{it} &= 0 \text{ and } \mathrm{Cov}(f_t, \varepsilon_{it}) = \mathbf{0}_{K\times 1}. \end{aligned} \tag{6.30}$$

and β_i is now an $K \times 1$ vector. The CAPM regression is a special case when the market excess return is the only factor. In other models like ICAPM (see Cochrane (2005) 9.2), we typically have several factors. We stack the returns for n assets to get

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_{11} & \dots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \dots & \beta_{nK} \end{bmatrix} \begin{bmatrix} f_{1t} \\ \vdots \\ f_{Kt} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}, \text{ or}$$

$$\begin{aligned} R_t^e &= \alpha + \beta f_t + \varepsilon_t, \text{ where} \\ \mathrm{E}\,\varepsilon_t &= \mathbf{0}_{n\times 1} \text{ and } \mathrm{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{K\times n}, \end{aligned} \tag{6.31}$$

where α is $n \times 1$ and β is $n \times K$. Notice that β_{ij} shows how the ith asset depends on the jth factor.

This is, of course, very similar to the CAPM (one-factor) model—and both the LS and GMM approaches are straightforward to extend.

6.4.2 Multi-Factor Model: Traditional LS (SURE)

The results from the LS approach of testing CAPM generalizes directly. In particular, (6.9) still holds—but where the residuals are from the multi-factor regressions (6.30) and where the Sharpe ratio of the tangency portfolio (based on the factors) depends on the means and covariance matrix of all factors

$$T\hat{\alpha}'(1+SR^2)^{-1}\Sigma^{-1}\hat{\alpha} \sim \chi^2_n, \text{ where} \qquad (6.32)$$
$$SR^2 = \mathrm{E}\, f' \,\mathrm{Cov}(f)^{-1}\,\mathrm{E}\, f.$$

This result is well known, but some properties of SURE models are found in Appendix A.

6.4.3 Multi-Factor Model: GMM

The moment conditions are

$$\mathrm{E}\, g_t(\alpha,\beta) = \mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes \varepsilon_t\right) = \mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R^e_t - \alpha - \beta f_t)\right) = \mathbf{0}_{n(1+K)\times 1}. \qquad (6.33)$$

Note that this expression looks similar to (6.15)—the only difference is that f_t may now be a vector (and we therefore need to use the Kronecker product). It is then intuitively clear that the expressions for the asymptotic covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ will look very similar too.

When the system is exactly identified, the GMM estimator solves

$$\bar{g}(\alpha,\beta) = \mathbf{0}_{n(1+K)\times 1}, \qquad (6.34)$$

which is the same as LS equation by equation. The model can be tested by testing if all alphas are zero—as in (6.18).

Instead, when we restrict $\alpha = \mathbf{0}_{n\times 1}$ (overidentified system), then we either specify a weighting matrix W and solve

$$\min_\beta\, \bar{g}(\beta)' W \bar{g}(\beta), \qquad (6.35)$$

or we specify a matrix A to combine the moment conditions and solve

$$A_{nK\times n(1+K)}\bar{g}(\beta) = \mathbf{0}_{nK\times 1}. \tag{6.36}$$

For instance, to get the classical LS estimator without intercepts we specify

$$A = \begin{bmatrix} 0_{nK\times n} & I_{nK} \end{bmatrix} \mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \beta f_t) \right). \tag{6.37}$$

More generally, details on how to compute the estimates effectively are given in Appendix B.1.

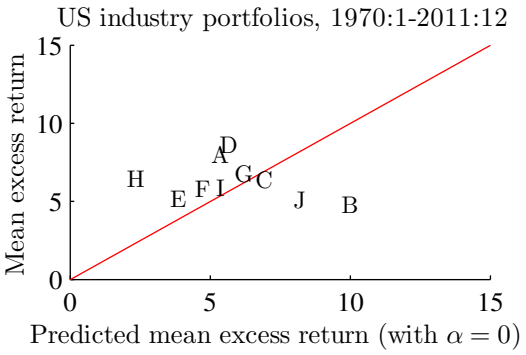
Example 6.12 *(Moment condition with two assets and two factors) The moment conditions for $n = 2$ and $K = 2$ are*

$$\mathrm{E}\, g_t(\alpha,\beta) = \mathrm{E} \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_{11} f_{1t} - \beta_{12} f_{2t} \\ R_{2t}^e - \alpha_2 - \beta_{21} f_{1t} - \beta_{22} f_{2t} \\ f_{1t}(R_{1t}^e - \alpha_1 - \beta_{11} f_{1t} - \beta_{12} f_{2t}) \\ f_{1t}(R_{2t}^e - \alpha_2 - \beta_{21} f_{1t} - \beta_{22} f_{2t}) \\ f_{2t}(R_{1t}^e - \alpha_1 - \beta_{11} f_{1t} - \beta_{12} f_{2t}) \\ f_{2t}(R_{2t}^e - \alpha_2 - \beta_{21} f_{1t} - \beta_{22} f_{2t}) \end{bmatrix} = \mathbf{0}_{6\times 1}.$$

Restricting $\alpha_1 = \alpha_2 = 0$ gives the moment conditions for the overidentified case.

Details on the Wald Test*

For the exactly identified case, we have the following results. The expressions for the Jacobian D_0 and its inverse are the same as in (6.19)–(6.20). Notice that in this Jacobian we differentiate the moment conditions (6.33) with respect to $\mathrm{vec}(\alpha, \beta)$, that is, where the parameters are stacked in a column vector with the alphas first, then the betas for the first factor, followed by the betas for the second factor etc. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used. The covariance matrix of the average moment conditions are as in (6.21)–(6.23).



	alpha	pval	StdErr
all	NaN	0.00	NaN
A (NoDur)	2.94	0.03	8.64
B (Durbl)	-4.92	0.01	12.23
C (Manuf)	-0.23	0.80	6.03
D (Enrgy)	3.26	0.14	14.14
E (HiTec)	1.59	0.32	10.09
F (Telcm)	1.37	0.45	11.04
G (Shops)	0.91	0.55	9.74
H (Hlth)	4.41	0.01	10.86
I (Utils)	0.63	0.71	10.47
J (Other)	-2.88	0.00	6.12

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

Figure 6.7: Three-factor model, US industry portfolios

6.4.4 Empirical Evidence

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 (see Campbell, Lo, and MacKinlay (1997) Table 6.1 or Fama and French (1993) Table 9c), but it can still capture a fair amount of the variation of expected returns—see Figures 6.7–6.10.

6.5 Testing Multi-Factor Models (General Factors)

Reference: Cochrane (2005) 12.2; Campbell, Lo, and MacKinlay (1997) 6.2.3 and 6.3

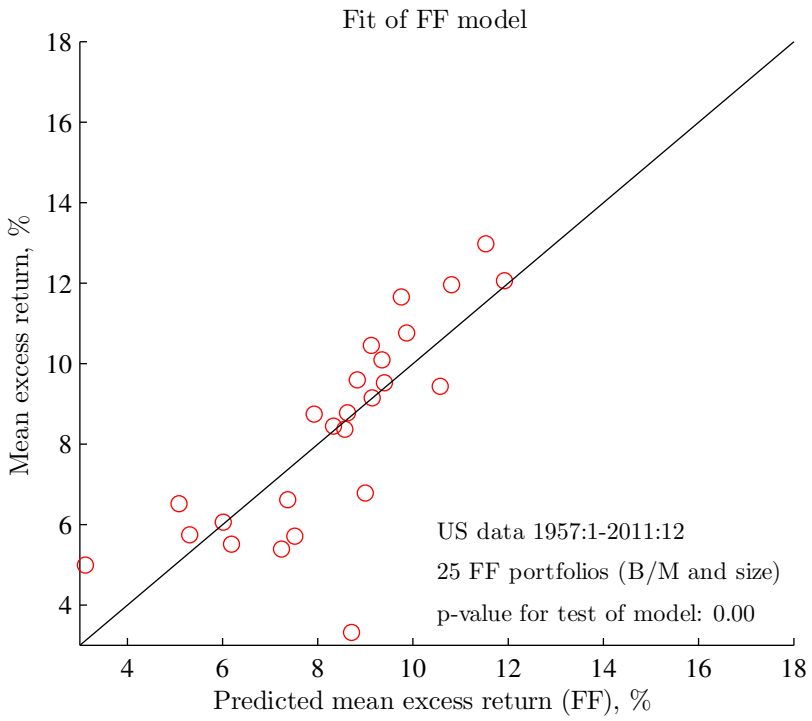


Figure 6.8: FF, FF portfolios

6.5.1 GMM Estimation with General Factors

Linear factor models imply that all expected excess returns are linear functions of the same vector of factor risk premia (λ)

$$\mathrm{E}\, R_{it}^e = \beta_i' \lambda, \text{ where } \lambda \text{ is } K \times 1, \text{ for } i = 1, \ldots n. \tag{6.38}$$

Stacking the test assets gives

$$\mathrm{E}\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \beta_{11} & \ldots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \ldots & \beta_{nK} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_K \end{bmatrix}, \text{ or}$$

$$\mathrm{E}\, R_t^e = \beta\lambda, \tag{6.39}$$

where β is $n \times K$.

When the factors are excess returns, then the factor risk premia must equal the ex-

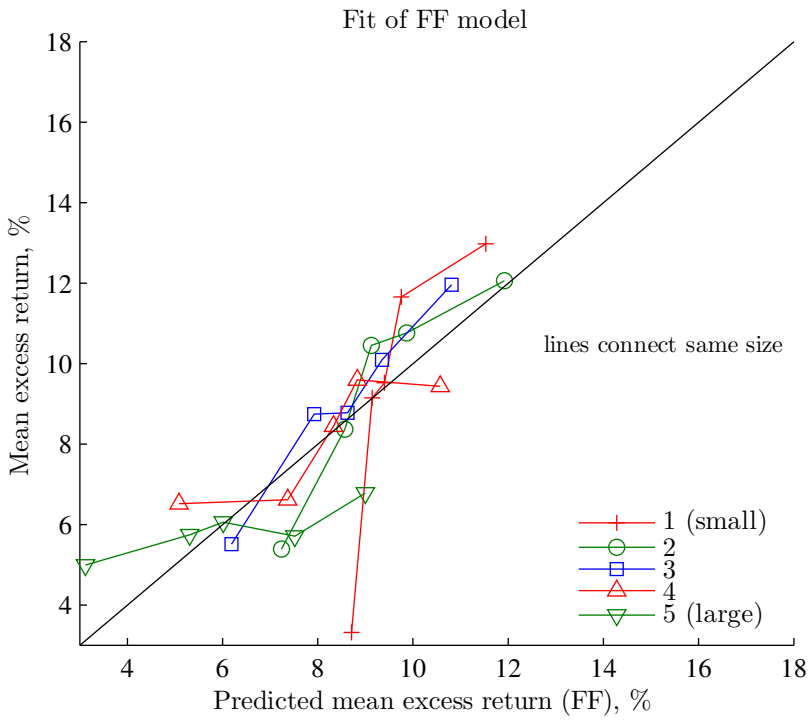


Figure 6.9: FF, FF portfolios

pected excess returns of those factors. (To see this, let the factor also be one of the test assets. It will then get a beta equal to unity on itself (for instance, regressing R^e_{mt} on itself must give a coefficient equal to unity). This shows that for factor k, $\lambda_k = \mathrm{E}\, R^e_{kt}$. More generally, the factor risk premia can be interpreted as follows. Consider an asset that has a beta of unity against factor k and zero betas against all other factors. This asset will have an expected excess return equal to λ_k. For instance, if a factor risk premium is negative, then assets that are positively exposed to it (positive betas) will have a negative risk premium—and vice versa.

The old way of testing this is to do a two-step estimation: first, estimate the β_i vectors in a time series model like (6.31) (equation by equation); second, use $\hat{\beta}_i$ as regressors in a regression equation of the type (6.38) with a residual added

$$\Sigma_{t=1}^{T} R^e_{it}/T = \hat{\beta}_i'\lambda + u_i. \tag{6.40}$$

It is then tested if $u_i = 0$ for all assets $i = 1, \ldots, n$. This approach is often called a

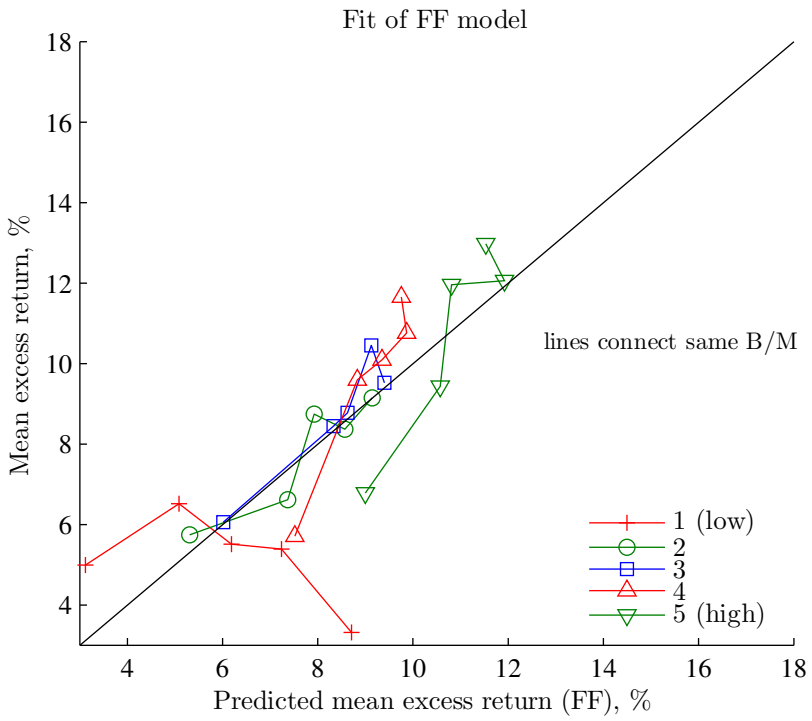


Figure 6.10: FF, FF portfolios

cross-sectional regression while the previous tests are time series regressions. The main problem of the cross-sectional approach is that we have to account for the fact that the regressors in the second step, $\hat{\beta}_i$, are just estimates and therefore contain estimation errors. This errors-in-variables problem is likely to have two effects *(i)* it gives a downwards bias of the estimates of λ and an upward bias of the mean of the fitted residuals; and *(ii)* invalidates the standard expression of the test of λ.

A way to handle these problems is to combine the moment conditions for the regression function (6.33) (to estimate β) with (6.39) (to estimate λ) to get a joint system

$$\mathrm{E}\, g_t(\alpha, \beta, \lambda) = \mathrm{E}\begin{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \\ R_t^e - \beta\lambda \end{bmatrix} = \mathbf{0}_{n(1+K+1)\times 1}. \tag{6.41}$$

See Figures 6.11–6.13 for an empirical example of a co-skewness model.

We can then test the overidentifying restrictions of the model. There are $n(1 + K +$

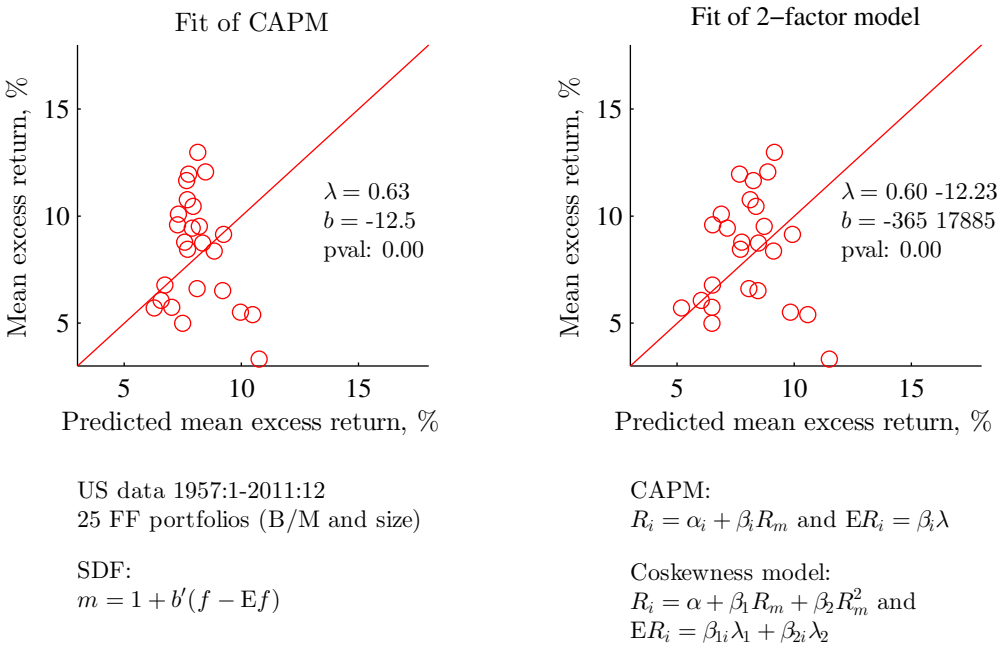
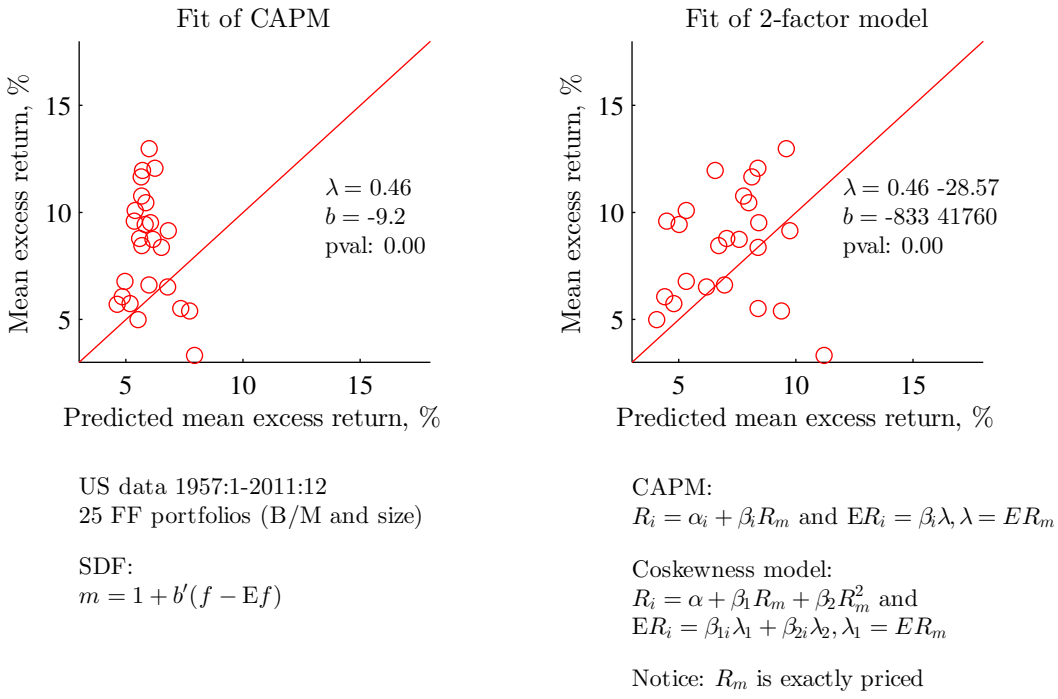


Figure 6.11: CAPM and quadratic model

1) moment condition (for each asset we have one moment condition for the constant, K moment conditions for the K factors, and one moment condition corresponding to the restriction on the linear factor model). There are only $n(1+K)+K$ parameters (n in α, nK in β and K in λ). We therefore have $n-K$ overidentifying restrictions which can be tested with a chi-square test. Notice that this is, in general, a non-linear estimation problem, since the parameters in β multiply the parameters in λ. From the GMM estimation using (6.41) we get estimates of the factor risk premia and also the variance-covariance of them. This allows us to not only test the moment conditions, but also to characterize the risk factors and to test if they are priced (each of them, or perhaps all jointly) by using a Wald test.

One approach to estimate the model is to specify a weighting matrix W and then solve a minimization problem like (6.35). The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used. In the special case of $W = S_0^{-1}$, the distribution is given by Remark 6.4. For other choices of the weighting matrix, the expression for the covariance matrix is more complicated.



US data 1957:1-2011:12
25 FF portfolios (B/M and size)

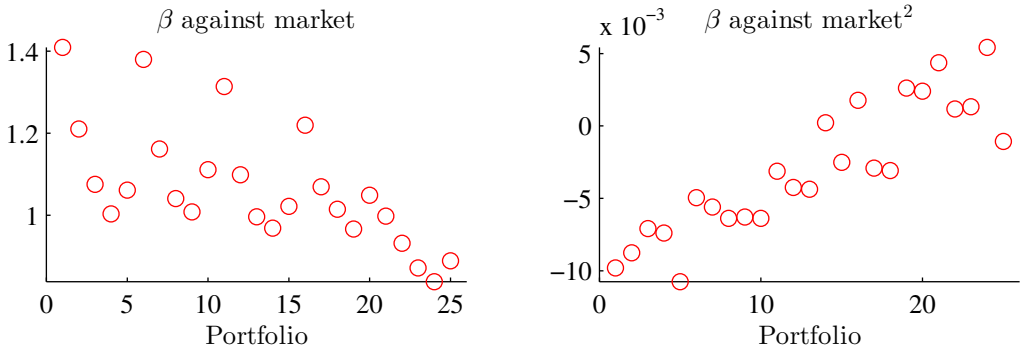
SDF:
$m = 1 + b'(f - \mathrm{E}f)$

CAPM:
$R_i = \alpha_i + \beta_i R_m$ and $\mathrm{E}R_i = \beta_i \lambda, \lambda = ER_m$

Coskewness model:
$R_i = \alpha + \beta_1 R_m + \beta_2 R_m^2$ and
$\mathrm{E}R_i = \beta_{1i}\lambda_1 + \beta_{2i}\lambda_2, \lambda_1 = ER_m$

Notice: R_m is exactly priced

Figure 6.12: CAPM and quadratic model, market excess is exactly priced



US data 1957:1–2011:12

25 FF portfolios (B/M and size)

Figure 6.13: CAPM and quadratic model

It is straightforward to show that the Jacobian of these moment conditions (with respect to vec(α, β, λ)) is

$$D_0 = -\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left(\begin{bmatrix}1\\ f_t\end{bmatrix}\begin{bmatrix}1\\ f_t\end{bmatrix}'\right)\otimes I_n & \mathbf{0}_{n(1+K)\times K} \\ \begin{bmatrix}0 & \lambda'\end{bmatrix}\otimes I_n & \beta_{n\times K} \end{bmatrix} \tag{6.42}$$

where the upper left block is similar to the expression for the case with excess return factors (6.19), while the other blocks are new.

Example 6.13 *(Two assets and one factor) we have the moment conditions*

$$\mathrm{E}\, g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \mathrm{E}\begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \\ R^e_{1t} - \beta_1\lambda \\ R^e_{2t} - \beta_2\lambda \end{bmatrix} = \mathbf{0}_{6\times 1}.$$

There are then 6 moment conditions and 5 parameters, so there is one overidentifying restriction to test. Note that with one factor, then we need at least two assets for this testing approach to work ($n - K = 2 - 1$). In general, we need at least one more asset than factors. In this case, the Jacobian is

$$\begin{aligned} \frac{\partial \bar{g}}{\partial[\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda]'} &= -\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} 1 & 0 & f_t & 0 & 0 \\ 0 & 1 & 0 & f_t & 0 \\ f_t & 0 & f_t^2 & 0 & 0 \\ 0 & f_t & 0 & f_t^2 & 0 \\ 0 & 0 & \lambda & 0 & \beta_1 \\ 0 & 0 & 0 & \lambda & \beta_2 \end{bmatrix} \\ &= -\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left(\begin{bmatrix}1\\ f_t\end{bmatrix}\begin{bmatrix}1\\ f_t\end{bmatrix}'\right)\otimes I_2 & \mathbf{0}_{4\times 1} \\ [0, \lambda]\otimes I_2 & \beta \end{bmatrix}. \end{aligned}$$

6.5.2 Traditional Cross-Sectional Regressions as Special Cases

Instead of estimating the overidentified model (6.41) (by specifying a weighting matrix), we could combine the moment equations so they become equal to the number of parameters. This can be done, by specifying a matrix A and combine as $A\,\mathrm{E}\,g_t = \mathbf{0}$. This does not generate any overidentifying restrictions, but it still allows us to test hypotheses about some moment conditions and about λ. One possibility is to let the upper left block of A be an identity matrix and just combine the last n moment conditions, $R_t^e - \beta\lambda$, to just K moment conditions

$$A\,\mathrm{E}\,g_t = \mathbf{0}_{[n(1+K)+K]\times 1} \tag{6.43}$$

$$\begin{bmatrix} I_{n(1+K)} & \mathbf{0}_{n(1+K)\times n} \\ \mathbf{0}_{K\times n(1+K)} & \theta_{K\times n} \end{bmatrix} \mathrm{E}\begin{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \\ R_t^e - \beta\lambda \end{bmatrix} = \mathbf{0} \tag{6.44}$$

$$\mathrm{E}\begin{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \\ \theta(R_t^e - \beta\lambda) \end{bmatrix} = \mathbf{0} \tag{6.45}$$

Here A has $n(1 + K) + K$ rows (which equals the number of parameters (α, β, λ)) and $n(1 + K + 1)$ columns (which equals the number of moment conditions). (Notice also that θ is $K \times n$, β is $n \times K$ and λ is $K \times 1$.)

Remark 6.14 *(Calculation of the estimates based on (6.44)) In this case, we can estimate α and β with LS equation by equation—as a standard time-series regression of a factor model. To estimate the $K \times 1$ vector λ, notice that we can solve the second set of K moment conditions as*

$$\theta\,\mathrm{E}(R_t^e - \beta\lambda) = \mathbf{0}_{K\times 1} \text{ or } \lambda = (\theta\beta)^{-1}\,\theta\,\mathrm{E}\,R_t^e,$$

which is just like a cross-sectional instrumental variables regression of $\mathrm{E}\,R_t^e = \beta\lambda$ (with β being the regressors, θ the instruments, and $\mathrm{E}\,R_t^e$ the dependent variable).

With $\theta = \beta'$, we get the traditional cross-sectional approach (6.38). The only difference is we here take the uncertainty about the generated betas into account (in the testing). Alternatively, let Σ be the covariance matrix of the residuals from the time-series estima-

tion of the factor model. Then, using $\theta = \beta'\Sigma$ gives a traditional GLS cross-sectional approach.

To test the asset pricing implications, we test if the moment conditions $\mathrm{E}\, g_t = \mathbf{0}$ in (6.43) are satisfied at the estimated parameters. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used (typically more complicated than in Remark 6.4).

Example 6.15 *(LS cross-sectional regression, two assets and one factor) With the moment conditions in Example (6.13) and the weighting vector $\theta = [\beta_1, \beta_2]$ (6.45) is*

$$A\,\mathrm{E}\, g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \mathrm{E}\begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \\ \beta_1(R^e_{1t} - \beta_1\lambda) + \beta_2(R^e_{2t} - \beta_2\lambda) \end{bmatrix} = \mathbf{0}_{5\times 1},$$

which has as many parameters as moment conditions. The test of the asset pricing model is then to test if

$$\mathrm{E}\, g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \mathrm{E}\begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \\ R^e_{1t} - \beta_1\lambda \\ R^e_{2t} - \beta_2\lambda \end{bmatrix} = \mathbf{0}_{6\times 1},$$

are satisfied at the estimated parameters.

Example 6.16 *(Structure of $\theta\,\mathrm{E}(R^e_t - \beta\lambda)$) If there are 2 factors and three test assets, then $0_{2\times 1} = \theta\,\mathrm{E}(R^e_t - \beta\lambda)$ is*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \left(\begin{bmatrix} \mathrm{E}\, R^e_{1t} \\ \mathrm{E}\, R^e_{2t} \\ \mathrm{E}\, R^e_{3t} \end{bmatrix} - \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \right).$$

6.5.3 Alternative Formulation of Moment Conditions*

The test of the general multi-factor models is sometimes written on a slightly different form (see, for instance, Campbell, Lo, and MacKinlay (1997) 6.2.3, but adjust for the fact that they look at returns rather than excess returns). To illustrate this, note that the regression equations (6.31) imply that

$$\mathrm{E}\, R_t^e = \alpha + \beta\, \mathrm{E}\, f_t. \tag{6.46}$$

Equate the expected returns of (6.46) and (6.38) to get

$$\alpha = \beta(\lambda - \mathrm{E}\, f_t), \tag{6.47}$$

which is another way of summarizing the restrictions that the linear factor model gives. We can then rewrite the moment conditions (6.41) as (substitute for α and skip the last set of moments)

$$\mathrm{E}\, g_t(\beta, \lambda) = \mathrm{E}\left[\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \beta(\lambda - \mathrm{E}\, f_t) - \beta f_t)\right] = \mathbf{0}_{n(1+K)\times 1}. \tag{6.48}$$

Note that there are $n(1+K)$ moment conditions and $nK + K$ parameters (nK in β and K in λ), so there are $n - K$ overidentifying restrictions (as before).

Example 6.17 *(Two assets and one factor) The moment conditions (6.48) are*

$$\mathrm{E}\, g_t(\beta_1, \beta_2, \lambda) = \mathrm{E}\begin{bmatrix} R_{1t}^e - \beta_1(\lambda - \mathrm{E}\, f_t) - \beta_1 f_t \\ R_{2t}^e - \beta_2(\lambda - \mathrm{E}\, f_t) - \beta_2 f_t \\ f_t[R_{1t}^e - \beta_1(\lambda - \mathrm{E}\, f_t) - \beta_1 f_t] \\ f_t[R_{2t}^e - \beta_2(\lambda - \mathrm{E}\, f_t) - \beta_2 f_t] \end{bmatrix} = \mathbf{0}_{4\times 1}.$$

This gives 4 moment conditions, but only three parameters, so there is one overidentifying restriction to test—just as with (6.44).

6.5.4 What If the Factors Are Excess Returns?

It would (perhaps) be natural if the tests discussed in this section coincided with those in Section 6.4 when the factors are in fact excess returns. That is *almost* so. The difference is that we here estimate the $K \times 1$ vector λ (factor risk premia) as a vector of free parameters,

while the tests in Section 6.4 *impose* $\lambda = \mathrm{E}\, f_t$. This can be done in (6.44)–(6.45) by doing two things. First, define a new set of test assets by stacking the original test assets and the excess return factors

$$\tilde{R}_t^e = \begin{bmatrix} R_t^e \\ f_t \end{bmatrix}, \tag{6.49}$$

which is an $(n + K) \times 1$ vector. Second, define the $K \times (n + K)$ matrix θ as

$$\tilde{\theta} = \begin{bmatrix} \mathbf{0}_{K\times n} & I_K \end{bmatrix}. \tag{6.50}$$

Together, this gives

$$\lambda = \mathrm{E}\, f_t. \tag{6.51}$$

It is also straightforward to show that this gives precisely the same test statistics as the Wald test on the multifactor model (6.30).

Proof. (of (6.51)) The betas of the $\tilde{R}_t^e$ vector are

$$\tilde{\beta} = \begin{bmatrix} \beta_{n\times K} \\ I_K \end{bmatrix}.$$

The expression corresponding to $\theta\, \mathrm{E}(R_t^e - \beta\lambda) = \mathbf{0}$ is then

$$\begin{bmatrix} \mathbf{0}_{K\times n} & I_K \end{bmatrix} \mathrm{E} \begin{bmatrix} R_t^e \\ f_t \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{K\times n} & I_K \end{bmatrix} \begin{bmatrix} \beta_{n\times K} \\ I_K \end{bmatrix} \lambda, \text{ or}$$
$$\mathrm{E}\, f_t = \lambda.$$

■

Remark 6.18 *(Two assets, one excess return factor) By including the factors among the test assets and using the weighting vector* $\theta = [0, 0, 1]$ *gives*

$$A\, \mathrm{E}\, g_t(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \lambda) = \mathrm{E} \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t - \alpha_3 - \beta_3 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ f_t(f_t - \alpha_3 - \beta_3 f_t) \\ 0(R_{1t}^e - \beta_1\lambda) + 0(R_{2t}^e - \beta_2\lambda) + 1(f_t - \beta_3\lambda) \end{bmatrix} = \mathbf{0}_{7\times 1}.$$

Since $\alpha_3 = 0$ and $\beta_3 = 1$, this gives the estimate $\lambda = \mathrm{E}\, f_t$. There are 7 moment conditions and as many parameters. To test the asset pricing model, test if the following moment conditions are satisfied at the estimated parameters

$$\mathrm{E}\, g_t(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \lambda) = \mathrm{E} \begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t - \alpha_3 - \beta_3 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \\ f_t(f_t - \alpha_3 - \beta_3 f_t) \\ R^e_{1t} - \beta_1 \lambda \\ R^e_{2t} - \beta_2 \lambda \\ f_t - \beta_3 \lambda \end{bmatrix} = \mathbf{0}_{9\times 1}.$$

In fact, this gives the same test statistic as when testing if α_1 and α_2 are zero in (6.18).

6.5.5 When Some (but Not All) of the Factors Are Excess Returns*

Partition the vector of factors as

$$f_t = \begin{bmatrix} Z_t \\ F_t \end{bmatrix}, \tag{6.52}$$

where Z_t is an $v \times 1$ vector of excess return factors and F_t is a $w \times 1$ vector of general factors ($K = v + w$).

It makes sense (and is econometrically efficient) to use the fact that the factor risk premia of the excess return factors are just their average excess returns (as in CAPM). This can be done in (6.44)–(6.45) by doing two things. First, define a new set of test assets by stacking the original test assets and the excess return factors

$$\tilde{R}^e_t = \begin{bmatrix} R^e_t \\ Z_t \end{bmatrix}, \tag{6.53}$$

which is an $(n + v) \times 1$ vector. Second, define the $K \times (n + K)$ matrix θ

$$\tilde{\theta} = \begin{bmatrix} \mathbf{0}_{v\times n} & I_v \\ \vartheta_{w\times n} & \mathbf{0}_{w\times v} \end{bmatrix}, \tag{6.54}$$

where ϑ is some $w \times n$ matrix. Together, this ensures that

$$\tilde{\lambda} = \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} = \begin{bmatrix} \mathrm{E}\, Z_t \\ (\vartheta\beta^F)^{-1}\vartheta(\mathrm{E}\, R_t^e - \beta^Z\lambda_Z) \end{bmatrix}, \tag{6.55}$$

where the β^Z and β^F are just betas of the original test assets on Z_t and F_t respectively—according to the partitioning

$$\beta_{n\times K} = \begin{bmatrix} \beta^Z_{n\times v} & \beta^F_{n\times w} \end{bmatrix}. \tag{6.56}$$

One possible choice of ϑ is $\vartheta = \beta^{F\prime}$, since then λ_F are the same as when running a cross-sectional regression of the expected "abnormal return" $(\mathrm{E}\, R_t^e - \beta^Z\lambda_Z)$ on the betas (β^F).

Proof. (of (6.55)) The betas of the $\tilde{R}_t^e$ vector are

$$\tilde{\beta} = \begin{bmatrix} \beta^Z_{n\times v} & \beta^F_{n\times w} \\ I_v & 0_{v\times w} \end{bmatrix}.$$

The expression corresponding to $\theta\,\mathrm{E}(R_t^e - \beta\lambda) = \mathbf{0}$ is then

$$\begin{aligned} \tilde{\theta}\,\mathrm{E}\,\tilde{R}_t^e &= \tilde{\theta}\tilde{\beta}\tilde{\lambda} \\ \begin{bmatrix} \mathbf{0}_{v\times n} & I_v \\ \vartheta_{w\times n} & \mathbf{0}_{w\times v} \end{bmatrix}\begin{bmatrix} \mathrm{E}\, R_t^e \\ \mathrm{E}\, Z_t \end{bmatrix} &= \begin{bmatrix} \mathbf{0}_{v\times n} & I_v \\ \vartheta_{w\times n} & \mathbf{0}_{w\times v} \end{bmatrix}\begin{bmatrix} \beta^Z_{n\times v} & \beta^F_{n\times w} \\ I_v & 0_{v\times w} \end{bmatrix}\begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} \\ \begin{bmatrix} \mathrm{E}\, Z_t \\ \vartheta_{w\times n}\,\mathrm{E}\, R_t^e \end{bmatrix} &= \begin{bmatrix} I_v & \mathbf{0}_{v\times w} \\ \vartheta_{w\times n}\beta^Z_{n\times v} & \vartheta_{w\times n}\beta^F_{n\times w} \end{bmatrix}\begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix}. \end{aligned}$$

The first v equations give

$$\lambda_Z = \mathrm{E}\, Z_t.$$

The remaining w equations give

$$\begin{aligned} \vartheta\,\mathrm{E}\, R_t^e &= \vartheta\beta^Z\lambda_Z + \vartheta\beta^F\lambda_F, \text{ so} \\ \lambda_F &= (\vartheta\beta^F)^{-1}\vartheta(\mathrm{E}\, R_t^e - \beta^Z\lambda_Z). \end{aligned}$$

■

Example 6.19 *(Structure of θ to identify λ for excess return factors) Continue Example 6.16 (where there are 2 factors and three test assets) and assume that $Z_t = R_{3t}^e$—so the*

first factor is really an excess return—which we have appended last to set of test assets. Then $\beta_{31} = 1$ *and* $\beta_{32} = 0$ *(regressing* Z_t *on* Z_t *and* F_t *gives the slope coefficients* 1 *and* $\dot{0}$*.) If we set* $(\theta_{11}, \theta_{12}, \theta_{13}) = (0, 0, 1)$*, then the moment conditions in Example 6.16 can be written*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \left(\begin{bmatrix} \mathrm{E}\, R_{1t}^e \\ \mathrm{E}\, R_{2t}^e \\ \mathrm{E}\, Z_t \end{bmatrix} - \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} \right).$$

The first line reads

$$0 = \mathrm{E}\, Z_t - \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix}, \text{ so } \lambda_Z = \mathrm{E}\, Z_t.$$

6.5.6 Empirical Evidence

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. Breeden, Gibbons, and Litzenberger (1989) and Lettau and Ludvigson (2001) estimate models where consumption growth is the factor—with mixed results.

6.6 Linear SDF Models

This section discusses how we can estimate and test the asset pricing equation

$$\mathrm{E}\, p_{t-1} = \mathrm{E}\, x_t m_t, \tag{6.57}$$

where x_t are the "payoffs" and p_{t-1} the "prices" of the assets. We can either interpret p_{t-1} as actual asset prices and x_t as the payoffs, or we can set $p_{t-1} = 1$ and let x_t be gross returns, or set $p_{t-1} = 0$ and x_t be excess returns.

Assume that the SDF is linear in the factors

$$m_t = \gamma' f_t, \tag{6.58}$$

where the $(1 + K) \times 1$ vector f_t contains a constant and the other factors. Combining

with (6.57) gives the sample moment conditions

$$\bar{g}(\gamma) = \sum_{t=1}^{T} g_t(\gamma)/T = \mathbf{0}_{n\times 1}, \text{ where} \tag{6.59}$$

$$g_t = x_t m_t - p_{t-1} = x_t f_t'\gamma - p_{t-1}. \tag{6.60}$$

There are $1 + K$ parameters and n moment conditions (the number of assets).

To estimate this model with a weighting matrix W, we minimize the loss function

$$J = \bar{g}(\gamma)' W \bar{g}(\gamma). \tag{6.61}$$

Alternatively, the moment conditions are combined into $1 + K$ effective conditions as

$$A_{(1+K)\times n}\bar{g}(\gamma) = \mathbf{0}_{(1+K)\times 1}. \tag{6.62}$$

See Appendix B.2 for details on how to calculate the estimates.

To test the asset pricing implications, we test if the moment conditions $\mathrm{E}\, g_t = \mathbf{0}$ are satisfied at the estimated parameters. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used.

This approach estimates all the parameters of the SDF freely. In particular, the mean of the SDF is estimated along with the other parameters. Nothing guarantees that the reciprocal of this mean is anywhere close to a reasonable proxy of a riskfree rate. This may have a large effect on the test of the asset pricing model: think of testing CAPM by using a very strange riskfree rate. (This is discussed in some detail in Dahlquist and Söderlind (1999).)

6.6.1 Restricting the Mean SDF

The model (6.57) does not put any restrictions on the riskfree rate, which may influence the test. The approach above is also incapable of handling the case when all payoffs are excess returns. The reason is that there is nothing to tie down the mean of the SDF. To demonstrate this, the model of the SDF (6.57) is here rewritten as

$$m_t = \bar{m} + b'(f_t - \mathrm{E}\, f_t), \tag{6.63}$$

so $\bar{m} = \mathrm{E}\, m$.

Remark 6.20 *(The SDF model (6.63) combined with excess returns) With excess returns, $x_t = R_t^e$ and $p_{t-1} = \mathbf{0}$. The asset pricing equation is then*

$$\mathbf{0} = \mathrm{E}(m_t R_t^e) = \mathrm{E}\, R_t^e \bar{m} + \mathrm{E}\, R_t^e (f_t - \mathrm{E}\, f_t)' b,$$

which would be satisfied by $(\bar{m}, b) = (0, \mathbf{0})$, which makes no sense.

To handle excess returns, we could add moment conditions for some gross returns (a "riskfree" return might be a good choice) or prices. Alternatively, we could restrict the mean of the SDF. The analysis below considers the latter.

The sample moment conditions for $\mathrm{E}\, x_t m_t = \mathrm{E}\, p_{t-1}$ with the SDF (6.63) are

$$\begin{aligned} \bar{g}(\gamma) &= \mathbf{0}_{n\times 1}, \text{ where} \\ g_t &= x_t m_t - p_{t-1} = x_t \bar{m} + x_t (f_t - \mathrm{E}\, f_t)' b - p_{t-1}, \end{aligned} \tag{6.64}$$

where $\bar{m}$ is given (our restriction). See Appendix B.2 for details on how to calculate the estimates.

Provided we choose $\bar{m} \neq 0$, this formulation works with payoffs, gross returns and also excess returns. It is straightforward to show that the choice of $\bar{m}$ does not matter for the test based on excess returns ($p = 0$, so $\Sigma_p = 0$).

6.6.2 SDF Models versus Linear Factor Models: The Tests

Reference: Ferson (1995); Jagannathan and Wang (2002) (theoretical results); Cochrane (2005) 15 (empirical comparison); Bekaert and Urias (1996); and Söderlind (1999)

The test of the linear factor model and the test of the linear SDF model are (generally) not the same: they test the same implications of the models, but in slightly different ways. The moment conditions look a bit different—and combined with non-parametric methods for estimating the covariance matrix of the sample moment conditions, the two methods can give different results (in small samples, at least). Asymptotically, they are always the same, as showed by Jagannathan and Wang (2002).

There is one case where we know that the tests of the linear factor model and the SDF model are identical: when the factors are excess returns and the SDF is constructed to price these factors as well. To demonstrate this, let R_{1t}^e be a vector of excess returns

on some benchmarks assets. Construct a stochastic discount factor as in Hansen and Jagannathan (1991):

$$m_t = \bar{m} + (R^e_{1t} - \bar{R}^e_{1t})'\theta, \tag{6.65}$$

where $\bar{m}$ is a constant and θ is chosen to make m_t "price" R^e_{1t} in the sample, that is, so

$$\Sigma_{t=1}^T \mathrm{E}\, R^e_{1t} m_t / T = \mathbf{0}. \tag{6.66}$$

Consider the test assets with excess returns R^e_{2t}, and "SDF performance"

$$\bar{g}_{2t} = \frac{1}{T}\sum_{t=1}^T R^e_{2t} m_t. \tag{6.67}$$

Let the factor portfolio model be the linear regression

$$R^e_{2t} = \alpha + \beta R^e_{1t} + \varepsilon_t, \tag{6.68}$$

where $\mathrm{E}\,\varepsilon_t = \mathbf{0}$ and $\mathrm{Cov}(R^e_{1t}, \varepsilon_t) = \mathbf{0}$. Then, the SDF-performance ("pricing error") is proportional to a traditional alpha

$$\bar{g}_{2t}/\bar{m} = \hat{\alpha}. \tag{6.69}$$

In both cases we are thus testing if α is zero or not.

Notice that (6.69) allows for the possibility that R^e_{1t} is the excess return on *dynamic portfolios*, $R^e_{1t} = s_{t-1} \otimes R^e_{0t}$, where s_{t-1} are some information variables (not payoffs as before), for instance, lagged returns or market volatility, and R^e_{0t} are some basic benchmarks (S&P500 and bond, perhaps). The reason is that if R^e_{0t} are excess returns, so are $R^e_{1t} = s_{t-1} \otimes R^e_{0t}$. Therefore, the typical cross-sectional test (of $\mathrm{E}\, R^e = \beta'\lambda$) coincides with the test of the alpha—and also of zero SDF pricing errors.

Notice also that R^e_{2t} could be the excess return on dynamic strategies in terms of the test assets, $R^e_{2t} = z_{t-1} \otimes R^e_{pt}$, where z_{t-1} are information variables and R^e_{pt} are basic test assets (mutual funds say). In this case, we are testing the performance of these dynamic strategies (in terms of mutual funds, say). For instance, suppose R_{1t} is a scalar and the α for $z_{t-1}R_{1t}$ is positive. This would mean that a strategy that goes long in R_{1t} when z_{t-1} is high (and vice versa) has a positive performance.

Proof. (of (6.69)) (Here written in terms of population moments, to simplify the notation.) It follows directly that $\theta = -\mathrm{Var}(R^e_{1t})^{-1}\left(\mathrm{E}\, R^e_{1t}\bar{m}\right)$. Using this and the expression

for m_t in (6.67) gives

$$\mathrm{E}\, g_{2t} = \mathrm{E}\, R^e_{2t}\bar{m} - \mathrm{Cov}\left(R^e_{2t}, R^e_{1t}\right)\mathrm{Var}(R^e_{1t})^{-1}\,\mathrm{E}\, R^e_{1t}\bar{m}.$$

We now rewrite this equation in terms of the parameters in the factor portfolio model (6.68). The latter implies $\mathrm{E}\, R^e_{2t} = \alpha + \beta\,\mathrm{E}\, R^e_{1t}$, and the least squares estimator of the slope coefficients is $\beta = \mathrm{Cov}\left(R^e_{2t}, R^e_{1t}\right)\mathrm{Var}\left(R^e_{1t}\right)^{-1}$. Using these two facts in the equation above—and replacing population moments with sample moments, gives (6.69). ■

6.7 Conditional Factor Models

Reference: Cochrane (2005) 8; Ferson and Schadt (1996)

The simplest way of introducing conditional information is to simply state that the factors are not just the usual market indices or macro economic series: the factors are non-linear functions of them (this is sometimes called "scaled factors" to indicate that we scale the original factors with instruments). For instance, if R^e_{mt} is the return on the market portfolio and z_{t-1} is something else which is thought to be important for asset pricing (use theory), then the factors could be

$$f_{1t} = R^e_{mt} \text{ and } f_{2t} = z_{t-1}R^e_{mt}. \tag{6.70}$$

Since the second factor is not an excess return, the test is done as in (6.41).

An alternative interpretation of this is that we have only one factor, but that the coefficient of the factor is time varying. This is easiest seen by plugging in the factors in the time-series regression part of the moment conditions (6.41), $R^e_{it} = \alpha + \beta f_t + \varepsilon_{it}$,

$$\begin{aligned} R^e_{it} &= \alpha + \beta_1 R^e_{mt} + \beta_2 z_{t-1}R^e_{mt} + \varepsilon_{it} \\ &= \alpha + (\beta_1 + \beta_2 z_{t-1})R^e_{mt} + \varepsilon_{it}. \end{aligned} \tag{6.71}$$

The first line looks like a two factor model with constant coefficients, while the second line looks like a one-factor model with a time-varying coefficient $(\beta_1 + \beta_2 z_{t-1})$. This is clearly just a matter of interpretation, since it is the same model (and is tested in the same way). This model can be estimated and tested as in the case of "general factors"—as $z_{t-1}R^e_{mt}$ is not a traditional excess return.

See Figure 6.14–6.15 for an empirical illustration.

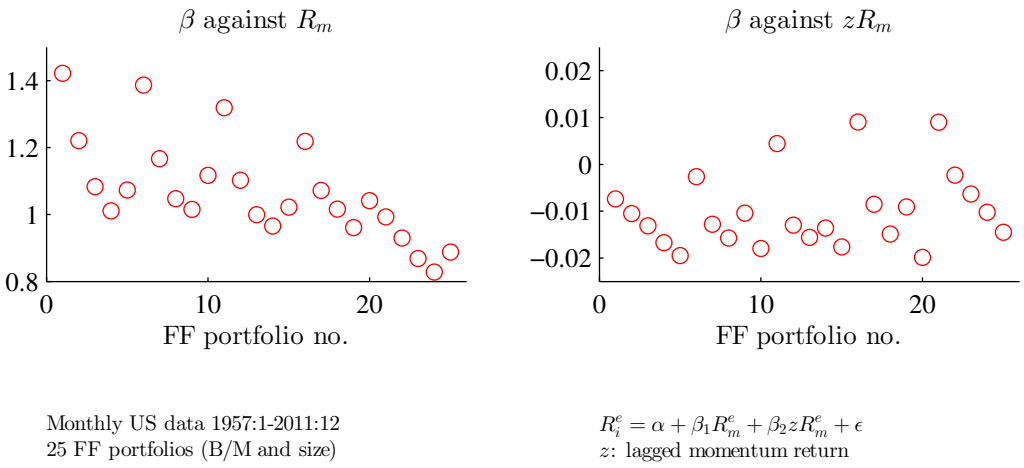


Figure 6.14: Conditional betas of the 25 FF portfolios

Remark 6.21 *(Figures 6.14–6.15, equally weighted 25 FF portfolios) Figure 6.14 shows the betas of the conditional model. It seems as if the small firms (portfolios with low numbers) have a somewhat higher exposure to the market in bull markets and vice versa, while large firms have pretty constant exposures. However, the time-variation is not marked. Therefore, the conditional (two-factor model) fits the cross-section of average returns only slightly better than CAPM—see Figure 6.15.*

Conditional models typically have more parameters than unconditional models, which is likely to give small samples issues (in particular with respect to the inference). It is important to remember some of the new factors (original factors times instruments) are probably not an excess returns, so the test is done with an LM test as in (6.41).

6.8 Conditional Models with "Regimes"

Reference: Christiansen, Ranaldo, and Söderlind (2010)

It is also possible to estimate non-linear factor models. The model could be piecewise linear or include higher order times. For instance, Treynor and Mazuy (1966) extends the CAPM regression by including a squared term (of the market excess return) to capture market timing.

Alternatively, the conditional model (6.71) could be changed so that the time-varying

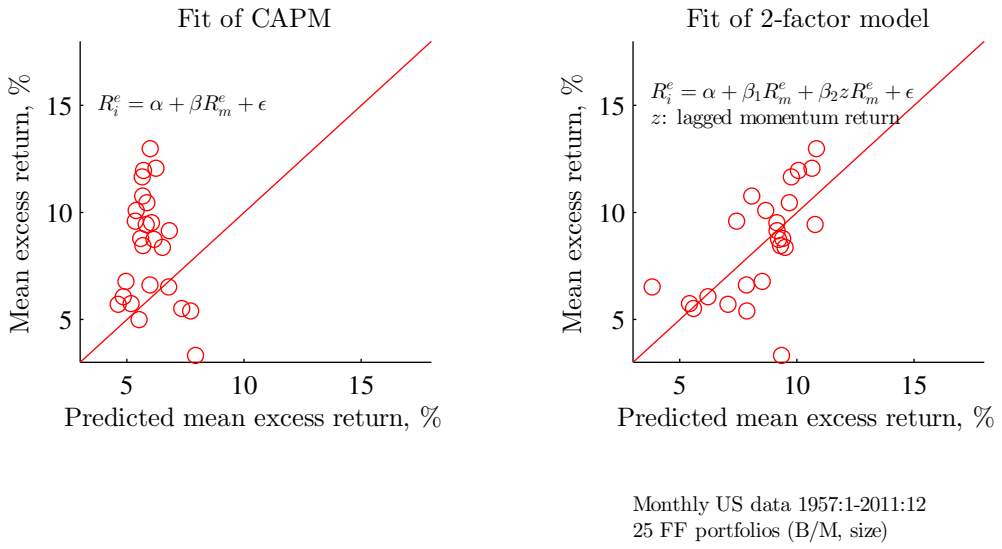


Figure 6.15: Unconditional and conditional CAPM tests of the 25 FF portfolios

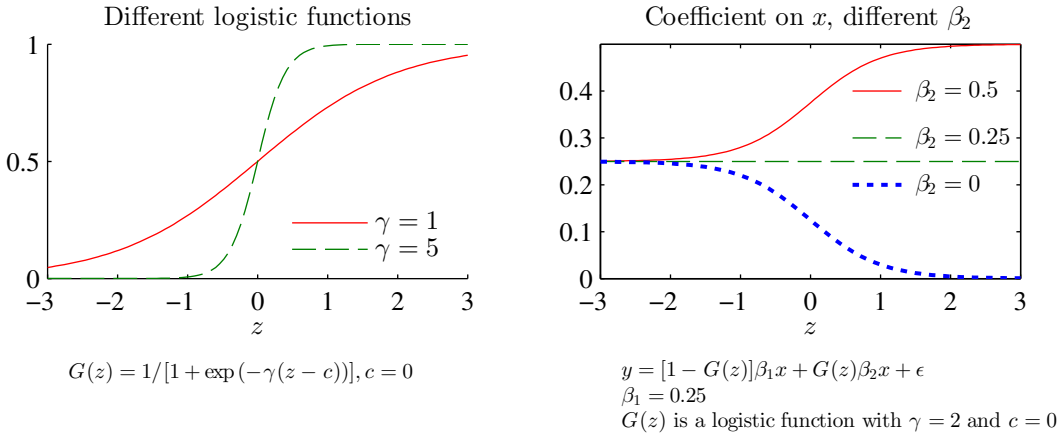


Figure 6.16: Logistic function and the effective slope coefficient in a Logistic smooth transition regression

coefficients are non-linear in the information variable. In the simplest case, this could be dummy variable regression where the definition of the regimes is exogenous.

More ambitiously, we could use a smooth transition regression, which estimates both the "abruptness" of the transition between regimes as well as the cutoff point. Let $G(z)$

be a logistic (increasing but "S-shaped") function

$$G(z) = \frac{1}{1 + \exp[-\gamma(z - c)]}, \tag{6.72}$$

where the parameter c is the central location (where $G(z) = 1/2$) and $\gamma > 0$ determines the steepness of the function (a high γ implies that the function goes quickly from 0 to 1 around $z = c$.) See Figure 6.16 for an illustration. A logistic smooth transition regression is

$$\begin{aligned} y_t &= \left\{[1 - G(z_t)]\,\beta_1' + G(z_t)\beta_2'\right\} x_t + \varepsilon_t \\ &= [1 - G(z_t)]\,\beta_1' x_t + G(z_t)\beta_2' x_t + \varepsilon_t. \end{aligned} \tag{6.73}$$

At low z_t values, the regression coefficients are (almost) β_1 and at high z_t values they are (almost) β_2. See Figure 6.16 for an illustration.

Remark 6.22 *(NLS estimation) The parameter vector $(\gamma, c, \beta_1, \beta_2)$ is easily estimated by Non-Linear least squares (NLS) by concentrating the loss function: optimize (numerically) over (γ, c) and let (for each value of (γ, c)) the parameters (β_1, β_2) be the OLS coefficients on the vector of "regressors"* $([1 - G(z_t)]\,x_t, G(z_t)x_t)$.

The most common application of this model is by letting $x_t = y_{t-s}$. This is the LSTAR model—logistic smooth transition auto regression model, see Franses and van Dijk (2000).

For an empirical application to a factor model, see Figures 6.17–6.18.

6.9 Fama-MacBeth*

Reference: Cochrane (2005) 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—although is seems most related to what we discussed in Section 6.5. The method has three steps, described below.

- First, estimate the betas β_i $(i = 1, \ldots, n)$ from (6.1) (this is a time-series regression). This is often done on the whole sample—assuming the betas are constant.

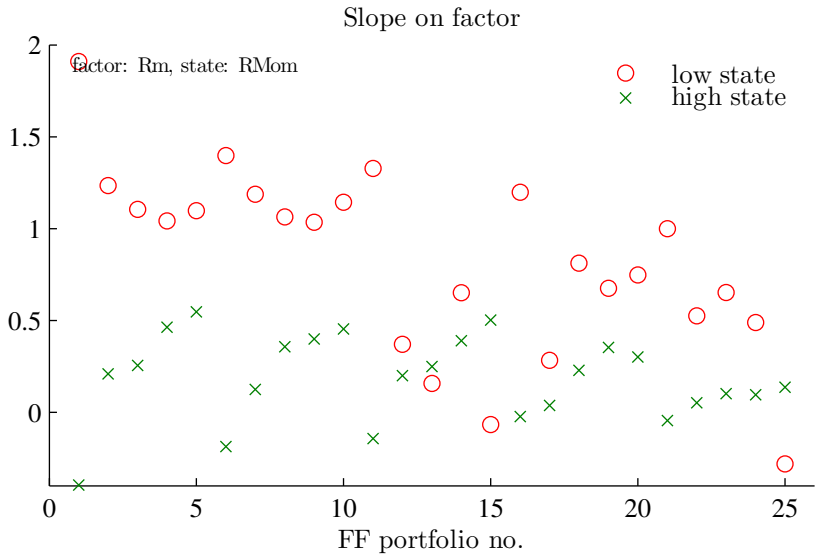


Figure 6.17: Betas on the market in the low and high regimes, 25 FF portfolios

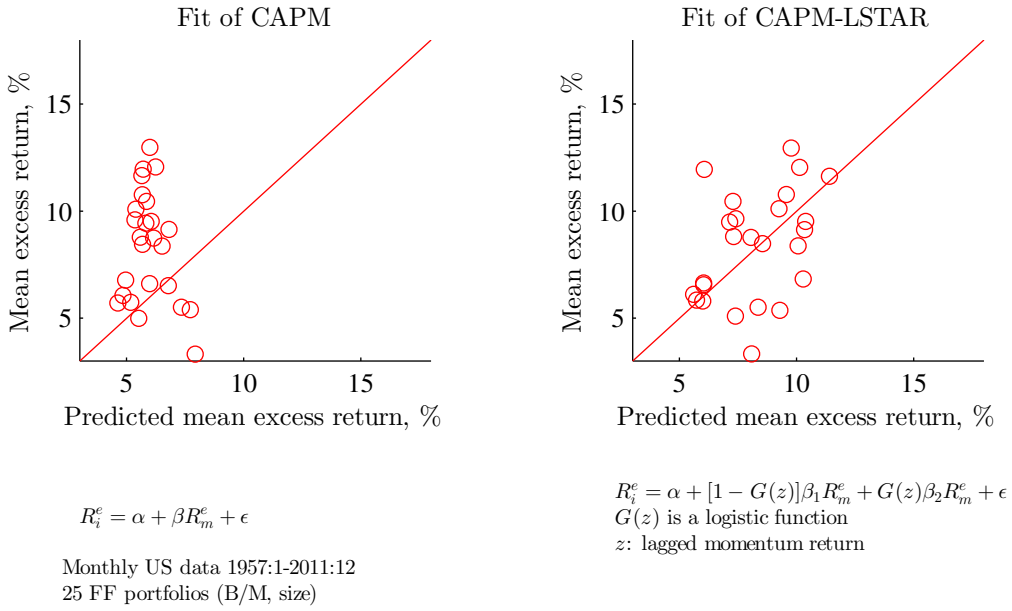


Figure 6.18: Test of 1 and 2-factor models, 25 FF portfolios

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

$$R_{it}^e = \lambda_t' \hat{\beta}_i + \varepsilon_{it}, \tag{6.74}$$

  where $\hat{\beta}_i$ are the regressors. (Note the difference to the traditional cross-sectional approach discussed in (6.14), where the second stage regression regressed E R_{it}^e on $\hat{\beta}_i$, while the Fama-French approach runs one regression for every time period.)

- Third, estimate the time averages

$$\hat{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_{it} \text{ for } i = 1, \ldots, n, \text{ (for every asset)} \tag{6.75}$$

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \hat{\lambda}_t. \tag{6.76}$$

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

We clearly want portfolios which have different betas, or else the second step regression (6.74) 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 adjust the tests. Jagannathan and Wang (1996) and Jagannathan and Wang (1998) discuss the asymptotic distribution of this estimator.

We can test the model by studying if $\varepsilon_i = 0$ (recall from (6.75) that ε_i is the time average of the residual for asset i, ε_{it}), by forming a t-test $\hat{\varepsilon}_i / \text{Std}(\hat{\varepsilon}_i)$. Fama and MacBeth

(1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\varepsilon}_{it}$. In particular, they suggest that the variance of $\hat{\varepsilon}_{it}$ (not $\hat{\varepsilon}_i$) can be estimated by the (average) squared variation around its mean

$$\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2 . \tag{6.77}$$

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,

$$\text{Var}(\hat{\varepsilon}_i) = \frac{1}{T}\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T^2}\sum_{t=1}^{T}(\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2 . \tag{6.78}$$

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

$$\text{Var}(\hat{\lambda}) = \frac{1}{T^2}\sum_{t=1}^{T}(\hat{\lambda}_t - \hat{\lambda})^2. \tag{6.79}$$

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 Details of SURE Systems

Proof. (of (6.8)) Write each of the regression equations in (6.7) on a traditional form

$$R_{it}^e = x_t'\theta_i + \varepsilon_{it}, \text{ where } x_t = \begin{bmatrix} 1 \\ f_t \end{bmatrix}.$$

Define

$$\Sigma_{xx} = \text{plim}\sum_{t=1}^{T} x_t x_t'/T, \text{ and } \sigma_{ij} = \text{plim}\sum_{t=1}^{T} \varepsilon_{it}\varepsilon_{jt}/T,$$

then the asymptotic covariance matrix of the vectors $\hat{\theta}_i$ and $\hat{\theta}_j$ (assets i and j) is $\sigma_{ij}\Sigma_{xx}^{-1}/T$ (see below for a separate proof). In matrix form,

$$\text{Cov}(\sqrt{T}\hat{\theta}) = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \dots & \hat{\sigma}_{nn} \end{bmatrix} \otimes \Sigma_{xx}^{-1},$$

where $\hat{\theta}$ stacks $\hat{\theta}_1, \ldots, \hat{\theta}_n$. As in (6.3), the upper left element of Σ_{xx}^{-1} equals $1 + SR^2$, where SR is the Sharpe ratio of the market. ■

Proof. (of distribution of SUR coefficients, used in proof of (6.8)*) To simplify, consider the SUR system

$$\begin{aligned} y_t &= \beta x_t + u_t \\ z_t &= \gamma x_t + v_t, \end{aligned}$$

where y_t, z_t and x_t are zero mean variables. We then know (from basic properties of LS) that

$$\begin{aligned} \hat{\beta} &= \beta + \frac{1}{\sum_{t=1}^{T} x_t x_t} (x_1 u_1 + x_2 u_2 + \ldots x_T u_T) \\ \hat{\gamma} &= \gamma + \frac{1}{\sum_{t=1}^{T} x_t x_t} (x_1 v_1 + x_2 v_2 + \ldots x_T v_T) . \end{aligned}$$

In the traditional LS approach, we treat x_t as fixed numbers ("constants") and also assume that the residuals are uncorrelated across and have the same variances and covariances across time. The covariance of $\hat{\beta}$ and $\hat{\gamma}$ is therefore

$$\begin{aligned} \text{Cov}(\hat{\beta}, \hat{\gamma}) &= \left(\frac{1}{\sum_{t=1}^{T} x_t x_t} \right)^2 \left[x_1^2 \text{Cov}(u_1, v_1) + x_2^2 \text{Cov}(u_2, v_2) + \ldots x_T^2 \text{Cov}(u_T, v_T) \right] \\ &= \left(\frac{1}{\sum_{t=1}^{T} x_t x_t} \right)^2 \left(\sum_{t=1}^{T} x_t x_t \right) \sigma_{uv}, \text{ where } \sigma_{uv} = \text{Cov}(u_t, v_t) , \\ &= \frac{1}{\sum_{t=1}^{T} x_t x_t} \sigma_{uv}. \end{aligned}$$

Divide and multiply by T to get the result in the proof of (6.8). (We get the same results if we relax the assumption that x_t are fixed numbers, and instead derive the asymptotic distribution.) ■

Remark A.1 *(General results on SURE distribution, same regressors) Let the regression equations be*

$$y_{it} = x_t' \theta_i + \varepsilon_{it}, \; i = 1, \ldots, n,$$

where x_t is a $K \times 1$ vector (the same in all n regressions). When the moment conditions

are arranged so that the first n are $x_{1t}\varepsilon_t$, then next are $x_{2t}\varepsilon_t$

$$\mathrm{E}\, g_t = \mathrm{E}(x_t \otimes \varepsilon_t),$$

then Jacobian (with respect to the coefs of x_{1t}, then the coefs of x_{2t}, etc) and its inverse are

$$D_0 = -\Sigma_{xx} \otimes I_n \text{ and } D_0^{-1} = -\Sigma_{xx}^{-1} \otimes I_n.$$

The covariance matrix of the moment conditions is as usual $S_0 = \sum_{s=-\infty}^{\infty} \mathrm{E}\, g_t g_{t-s}'$. As an example, let $n = 2$, $K = 2$ with $x_t' = (1, f_t)$ and let $\theta_i = (\alpha_i, \beta_i)$, then we have

$$\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \bar{g}_3 \\ \bar{g}_4 \end{bmatrix} = \frac{1}{T} \sum\nolimits_{t=1}^{T} \begin{bmatrix} y_{1t} - \alpha_1 - \beta_1 f_t \\ y_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(y_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(y_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix},$$

and

$$\begin{aligned} \frac{\partial \bar{g}}{\partial [\alpha_1, \alpha_2, \beta_1, \beta_2]'} &= \begin{bmatrix} \partial \bar{g}_1/\partial \alpha_1 & \partial \bar{g}_1/\partial \alpha_2 & \partial \bar{g}_1/\partial \beta_1 & \partial \bar{g}_1/\partial \beta_2 \\ \partial \bar{g}_2/\partial \alpha_1 & \partial \bar{g}_2/\partial \alpha_2 & \partial \bar{g}_2/\partial \beta_1 & \partial \bar{g}_2/\partial \beta_2 \\ \partial \bar{g}_3/\partial \alpha_1 & \partial \bar{g}_3/\partial \alpha_2 & \partial \bar{g}_3/\partial \beta_1 & \partial \bar{g}_3/\partial \beta_2 \\ \partial \bar{g}_4/\partial \alpha_1 & \partial \bar{g}_4/\partial \alpha_2 & \partial \bar{g}_4/\partial \beta_1 & \partial \bar{g}_4/\partial \beta_2 \end{bmatrix} \\ &= -\frac{1}{T} \sum\nolimits_{t=1}^{T} \begin{bmatrix} 1 & 0 & f_t & 0 \\ 0 & 1 & 0 & f_t \\ f_t & 0 & f_t^2 & 0 \\ 0 & f_t & 0 & f_t^2 \end{bmatrix} = \left(-\frac{1}{T} \sum\nolimits_{t=1}^{T} x_t x_t' \right) \otimes I_2. \end{aligned}$$

Remark A.2 *(General results on SURE distribution, same regressors, alternative ordering of moment conditions and parameters*) If instead, the moment conditions are arranged so that the first K are $x_t\varepsilon_{1t}$, the next are $x_t\varepsilon_{2t}$ as in*

$$\mathrm{E}\, g_t = \mathrm{E}(\varepsilon_t \otimes x_t),$$

then the Jacobian (wrt the coffecients in regression 1, then the coeffs in regression 2 etc.) and its inverse are

$$D_0 = I_n \otimes (-\Sigma_{xx}) \text{ and } D_0^{-1} = I_n \otimes (-\Sigma_{xx}^{-1}).$$

Reordering the moment conditions and parameters in Example A.1 gives

$$\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \bar{g}_3 \\ \bar{g}_4 \end{bmatrix} = \frac{1}{T}\sum\nolimits_{t=1}^{T} \begin{bmatrix} y_{1t} - \alpha_1 - \beta_1 f_t \\ f_t(y_{1t} - \alpha_1 - \beta_1 f_t) \\ y_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(y_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix},$$

and

$$\begin{aligned} \frac{\partial \bar{g}}{\partial[\alpha_1, \beta_1, \alpha_2, \beta_2]'} &= \begin{bmatrix} \partial\bar{g}_1/\partial\alpha_1 & \partial\bar{g}_1/\partial\beta_1 & \partial\bar{g}_1/\partial\alpha_2 & \partial\bar{g}_1/\partial\beta_2 \\ \partial\bar{g}_2/\partial\alpha_1 & \partial\bar{g}_2/\partial\beta_1 & \partial\bar{g}_2/\partial\alpha_2 & \partial\bar{g}_2/\partial\beta_2 \\ \partial\bar{g}_3/\partial\alpha_1 & \partial\bar{g}_3/\partial\beta_1 & \partial\bar{g}_3/\partial\alpha_2 & \partial\bar{g}_3/\partial\beta_2 \\ \partial\bar{g}_4/\partial\alpha_1 & \partial\bar{g}_4/\partial\beta_1 & \partial\bar{g}_4/\partial\alpha_2 & \partial\bar{g}_4/\partial\beta_2 \end{bmatrix} \\ &= -\frac{1}{T}\sum\nolimits_{t=1}^{T} \begin{bmatrix} 1 & f_t & 0 & 0 \\ f_t & f_t^2 & 0 & 0 \\ 0 & 0 & 1 & f_t \\ 0 & 0 & f_t & f_t^2 \end{bmatrix} = I_2 \otimes \left(-\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t x_t' \right). \end{aligned}$$

B Calculating GMM Estimator

B.1 Coding of the GMM Estimation of a Linear Factor Model

This section describes how the GMM problem can be programmed. We treat the case with n assets and K Factors (which are all excess returns). The moments are of the form

$$g_t = \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \right)$$

$$g_t = \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \beta f_t) \right)$$

for the exactly identified and overidentified case respectively

Suppose we could write the moments on the form

$$g_t = z_t \left(y_t - x_t' b \right),$$

to make it easy to use matrix algebra in the calculation of the estimate (see below for how

to do that). These moment conditions are similar to those for the instrumental variable method. In that case we could let

$$\Sigma_{zy} = \frac{1}{T}\sum_{t=1}^{T} z_t y_t \text{ and } \Sigma_{zx} = \frac{1}{T}\sum_{t=1}^{T} z_t x_t', \text{ so } \frac{1}{T}\sum_{t=1}^{T} g_t = \Sigma_{zy} - \Sigma_{zx} b.$$

In the exactly identified case, we then have

$$\bar{g}_t = \Sigma_{zy} - \Sigma_{zx} b = \mathbf{0}, \text{ so } \hat{b} = \Sigma_{zx}^{-1}\Sigma_{zy}.$$

(It is straightforward to show that this can also be calculated equation by equation.) In the overidentified case with a weighting matrix, the loss function can be written

$$\begin{aligned} \bar{g}' W \bar{g} &= (\Sigma_{zy} - \Sigma_{zx} b)' W (\Sigma_{zy} - \Sigma_{zx} b), \text{ so} \\ \Sigma_{zx}' W \Sigma_{zy} - \Sigma_{zx}' W \Sigma_{zx} \hat{b} &= \mathbf{0} \text{ and } \hat{b} = (\Sigma_{zx}' W \Sigma_{zx})^{-1} \Sigma_{zx}' W \Sigma_{zy}. \end{aligned}$$

In the overidentified case when we premultiply the moment conditions by A, we get

$$A\bar{g} = A\Sigma_{zy} - A\Sigma_{zx} b = \mathbf{0}, \text{ so } b = (A\Sigma_{zx})^{-1} A \Sigma_{zy}.$$

In practice, we never perform an explicit inversion—it is typically much better (in terms of both speed and precision) to let the software solve the system of linear equations instead.

To rewrite the moment conditions as $g_t = z_t \left(y_t - x_t' b\right)$, notice that

$$g_t = \underbrace{\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes I_n\right)}_{z_t} \left(R_t^e - \underbrace{\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}' \otimes I_n\right)}_{x_t'} b \right), \text{ with } b = \text{vec}(\alpha, \beta)$$

$$g_t = \underbrace{\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes I_n\right)}_{z_t} \left(R_t^e - \underbrace{\left(f_t' \otimes I_n\right)}_{x_t'} b \right), \text{ with } b = \text{vec}(\beta)$$

for the exactly identified and overidentified case respectively. Clearly, z_t and x_t are matrices, not vectors. (z_t is $n(1 + K) \times n$ and x_t' is either of the same dimension or has n rows less, corresponding to the intercept.)

Example B.1 *(Rewriting the moment conditions) For the moment conditions in Example*

6.12 we have

$$g_t(\alpha,\beta) = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ f_{1t} & 0 \\ 0 & f_{1t} \\ f_{2t} & 0 \\ 0 & f_{2t} \end{bmatrix}}_{z_t} \left(\begin{bmatrix} R^e_{1t} \\ R^e_{2t} \end{bmatrix} - \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ f_{1t} & 0 \\ 0 & f_{1t} \\ f_{2t} & 0 \\ 0 & f_{2t} \end{bmatrix}'}_{x_t'} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{21} \\ \beta_{12} \\ \beta_{22} \end{bmatrix} \right).$$

Proof. (of rewriting the moment conditions) From the properties of Kronecker products, we know that *(i)* $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$; and *(ii)* if a is $m \times 1$ and c is $n \times 1$, then $a \otimes c = (a \otimes I_n)c$. The first rule allows to write

$$\alpha + \beta f_t = I_n \begin{bmatrix} \alpha & \beta \end{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \text{ as } \underbrace{\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}' \otimes I_n \right)}_{x_t'} \underbrace{\text{vec}(\begin{bmatrix} \alpha & \beta \end{bmatrix})}_{b}.$$

The second rule allows us two write

$$\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R^e_t - \alpha - \beta f_t) \text{ as } \underbrace{\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes I_n \right)}_{z_t} (R^e_t - \alpha - \beta f_t).$$

(For the exactly identified case, we could also use the fact $(A \otimes B)' = A' \otimes B'$ to notice that $z_t = x_t$.) ■

Remark B.2 *(Quick matrix calculations of Σ_{zx} and Σ_{zy}) Although a loop wouldn't take too long time to calculate Σ_{zx} and Σ_{zy}, there is a quicker way. Put $\begin{bmatrix} 1 & f_t' \end{bmatrix}$ in row t of the matrix $Z_{T\times(1+K)}$ and $R^{e\prime}_t$ in row t of the matrix $R_{T\times n}$. For the exactly identified case, let $X = Z$. For the overidentified case, put f_t' in row t of the matrix $X_{T\times K}$. Then, calculate*

$$\Sigma_{zx} = (Z'X/T) \otimes I_n \text{ and } vec(R'Z/T) = \Sigma_{zy}.$$

B.2 Coding of the GMM Estimation of a Linear SDF Model

B.2.1 No Restrictions on the Mean SDF

To simplify the notation, define

$$\Sigma_{xf} = \sum_{t=1}^{T} x_t f_t'/T \text{ and } \Sigma_p = \sum_{t=1}^{T} p_{t-1}/T.$$

The moment conditions can then be written

$$\bar{g}(\gamma) = \Sigma_{xf}\gamma - \Sigma_p,$$

and the loss function as

$$J = \left(\Sigma_{xf}\gamma - \Sigma_p\right)' W \left(\Sigma_{xf}\gamma - \Sigma_p\right).$$

The first order conditions are

$$\begin{aligned} \mathbf{0}_{(1+K)\times 1} = \frac{\partial J}{\partial \gamma} &= \left(\frac{\partial \bar{g}(\hat{\gamma})}{\partial \gamma'}\right)' W \bar{g}(\hat{\gamma}) \\ &= \Sigma_{xf}' W \left(\Sigma_{xf}\hat{\gamma} - \Sigma_p\right), \text{ so} \\ \hat{\gamma} &= \left(\Sigma_{xf}' W \Sigma_{xf}\right)^{-1} \Sigma_{xf}' W \Sigma_p. \end{aligned}$$

In can also be noticed that the Jacobian is

$$\frac{\partial \bar{g}(\gamma)}{\partial \gamma'} = \Sigma_{xf}.$$

Instead, with $A\bar{g}(\gamma) = \mathbf{0}$, we have

$$\begin{aligned} A\Sigma_{xf}\gamma - A\Sigma_p = \mathbf{0}, \text{ so} \\ \gamma = (A\Sigma_{xf})^{-1} A\Sigma_p. \end{aligned}$$

B.2.2 Restrictions on the Mean SDF

To simplify the notation, let

$$\Sigma_x = \sum_{t=1}^{T} x_t/T,\ \Sigma_{xf} = \sum_{t=1}^{T} x_t(f_t - \mathrm{E}\, f_t)'/T \text{ and } \Sigma_p = \sum_{t=1}^{T} p_{t-1}/T.$$

The moment conditions are

$$\bar{g}(b) = \Sigma_x \bar{m} + \Sigma_{xf} b - \Sigma_p$$

With a weighting matrix W, we minimize

$$J = \left(\Sigma_x \bar{m} + \Sigma_{xf} b - \Sigma_p\right)' W \left(\Sigma_x \bar{m} + \Sigma_{xf} b - \Sigma_p\right).$$

The first order conditions (with respect to b only, since $\bar{m}$ is given) are

$$\begin{aligned} \mathbf{0}_{K\times 1} &= \Sigma_{xf}' W \left(\Sigma_x \bar{m} + \Sigma_{xf} \hat{b} - \Sigma_p\right), \text{ so} \\ \hat{b} &= \left(\Sigma_{xf}' W \Sigma_{xf}\right)^{-1} \Sigma_{xf}' W \left(\Sigma_p - \Sigma_x \bar{m}\right). \end{aligned}$$

Instead, with $A\bar{g}(\gamma) = \mathbf{0}$, we have

$$\begin{aligned} A\Sigma_x \bar{m} + A\Sigma_{xf} b - A\Sigma_p = \mathbf{0}, \text{ so} \\ b = (A\Sigma_{xf})^{-1} A \left(\Sigma_p - \Sigma_x \bar{m}\right). \end{aligned}$$

Bibliography

Bekaert, G., and M. S. Urias, 1996, "Diversification, integration and emerging market closed-end funds," *Journal of Finance*, 51, 835–869.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger, 1989, "Empirical tests of the consumption-oriented CAPM," *Journal of Finance*, 44, 231–262.

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.

Christiansen, C., A. Ranaldo, and P. Söderlind, 2010, "The time-varying systematic risk of carry trade strategies," *Journal of Financial and Quantitative Analysis,* forthcoming.

Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.

Dahlquist, M., and P. Söderlind, 1999, "Evaluating portfolio performance with stochastic discount factors," *Journal of Business*, 72, 347–383.

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.

Ferson, W. E., 1995, "Theory and empirical testing of asset pricing models," in Robert A. Jarrow, Vojislav Maksimovic, and William T. Ziemba (ed.), *Handbooks in Operations Research and Management Science* . pp. 145–200, North-Holland, Amsterdam.

Ferson, W. E., and R. Schadt, 1996, "Measuring fund strategy and performance in changing economic conditions," *Journal of Finance*, 51, 425–461.

Franses, P. H., and D. van Dijk, 2000, *Non-linear time series models in empirical finance*, Cambridge University Press.

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

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

Hansen, L. P., and R. Jagannathan, 1991, "Implications of security market data for models of dynamic economies," *Journal of Political Economy*, 99, 225–262.

Jagannathan, R., and Z. Wang, 1996, "The conditional CAPM and the cross-section of expectd returns," *Journal of Finance*, 51, 3–53.

Jagannathan, R., and Z. Wang, 1998, "A note on the asymptotic covariance in Fama-MacBeth regression," *Journal of Finance*, 53, 799–801.

Jagannathan, R., and Z. Wang, 2002, "Empirical evaluation of asset pricing models: a comparison of the SDF and beta methods," *Journal of Finance*, 57, 2337–2367.

Lettau, M., and S. Ludvigson, 2001, "Resurrecting the (C)CAPM: a cross-sectional test when risk premia are time-varying," *Journal of Political Economy*, 109, 1238–1287.

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

Söderlind, P., 1999, "An interpretation of SDF based performance measures," *European Finance Review*, 3, 233–237.

Treynor, J. L., and K. Mazuy, 1966, "Can Mutual Funds Outguess the Market?," *Harvard Business Review*, 44, 131–136.

7 Consumption-Based Asset Pricing

Reference: Bossaert (2002); Campbell (2003); Cochrane (2005); Smith and Wickens (2002)

7.1 Consumption-Based Asset Pricing

7.1.1 The Basic Asset Pricing Equation

The basic asset pricing equation says

$$\mathrm{E}_{t-1}\, R_t M_t = 1. \tag{7.1}$$

where R_t is the gross return of holding an asset from period $t-1$ to t, M_t is a stochastic discount factor (SDF). E_{t-1} denotes the expectations conditional on the information in period $t-1$, that is, when the investment decision is made. This equation holds for any assets that are freely traded without transaction costs (or taxes), even if markets are incomplete.

In a consumption-based model, (7.1) is the Euler equation for optimal saving in $t-1$ where M_t is the ratio of marginal utilities in t and $t-1$, $M_t = \beta u'(C_t)/u'(C_{t-1})$. I will focus on the case where the marginal utility of consumption is a function of consumption only, which is by far the most common formulation. This allows for other terms in the utility function, for instance, leisure and real money balances, but they have to be additively separable from the consumption term. With constant relative risk aversion (CRRA) γ, the stochastic discount factor is

$$M_t = \beta(C_t/C_{t-1})^{-\gamma}, \text{ so} \tag{7.2}$$

$$\ln M_t = \ln\beta - \gamma\Delta c_t, \text{ where } \Delta c_t = \ln C_t/C_{t-1}. \tag{7.3}$$

The second line is only there to introduce the convenient notation Δc_t for the consumption growth rate.

The next few sections study if the pricing model consisting of (7.1) and (7.2) can fit

historical data. To be clear about what this entails, note the following. First, general equilibrium considerations will not play any role in the analysis: the production side will not be even mentioned. Instead, the focus is on one of the building blocks of an otherwise unspecified model. Second, complete markets are not assumed. The key assumption is rather that the basic asset pricing equation (7.1) holds for the assets I analyse. This means that the representative investor can trade in these assets without transaction costs and taxes (clearly an approximation). Third, the properties of historical (ex post) data are assumed to be good approximations of what investors expected. In practice, this assumes both rational expectations and that the sample is large enough for the estimators (of various moments) to be precise.

To highlight the basic problem with the consumption-based model and to simplify the exposition, I assume that the excess return, R_t^e, and consumption growth, Δc_t, have a bivariate normal distribution. By using Stein's lemma, we can write the the risk premium as

$$\mathrm{E}_{t-1}\, R_t^e = \mathrm{Cov}_{t-1}(R_t^e, \Delta c_t)\gamma. \tag{7.4}$$

The intuition for this expressions is that an asset that has a high payoff when consumption is high, that is, when marginal utility is low, is considered risky and will require a risk premium. This expression also holds in terms of unconditional moments. (To derive that, start by taking unconditional expectations of (7.1).)

We can relax the assumption that the excess return is normally distributed: (7.4) holds also if R_t^e and Δc_t have a bivariate mixture normal distribution—provided Δc_t has the same mean and variance in all the mixture components (see Section 7.1.1 below). This restricts consumption growth to have a normal distribution, but allows the excess return to have a distribution with fat tails and skewness.

Remark 7.1 *(Stein's lemma) If x and y have a bivariate normal distribution and $h(y)$ is a differentiable function such that* $\mathrm{E}[|h'(y)|] < \infty$*, then* $\mathrm{Cov}[x, h(y)] = \mathrm{Cov}(x, y)\,\mathrm{E}[h'(y)]$.

Proof. (of (7.4)) For an excess return R^e, (7.1) says $\mathrm{E}\, R^e M = 0$, so

$$\mathrm{E}\, R^e = -\mathrm{Cov}(R^e, M)/\,\mathrm{E}\, M.$$

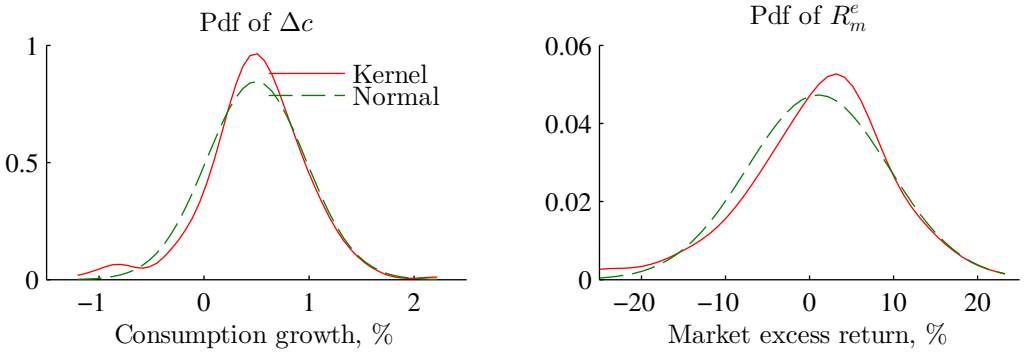
Stein's lemma gives $\mathrm{Cov}[R^e, \exp(\ln M)] = \mathrm{Cov}(R^e, \ln M)\,\mathrm{E}\, M$. (In terms of Stein's lemma, $x = R^e$, $y = \ln M$ and $h() = \exp()$.) Finally, notice that $\mathrm{Cov}(R^e, \ln M) = -\gamma\,\mathrm{Cov}(R^e, \Delta c)$. ■

The Gains and Losses from Using Stein's Lemma

The gain from using (the extended) Stein's lemma is that the unknown relative risk aversion, γ, does not enter the covariances. This facilitates the empirical analysis considerably. Otherwise, the relevant covariance would be between R_t^e and $(C_t/C_{t-1})^{-\gamma}$.

The price of using (the extended) Stein's lemma is that we have to assume that consumption growth is normally distributed and that the excess return have a mixture normal distribution. The latter is not much of a price, since a mixture normal can take many shapes and have both skewness and excess kurtosis.

In any case, *Figure 7.1* suggests that these assumptions might be reasonable. The upper panel shows unconditional distributions of the growth of US real consumption per capita of nondurable goods and services and of the real excess return on a broad US equity index. The non-parametric kernel density estimate of consumption growth is quite similar to a normal distribution, but this is not the case for the US market excess return which has a lot more skewness.



US quaterly data 1957Q1-2008Q4

Figure 7.1: **Density functions of consumption growth and equity market excess returns.** The kernel density function of a variable x is estimated by using a $N(0,\sigma)$ kernel with $\sigma = 1.06\,\text{Std}(x)T^{-1/5}$. The normal distribution is calculated from the estimated mean and variance of the same variable.

An Extended Stein's Lemma for Asset Pricing*

To allow for a non-normal distribution of the asset return, an extension of Stein's lemma is necessary. The following proposition shows that this is possible—if we restrict the

distribution of the log SDF to be gaussian.

Figure 7.2 gives an illustration.

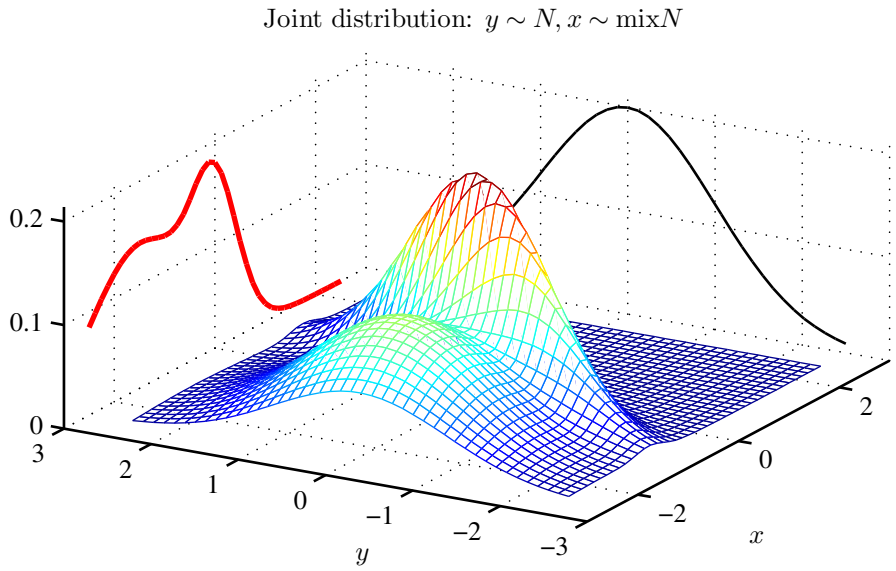


Figure 7.2: **Example of a bivariate mixed-normal distribution** The marginal distributions are drawn at the back.

Proposition 7.2 *Assume* (a) *the joint distribution of* x *and* y *is a mixture of* n *bivariate normal distributions;* (b) *the mean and variance of* y *is the same in each of the* n *components;* (c) $h(y)$ *is a differentiable function such that* $\mathrm{E}\,|h'(y)| < \infty$. *Then* $\mathrm{Cov}[x, h(y)] = \mathrm{E}\,h'(y)\,\mathrm{Cov}(x, y)$. *(See Söderlind (2009) for a proof.)*

7.2 Asset Pricing Puzzles

7.2.1 The Equity Premium Puzzle

This section studies if the consumption-based asset pricing model can explain the historical risk premium on the US stock market.

To discuss the historical average excess returns, it is convenient to work with the unconditional version of the pricing expression (7.4)

$$\mathrm{E}\,R_t^e = \mathrm{Cov}(R_t^e, \Delta c_t)\gamma. \tag{7.5}$$

Table 7.1 shows the key statistics for quarterly US real returns and consumption growth.

	Mean	Std	Autocorr	Corr with Δc
Δc	1.984	0.944	0.362	1.000
R_m^e	5.369	16.899	0.061	0.211
Riskfree	1.213	2.429	0.642	0.196

Table 7.1: US quarterly data, 1957Q1-2008Q4 , (annualized, in %, in real terms)

We see, among other things, that consumption has a standard deviation of only 1% (annualized), the stock market has had an average excess return (over a T-bill) of 6–8% (annualized), and that returns are only weakly correlated with consumption growth. These figures will be important in the following sections. Two correlations with consumption growth are shown, since it is unclear if returns should be related to what is recorded as consumption this quarter or the next. The reason is that consumption is measured as a flow during the quarter, while returns are measured at the end of the quarter.

Table 7.1 shows that we can write (7.5) as

$$\mathrm{E}\, R_t^e = \mathrm{Corr}(R_t^e, \Delta c_t) \times \mathrm{Std}(R_t^e) \times \mathrm{Std}(\Delta c_t)\gamma \tag{7.6}$$

$$0.06 \approx 0.15 \times 0.17 \times 0.01\gamma. \tag{7.7}$$

which requires a value of $\gamma \approx 236$ for the equation to fit.

The basic problem with the consumption-based asset pricing model is that investors enjoy a fairly stable consumption series (either because income is smooth or because it is easy/inexpensive to smooth consumption by changing savings), so only an extreme risk aversion can motivate why investors require such a high equity premium. This is the *equity premium puzzle* stressed by Mehra and Prescott (1985) (although they approach the issue from another angle). Indeed, even if the correlation was one, (7.7) would require $\gamma \approx 35$.

7.2.2 The Equity Premium Puzzle over Time

In contrast to the traditional interpretation of "efficient markets," it has been found that excess returns might be somewhat predictable—at least in the long run (a couple of years). In particular, Fama and French (1988a) and Fama and French (1988b) have argued that

future long-run returns can be predicted by the current dividend-price ratio and/or current returns.

Figure 7.3 illustrates this by showing results the regressions

$$R^e_{t+k}(k) = a_0 + a_1 x_t + u_{t+k}, \text{ where } x_t = E_t/P_t \text{ or } R^e_t(k), \tag{7.8}$$

where $R^e_t(k)$ is the annualized k-quarter excess return of the aggregate US stock market and E_t/P_t is the earnings-price ratio.

It seems as if the earnings-price ratio has some explanatory power for future returns—at least for long horizons. In contrast, the lagged return is a fairly weak predictor.

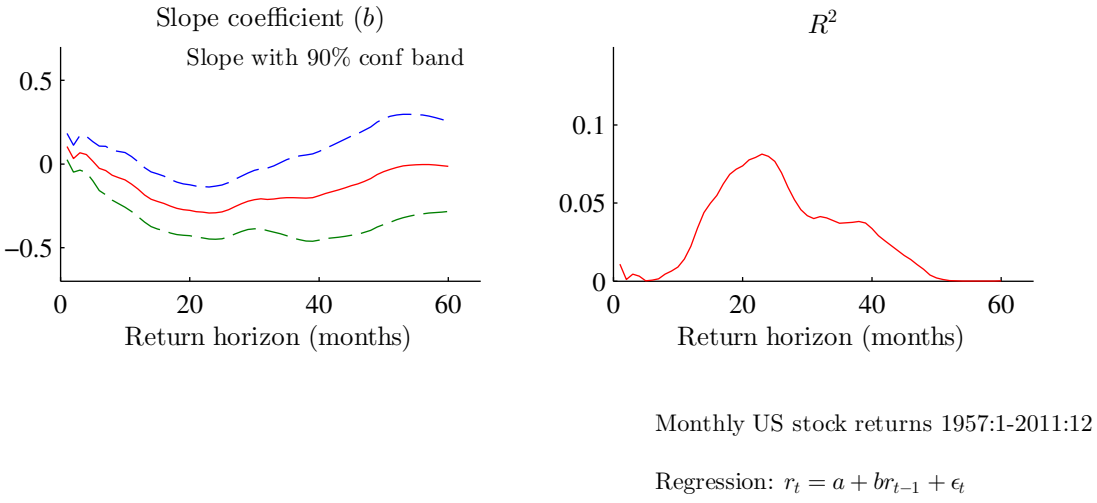


Figure 7.3: Predictability of US stock returns

This evidence suggests that excess returns may perhaps have a predictable component, that is, that (ex ante) risk premia are changing over time. To see how that fits with the consumption-based model, (7.4) says that the conditional expected excess return should equal the conditional covariance times the risk aversion.

Figure 7.4.a shows recursive estimates of the mean return of the aggregate US stock market and the covariance with consumption growth (dated $t+1$). The recursive estimation means that the results for (say) 1965Q2 use data for 1955Q2–1965Q2, the results for 1965Q3 add one data point, etc. The second subfigure shows the same statistics, but estimated on a moving data window of 10 years. For instance, the results for 1980Q2 are for the sample 1971Q3–1980Q2. Finally, the third subfigure uses a moving data window

of 5 years.

Together these figures give the impression that there are fairly long swings in the data. This fundamental uncertainty should serve as a warning against focusing on the fine details of the data. It could also be used as an argument for using longer data series—provided we are willing to assume that the economy has not undergone important regime changes.

It is clear from the earlier Figure 7.4 that the consumption-based model probably cannot generate plausible movements in risk premia. In that figure, the conditional moments are approximated by estimates on different data windows (that is, different subsamples). Although this is a crude approximation, the results are revealing: the actual average excess return and the covariance move in different directions on all frequencies.

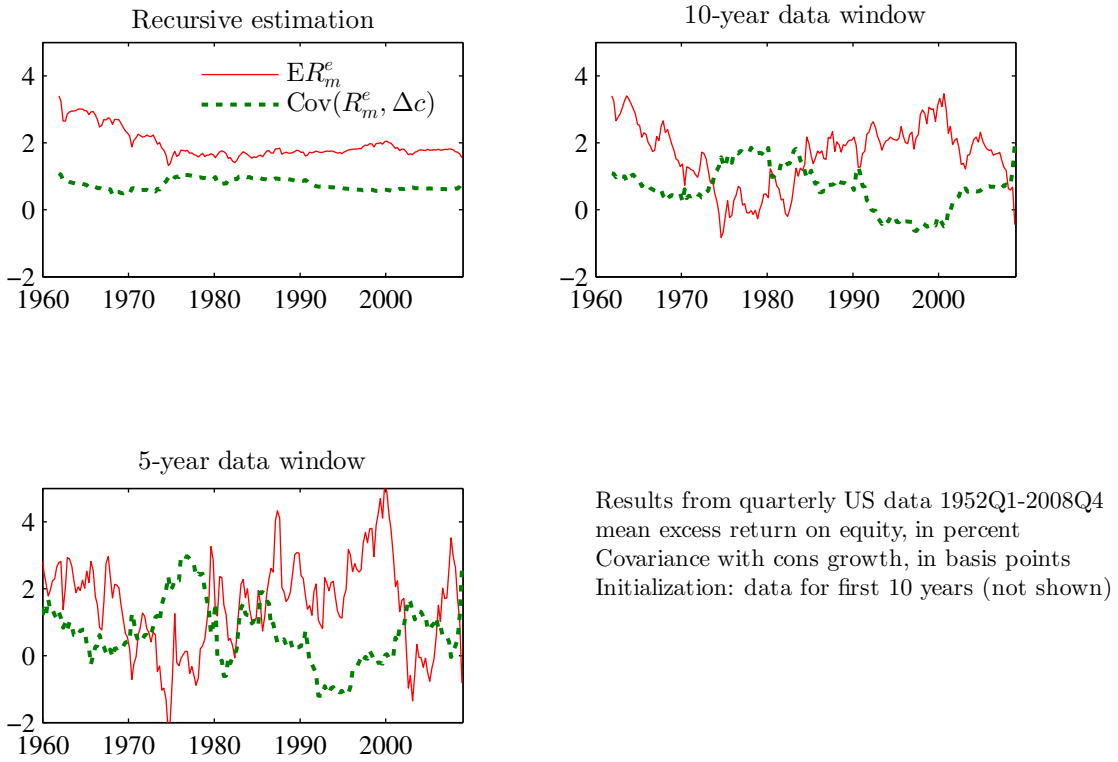


Figure 7.4: The equity premium puzzle for different samples.

7.2.3 The Riskfree Rate Puzzle

The CRRA utility function has the special feature that the intertemporal elasticity of substitution is the inverse of the risk aversion, that is, $1/\gamma$. Choosing the risk aversion parameter, for instance, to fit the equity premium, will therefore have direct effects on the riskfree rate.

A key feature of any consumption-based asset pricing model, or any consumption/saving model for that matter, is that the riskfree rate governs the time slope of the consumption profile. From the asset pricing equation for a riskfree asset (7.1) we have $\mathrm{E}_{t-1}(R_{ft})\,\mathrm{E}_{t-1}(M_t) = 1$. Note that we must use the conditional asset pricing equation—at least as long as we believe that the riskfree asset is a random variable. A riskfree asset is defined by having a zero conditional covariance with the SDF, which means that it is regarded as riskfree at the time of investment $(t-1)$. In practice, this means a real interest rate (perhaps approximated by the real return on a T-bill since the innovations in inflation are small), which may well have a nonzero unconditional covariance with the SDF.[1] Indeed, in Table 7.1 the real return on a T-bill is as correlated with consumption growth as the aggregate US stockmarket.

When the log SDF is normally distributed (the same assumption as before), then the log expected riskfree rate is

$$\ln \mathrm{E}_{t-1}\, R_{ft} = -\ln\beta + \gamma\,\mathrm{E}_{t-1}\,\Delta c_t - \gamma^2\,\mathrm{Var}_{t-1}(\Delta c_t)/2. \tag{7.9}$$

To relate this equation to historical data, we take unconditional expectations to get

$$\mathrm{E}\ln \mathrm{E}_{t-1}\, R_{ft} = -\ln\beta + \gamma\,\mathrm{E}\,\Delta c_t - \gamma^2\,\mathrm{E}\,\mathrm{Var}_{t-1}(\Delta c_t)/2. \tag{7.10}$$

Before we try to compare (7.10) with data, several things should be noted. First, the log gross rate is very close to a traditional net rate ($\ln(1+z) \approx z$ for small z), so it makes sense to compare with the data in Table 7.1. Second, we can safely disregard the variance term since it is very small, at least as long as we are considering reasonable values of γ. Although the average conditional variance is not directly observable, we know that it must be smaller than the unconditional variance[2], which is very small in Table 7.1. In fact, the

[1] As a very simple example, let $x_t = z_{t-1} + \varepsilon_t$ and $y_t = z_{t-1} + u_t$ where ε_t are u_t uncorrelated with each other and with z_{t-1}. If z_{t-1} is observable in $t-1$, then $\mathrm{Cov}_{t-1}(x_t, y_t) = 0$, but $\mathrm{Cov}(x_t, y_t) = \sigma^2(z_{t-1})$.

[2] Let $\mathrm{E}(y|x)$ and $\mathrm{Var}(y|x)$ be the expectation and variance of y conditional on x. The unconditional variance is then $\mathrm{Var}(y) = \mathrm{Var}[\mathrm{E}(y|x)] + \mathrm{E}[\mathrm{Var}(y|x)]$.

variance is around 0.0001 whereas the mean is around 0.02.

Proof. (of (7.9)) For a riskfree gross return R_f, (7.1) with the SDF (7.2) says $\mathrm{E}_{t-1}(R_{ft})\,\mathrm{E}_{t-1}[\beta(C_t/C_{t-1})^{-\gamma}] = 1$. Recall that if $x \sim N(\mu, \sigma^2)$ and $y = \exp(x)$ then $\mathrm{E}\,y = \exp(\mu + \sigma^2/2)$. When Δc_t is conditionally normally distributed, the log of $\mathrm{E}_{t-1}[\beta(C_t/C_{t-1})^{-\gamma}]$ equals $\ln\beta - \gamma\,\mathrm{E}_{t-1}\,\Delta c_t + \gamma^2\,\mathrm{Var}_{t-1}(\Delta c_t)/2)$. ■

According to (7.10) there are two ways to reconcile a positive consumption growth rate with a low real interest rate (around 1% in Table 7.1): investors may prefer to consume later rather than sooner ($\beta > 1$) or they are willing to substitute intertemporally without too much compensation ($1/\gamma$ is high, that is, γ is low). However, fitting the equity premium requires a high value of γ, so investors must be implausibly patient if (7.10) is to hold. For instance, with $\gamma = 25$ (which is a very conservative guess of what we need to fit the equity premium) equation (7.10) says

$$0.01 = -\ln\beta + 25 \times 0.02 \tag{7.11}$$

(ignoring the variance terms), which requires $\beta \approx 1.6$. This is the *riskfree rate puzzle* stressed by Weil (1989). The basic intuition for this result is that it is hard to reconcile a steep slope of the consumption profile and a low compensation for postponing consumption if people are insensitive to intertemporal prices—unless they are extremely patient (actually, unless they prefer to consume later rather than sooner).

Another implication of a high risk aversion is that the real interest rate should be very volatile, which it is not. According to Table 7.1 the standard deviation of the real interest rate is perhaps twice the standard deviation of consumption growth. From (7.9) the volatility of the (expected) riskfree rate should be

$$\mathrm{Std}[\ln \mathrm{E}_{t-1}\,R_{ft}] = \gamma\,\mathrm{Std}[\mathrm{E}_{t-1}\,\Delta c_t], \tag{7.12}$$

if the conditional variance of consumption growth is constant. This expression says that the standard deviation of expected real interest rate is γ times the standard deviation of expected consumption growth. We cannot observe the conditional expectations directly, and therefore not estimate their volatility. However, a simple example is enough to demonstrate that high values of γ are likely to imply counterfactually high volatility of the real interest rate.

As an approximation, suppose both the riskfree rate and consumption growth are

AR(1) processes. Then (7.12) can be written

$$\text{Corr}[\ln \text{E}_{t-1}(R_{ft}), \ln \text{E}_{t-1}(R_{ft})] \times \text{Std}[\ln \text{E}_{t-1}(R_{ft})] = \gamma \times \text{Corr}(\Delta c_t, \Delta c_{t+1}) \times \text{Std}(\Delta c_t) \tag{7.13}$$

$$0.75 \times 0.02 \approx \gamma \times 0.3 \times 0.01 \tag{7.14}$$

where the second line uses the results in Table 7.1. With $\gamma = 25$, (7.14) implies that the RHS is much too volatile This shows that an intertemporal elasticity of substitution of 1/25 is not compatible with the relatively stable real return on T-bills.

Proof. (of (7.13)) If $x_t = \alpha x_{t-1} + \varepsilon_t$, where ε_t is iid, then $\text{E}_{t-1}(x_t) = \alpha x_{t-1}$, so $\sigma(\text{E}_{t-1}\, x_t) = \alpha\sigma(x_{t-1})$. ■

7.3 The Cross-Section of Returns: Unconditional Models

The previous section demonstrated that the consumption-based model has a hard time explaining the risk premium on a broad equity portfolio—essentially because consumption growth is too smooth to make stocks look particularly risky. However, the model *does* predict a positive equity premium, even if it is not large enough. This suggests that the model may be able to explain the relative risk premia across assets, even if the scale is wrong. In that case, the model would still be useful for some issues. This section takes a closer look at that possibility by focusing on the relation between the average return and the covariance with consumption growth in a cross-section of asset returns.

The key equation is (7.5), which I repeat here for ease of reading

$$\text{E}\, R_t^e = \text{Cov}(R_t^e, \Delta c_t)\gamma. \qquad \text{(EPPn2 again)}$$

This can be tested with a GMM framework or a to the traditional cross-sectional regressions of returns on factors with unknown factor risk premia (see, for instance, Cochrane (2005) chap 12 or Campbell, Lo, and MacKinlay (1997) chap 6).

Remark 7.3 *(GMM estimation of (7.5)) Let there be N assets. The original moment*

conditions are

$$g_T(\beta) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} (\Delta c_t - \mu_{\Delta c}) = 0 \\ (R^e_{it} - \mu_i) = 0 \, for \, i = 1, 2, ..., N \\ [(\Delta c_t - \mu_c)(R^e_{it} - \mu_i) - \sigma_{ci}] = 0 \, for \, i = 1, 2, ..., N \\ (R^e_{it} - \alpha - \sigma_{ci}\kappa) = 0 \, for \, i = 1, 2, ..., N, \end{bmatrix}$$

where $\mu_{\Delta c}$ is the mean of Δc_t, μ_i the mean of R^e_{it}, σ_{ci} the covariance of Δc_t and R^e_{it}. This gives $1 + 3N$ moment conditions and $2N + 3$ parameters, so there are $N - 2$ overidentifying restrictions.
To estimate, we define the combined moment conditions as

$$Ag_T(\beta) = \mathbf{0}_{(2N+3)\times 1}, \text{ where}$$

$$A_{(2N+3)\times(1+3N)} = \begin{bmatrix} 1 & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} \\ \mathbf{0}_{N\times 1} & I_N & \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times 1} & \mathbf{0}_{N\times N} & I_N & \mathbf{0}_{N\times N} \\ 0 & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} & \sigma'_{ic} \\ 0 & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} & \mathbf{1}_{1\times N} \end{bmatrix},$$

where σ'_{ic} is an $1 \times N$ vector of covariances of the returns with consumption growth.
These moment conditions mean that means and covariances are estimated in the traditional way, and that κ is estimated by a LS regression of $\mathrm{E}\, R^e_{it}$ *on a constant and σ_{ci}. The test that the pricing errors are all zero is a Wald test that $g_T(\beta)$ are all zero, where the covariance matrix of the moments are estimated by a Newey-West method (using one lag). This covariance matrix is singular, but that does not matter (as we never have to invert it).*

It can be shown (see Söderlind (2006)) that *(i)* the recursive utility function in Epstein and Zin (1991); *(ii)* the habit persistence model of Campbell and Cochrane (1999) in the case of no return predictability, as well as the *(iii)* models of idiosyncratic risk by Mankiw (1986) and Constantinides and Duffie (1996) also in the case of no return predictability, all imply that (7.5) hold. There only difference is that the effective risk aversion (γ) differs. Still, the basic asset pricing implication is the same: expected returns are linearly related to the covariance.

Figure 7.5 shows the results of both C-CAPM and the standard CAPM—for the 25

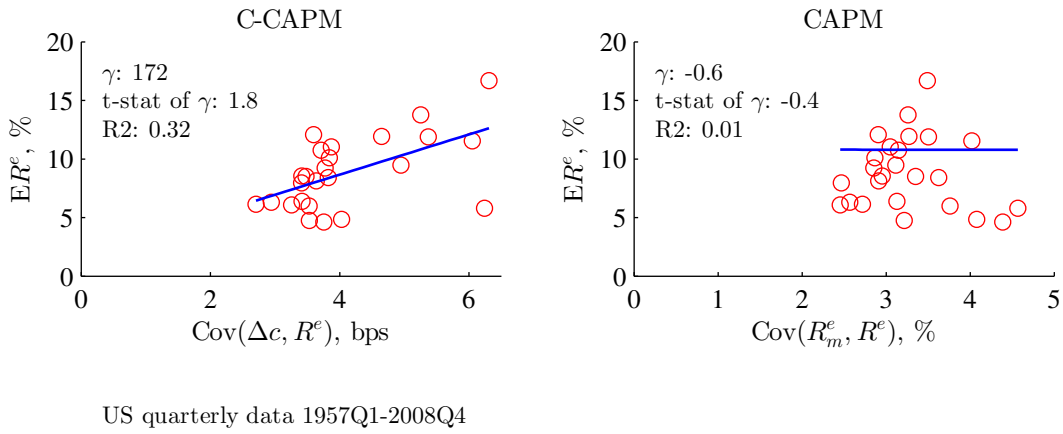


Figure 7.5: Test of C-CAPM and CAPM on 25 FF portfolios

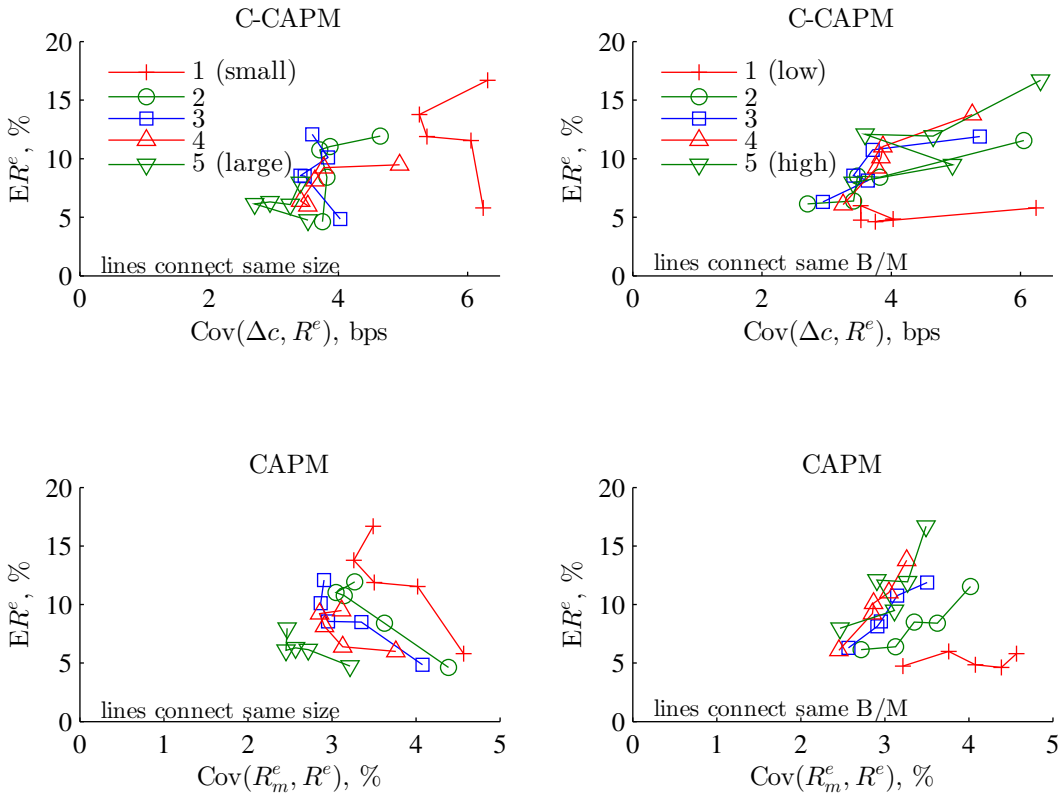


Figure 7.6: Diagnosing C-CAPM and CAPM, 25 FF portfolios

Fama and French (1993) portfolios. It is clear that both models work badly, but CAPM actually worse.

Figure 7.6 takes a careful look at how the C-CAPM and CAPM work in different smaller cross-sections. A common feature of both models is that growth firms (low book-to-market ratios) have large pricing errors (in the figures with lines connecting the same B/M categories, they are the lowest lines for both models). See also Table 7.2–7.4)

In contrast, a major difference between the models is that CAPM shows a very strange pattern when we compare across B/M categories (lines connecting the same size category): mean excess returns are decreasing in the covariance with the market—the wrong *sign* compared to the CAPM prediction. This is not the case for C-CAPM.

The conclusion is that the consumption-based model is not good at explaining the cross-section of returns, but it is no worse than CAPM—if it is any comfort.

	B/M				
	1	2	3	4	5
Size 1	−6.6	−1.2	1.0	3.0	4.1
2	−3.4	−0.1	2.6	2.6	2.2
3	−4.1	0.7	1.0	1.7	4.1
4	−1.8	−1.3	0.2	1.1	−0.7
5	−3.1	−0.5	−0.7	−1.3	0.3

Table 7.2: **Historical minus fitted risk premia (annualised %) from the unconditional model.** Results are shown for the 25 equally-weighted Fama-French portfolios, formed according to size and book-to-market ratios (B/M). Sample: 1957Q1-2008Q4

7.4 The Cross-Section of Returns: Conditional Models

The basic asset pricing model is about conditional moment and it can be summarizes as in (7.4) which is given here again

$$\mathrm{E}_{t-1}\, R_t^e = \mathrm{Cov}_{t-1}(R_t^e, \Delta c_t)\gamma. \qquad \text{(EPP3c again)}$$

Expression this in terms of unconditional moments as in (7.5) shows only part of the story. It is, however, fair to say that if the model does not hold unconditionally, then that is enough to reject the model.

	B/M				
	1	2	3	4	5
Size 1	5.8	11.0	11.9	13.8	16.6
2	4.7	8.4	10.7	11.0	12.0
3	4.8	8.4	8.6	10.2	12.0
4	6.0	6.4	8.2	9.3	9.6
5	4.7	6.1	6.3	6.1	8.0

Table 7.3: **Historical risk premia (annualised %).** Results are shown for the 25 equally-weighted Fama-French portfolios, formed according to size and book-to-market ratios (B/M) Sample: 1957Q1-2008Q4

	B/M				
	1	2	3	4	5
Size 1	−114.5	−10.5	8.5	21.9	24.9
2	−73.5	−0.7	23.8	23.6	18.2
3	−85.1	8.7	11.5	16.8	33.7
4	−30.4	−19.6	1.8	12.3	−6.8
5	−65.2	−7.8	−11.0	−22.1	4.2

Table 7.4: **Relative errors of risk premia (in %) of the unconditional model.** The relative errors are defined as historical minus fitted risk premia, divided by historical risk premia. Results are shown for the 25 equally-weighted Fama-French portfolios, formed according to size and book-to-market ratios (B/M). Sample: 1957Q1-2008Q4

However, it can be shown (see Söderlind (2006)) that several refinements of the consumption based model (the habit persistence model of Campbell and Cochrane (1999) and also the model with idiosyncratic risk by Mankiw (1986) and Constantinides and Duffie (1996)) also imply that (7.4) holds, but with a time varying effective risk aversion coefficient (so γ should carry a time subscript).

7.4.1 Approach 1 of Testing the Conditional CCAPM: A Scaled Factor Model

Reference: Lettau and Ludvigson (2001b), Lettau and Ludvigson (2001a)

Lettau and Ludvigson (2001b) use a scaled factor model, where they impose the restriction that the time variation (using a beta representation) is a linear function of some conditioning variables (specifically, the cay variable) only.

The *cay* variable is defined as the log consumption/wealth ratio. Wealth consists of both financial assets and human wealth. The latter is not observable, but is assumed to be proportional to current income (this would, for instance, be true if income follows and AR(1) process). Therefore, cay is modelled as

$$cay_t = c_t - \omega a_t - (1-\omega)y_t, \tag{7.15}$$

where c_t is log consumption, a_t log financial wealth and y_t is log income. The coefficient ω is estimated with LS to be around 0.3. Although (7.15) contains non-stationary variables, it is interpreted as a cointegrating relation so LS is an appropriate estimation method. Lettau and Ludvigson (2001a) shows that cay is able to forecast stock returns (at least, in-sample). Intuitively, cay should be a signal of investor expectations about future returns (or wage earnings...): a high value is probably driven by high expectations.

The SDF is modelled as time-varying function of consumption growth

$$M_t = a_t + b_t \Delta c_t, \text{ where} \tag{7.16}$$

$$a_t = \gamma_0 + \gamma_1 cay_{t-1} \text{ and } b_t = \eta_0 + \eta_1 cay_{t-1}. \tag{7.17}$$

This is a *conditional C-CAPM*. It is clearly the same as specifying a linear factor model

$$R^e_{it} = \alpha + \beta_{i1} cay_{t-1} + \beta_{i2}\Delta c_t + \beta_{i3}(\Delta c_t \times cay_{t-1}) + \varepsilon_{it}, \tag{7.18}$$

where the coefficients are estimated in time series regression (this is also called a scaled factor model since the "true" factor, Δc, is scaled by the instrument, cay). Then, the cross-sectional pricing implications are tested by

$$\mathrm{E}\, R^e_t = \beta\lambda, \tag{7.19}$$

where $(\beta_{i2}, \beta_{i2}, \beta_{i3})$ is row i of the β matrix and λ is a 3×1 vector of factor risk premia.

Lettau and Ludvigson (2001b) use the 25 Fama-French portfolios as test assets and compare the results from (7.18)–(7.19) with several other models, for instance, a traditional CAPM (the SDF is linear in the market return), a conditional CAPM (the SDF is linear in the market return, cay and their product), a traditional C-CAPM (the SDF is linear in consumption growth) and a Fama-French model (the SDF is linear in the market return, SMB and HML). It is found that the conditional CAPM and C-CAPM provides a much better fit of the cross-sectional returns that the unconditional models (including the

Fama-French model)—and that the C-CAPM is actually a pretty good model.

7.4.2 Approach 2 of Testing the Conditional CCAPM: An Explicit Volatility Model

Reference: Duffee (2005)

Duffee (2005) estimates the conditional model (7.4) by projecting both ex post returns and covariances on a set of instruments—and then studies if there is a relation between these projections.

A conditional covariance (here of the asset return and consumption growth) is the covariance of the innovations. To create innovations (denoted $e_{R,t}$ and $e_{c,t}$ below), the paper uses the following prediction equations

$$R_t^e = \alpha_R' Y_{R,t-1} + e_{R,t} \tag{7.20}$$

$$\Delta c_t = \alpha_c' Y_{c,t-1} + e_{c,t}. \tag{7.21}$$

In practice, only three lags of lagged consumption growth is used to predict consumption growth and only the cay variable is used to predict the asset return.

Then, the return is related to the covariance as

$$R_t^e = b_0 + (b_1 + b_2 p_{t-1})\, e_{R,t} e_{c,t} + w_t, \tag{7.22}$$

where $(b_1 + b_2 p_{t-1})$ is a model of the effective risk aversion. In the CRRA model, $b_2 = 0$, so b_1 measures the relative risk aversion as in (7.4). In contrast, in Campbell and Cochrane (1999) p_{t-1} is an observable proxy of the "surplus ratio" which measure how close consumption is to the habit level.

The model (7.20)–(7.22) is estimated with GMM, using a number of instruments (Z_{t-1}): lagged values of stock market value/consumption, stock market returns, cay and the product of demeaned consumption and returns. This can be thought of as first finding proxies for

$$\widehat{\mathrm{E}_{t-1}} R_t^e = \alpha_R' Y_{R,t-1} \text{ and } \widehat{\mathrm{Cov}_{t-1}}(e_{R,t}, e_{c,t}) = \alpha_v' Z_{t-1} \tag{7.23}$$

and then relating this proxies as

$$\widehat{\mathrm{E}_{t-1}} R_t^e = b_0 + (b_1 + b_2 p_{t-1})\, \widehat{\mathrm{Cov}_{t-1}}(e_{R,t}, e_{c,t}) + u_t. \tag{7.24}$$

The point of using a (GMM) system is that this allows handling the estimation uncer-

tainty of the prediction equations in the testing of the relation between the predictions.

The empirical results (using monthly returns on the broad U.S. stock market and per capita expenditures in nondurables and services, 1959–2001) suggest that there is a strong negative relation between the conditional covariance and the conditional expected market return—which is clearly at odds with a CRRA utility function (compare (7.4)). In addition, typical proxies of the p_{t-1} variable do not seem to any important (economic) effects.

In an extension, the paper also studies other return horizons and tries other ways to model volatility (including a DCC model).

(See also Söderlind (2006) for a related approach applied to a cross-section of returns.)

7.5 Ultimate Consumption

Reference: Parker and Julliard (2005)

Parker and Julliard (2005) suggest using a measure of long-run changes in consumption instead of just a one-period change. This turns out to give a much better empirical fit of the cross-section of risk premia.

To see the motivation for this approach, consider the asset pricing equation based on a CRRA utility function. It says that an excess return satisfies

$$\mathrm{E}_{t-1}\, R_t^e (C_t/C_{t-1})^{-\gamma} = 0 \tag{7.25}$$

Similarly, an n-period bond price ($P_{n,t}$) satisfies

$$\mathrm{E}_t\, \beta^n (C_{t+n}/C_t)^{-\gamma} = P_{nt}, \text{ so} \tag{7.26}$$

$$C_t^{-\gamma} = \mathrm{E}_t\, \beta^n C_{t+n}^{-\gamma}/P_{n,t}. \tag{7.27}$$

Use in (7.25) to get

$$\mathrm{E}_{t-1}\, R_t^e M_{n,t} = 0, \text{ where } M_{n,t} = (1/P_{n,t})(C_{t+n}/C_{t-1})^{-\gamma}. \tag{7.28}$$

This expression relates the one-period excess return to an n-period SDF—which involves the interest rate ($1/P_{n,t}$) and ratio of marginal utilities n periods apart.

If we can apply Stein's lemma (possibly extended) and use $y_{n,t} = \ln 1/P_{nt}$ to denote

the n-period log riskfree rate, then we get

$$\begin{aligned} \mathrm{E}_{t-1}\, R_t^e &= -\mathrm{Cov}_{t-1}(R_t^e, \ln M_{n,t}) \\ &= \mathrm{Cov}_{t-1}[R_t^e, \gamma \ln(C_{t+n}/C_{t-1})] - \mathrm{Cov}_{t-1}[R_t^e, y_{n,t}]. \end{aligned} \tag{7.29}$$

This first term is very similar to the traditional expression (7.2), except that we here have the $(n+1)$-period (instead of the 1-period) consumption growth. The second term captures the covariance between the excess return and the n-period interest rate in period t (both are random as seen from $t-1$). If we set $n = 0$, then this equation simplifies to the traditional expression (7.2). Clearly, the moments in (7.29) could be unconditional instead of conditional.

The empirical approach in Parker and Julliard (2005) is to estimate (using GMM) and test the cross-sectional implications of this model. (They do not use Stein's lemma.) They find that the model fits data much better with a high value of n ("ultimate consumption") than with $n = 0$ (the traditional model). Possible reasons could be: *(i)* long-run changes in consumption are better measured in national accounts data; *(ii)* the CRRA model is a better approximation for long-run movements.

Proof. (of (7.26)–(7.28)) To prove (7.26), let $M_{t+1} = \beta(C_{t+1}/C_t)^{-\gamma}$ denote the SDF and P_{nt} the price of an n-period bond. Clearly, $P_{2t} = \mathrm{E}_t\, M_{t+1} P_{1,t+1}$, so $P_{2t} = \mathrm{E}_t\, M_{t+1}\, \mathrm{E}_{t+1}(M_{t+2} P_{0,t+2})$. Use the law of iterated expectations (LIE) and $P_{0,t+2} = 1$ to get $P_{2t} = \mathrm{E}_t\, M_{t+2} M_{t+1}$. The extension from 2 to n is straightforward, which gives (7.26). To prove (7.28), use (7.27) in (7.25), apply LIE and simplify. ■

Bibliography

Bossaert, P., 2002, *The paradox of asset pricing*, Princeton University Press.

Campbell, J. Y., 2003, "Consumption-based asset pricing," in George Constantinides, Milton Harris, and Rene Stultz (ed.), *Handbook of the Economics of Finance* . chap. 13, pp. 803–887, North-Holland, Amsterdam.

Campbell, J. Y., and J. H. Cochrane, 1999, "By force of habit: a consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, 107, 205–251.

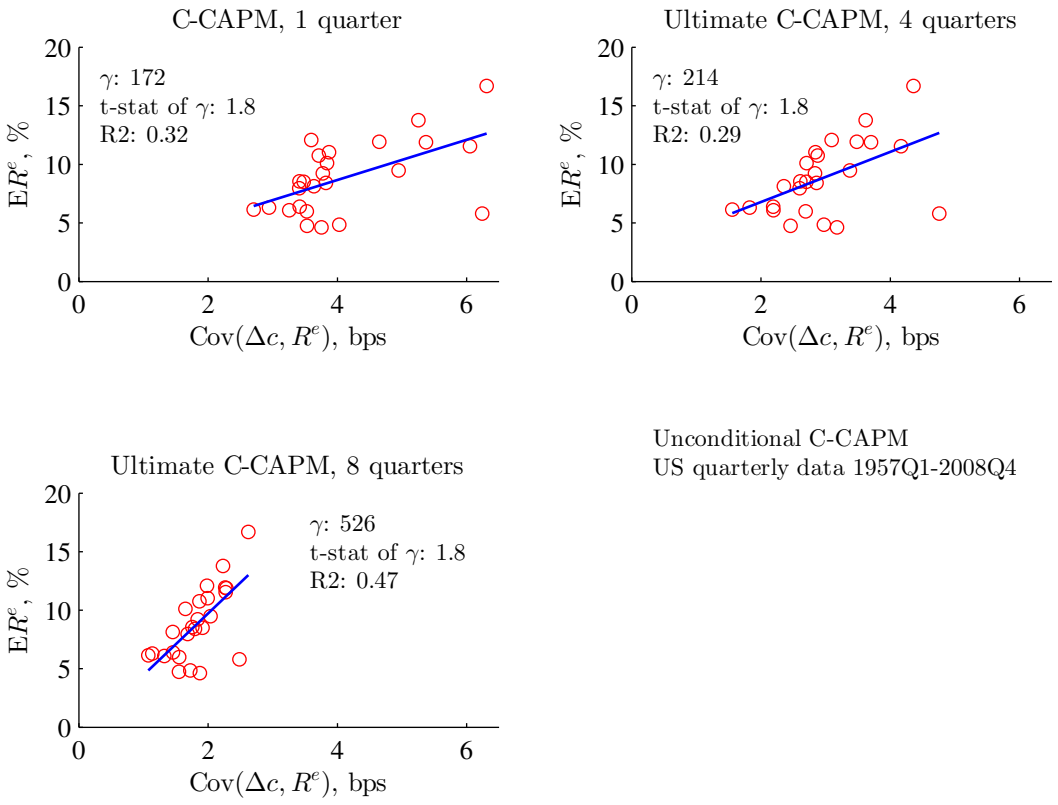


Figure 7.7: **C-CAPM and ultimate consumption, 25 FF portfolio.**

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.

Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.

Constantinides, G. M., and D. Duffie, 1996, "Asset pricing with heterogeneous consumers," *The Journal of Political Economy*, 104, 219–240.

Duffee, G. R., 2005, "Time variation in the covariance between stock returns and consumption growth," *Journal of Finance*, 60, 1673–1712.

Engle, R. F., 2002, "Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business and Economic Statistics*, 20, 339–351.

Epstein, L. G., and S. E. Zin, 1991, "Substitution, risk aversion, and the temporal behavior of asset returns: an empirical analysis," *Journal of Political Economy*, 99, 263–286.

Fama, E. F., and K. R. French, 1988a, "Dividend yields and expected stock returns," *Journal of Financial Economics*, 22, 3–25.

Fama, E. F., and K. R. French, 1988b, "Permanent and temporary components of stock prices," *Journal of Political Economy*, 96, 246–273.

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.

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.

Lettau, M., and S. Ludvigson, 2001a, "Consumption, wealth, and expected stock returns," *Journal of Finance*, 56, 815–849.

Lettau, M., and S. Ludvigson, 2001b, "Resurrecting the (C)CAPM: a cross-sectional test when risk premia are time-varying," *Journal of Political Economy*, 109, 1238–1287.

Mankiw, G. N., 1986, "The equity premium and the concentration of aggregate shocks," *Journal of Financial Economics*, 17, 211–219.

Mehra, R., and E. Prescott, 1985, "The equity premium: a puzzle," *Journal of Monetary Economics*, 15, 145–161.

Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric foundations*, Cambridge University Press, Cambridge.

Parker, J., and C. Julliard, 2005, "Consumption risk and the cross section of expected returns," *Journal of Political Economy*, 113, 185–222.

Smith, P. N., and M. R. Wickens, 2002, "Asset pricing with observable stochastic discount factors," Discussion Paper No. 2002/03, University of York.

Söderlind, P., 2006, "C-CAPM Refinements and the cross-section of returns," *Financial Markets and Portfolio Management*, 20, 49–73.

Söderlind, P., 2009, "An extended Stein's lemma for asset pricing," *Applied Economics Letters,* forthcoming, 16, 1005–1008.

Weil, P., 1989, "The equity premium puzzle and the risk-free rate puzzle," *Journal of Monetary Economics*, 24, 401–421.

8 Expectations Hypothesis of Interest Rates

8.1 Term (Risk) Premia

Term risk premia can be defined in several ways. All these premia are zero (or at least constant) under the expectations hypothesis.

A *yield term premium* is defined as the difference between a long (n-period) interest rate and the expected average future short (m-period) rates over the same period

$$\varphi_t^y(n,m) = y_{nt} - \frac{1}{k}\sum_{s=0}^{k-1} \mathrm{E}_t\, y_{m,t+sm}, \text{ with } k = n/m. \tag{8.1}$$

Figure 8.1 illustrates the timing.

Example 8.1 *(Yield term premium, rolling over 3-month rates for a year)*

$$\varphi_t^y(1,1/4) = y_{1y,t} - \frac{1}{4}\mathrm{E}_t\left(y_{3m,t} + y_{3m,t+3m} + y_{3m,t+6m} + y_{3m,t+9m}\right).$$

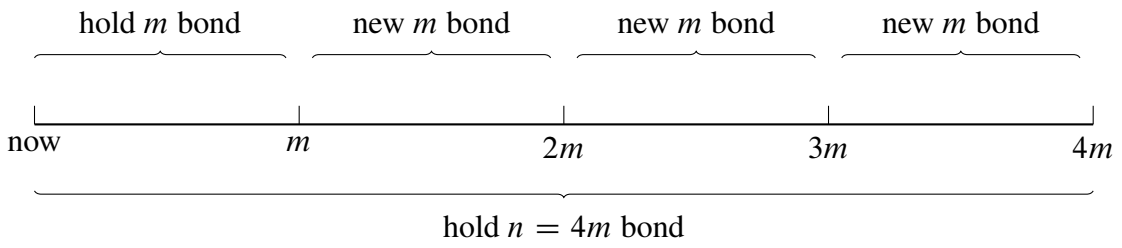


Figure 8.1: Timing for yield term premium

The (m-period) *forward term premium* is the difference between a forward rate for an m-period investment (starting in k periods ahead) and the expected short interest rate.

$$\varphi_t^f(k,m) = f_t(k,k+m) - \mathrm{E}_t\, y_{m,t+k}, \tag{8.2}$$

where $f_t(k,k+m)$ is a forward rate that applies for the period $t+k$ to $t+k+m$. Figure 8.2 illustrates the timing.

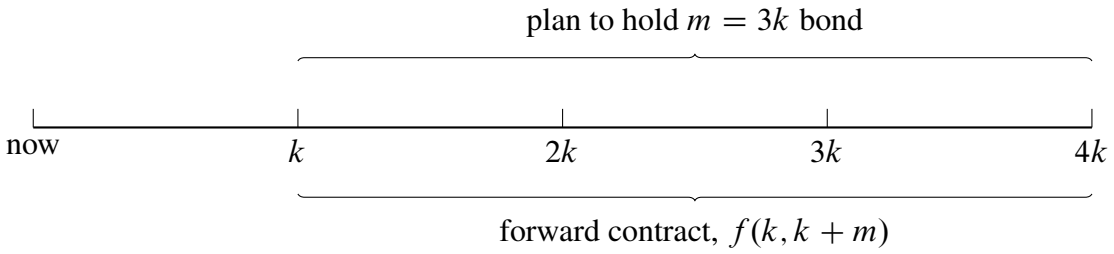


Figure 8.2: Timing for forward term premium

Finally, the *holding-period premium* is the expected excess return of holding an n-period bond between t and $t+m$ (buy it in t for P_{nt} and sell it in $t+m$ for $P_{n-m,t+m}$)—in excess of holding an m-period bond over the same period

$$\begin{aligned}\varphi_t^h(n,m) &= \frac{1}{m}\,\mathrm{E}_t \ln(P_{n-m,t+m}/P_{nt}) - y_{mt} \\ &= \frac{1}{m}\left[ny_{nt} - (n-m)\,\mathrm{E}_t\, y_{n-m,t+m}\right] - y_{mt}. \end{aligned} \tag{8.3}$$

Figure 8.3 illustrates the timing. This definition is perhaps most similar to the definition of risk premia of other assets (for instance, equity).

Example 8.2 *(Holding-period premium, holding a 10-year bond for one year).*

$$\begin{aligned}\varphi_t^h(10,1) &= \mathrm{E}_t \ln(P_{9,t+1}/P_{10,t}) - y_{1t} \\ &= \left[10y_{10,t} - 9\,\mathrm{E}_t\, y_{9,t+1}\right] - y_{1t}.\end{aligned}$$

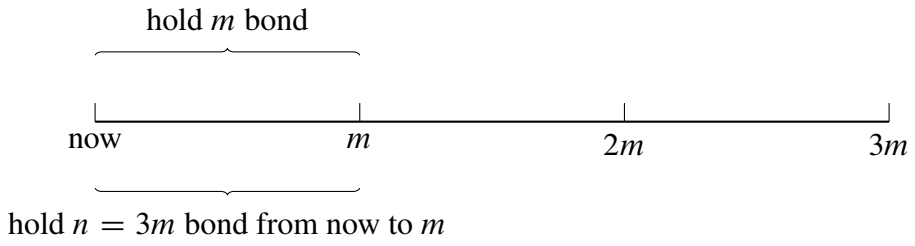


Figure 8.3: Timing for holding-period premium

Notice that these risk premia are all expressed relative to a short(er) rate—they are term premia. Nothing rules out the possibility that the short rate(-er) also includes risk

premia. For instance, a short nominal interest rate is likely to include an inflation risk premium since inflation over the next period is risky. However, this is not the focus here.

The (pure) *expectations hypothesis of interest rates* says that all these risk premia should be constant (or zero if the pure theory).

8.2 Testing the Expectations Hypothesis of Interest Rates

8.2.1 Basic Tests

The basic tests of the expectations hypothesis (EH) is that the realized values of the term premia (replace the expected values by realized values) in (8.1)–(8.3) should be unpredictable. In this case, the regressions of the realized premia on variables that are known in t should have zero slopes ($b_1 = 0, b_2 = 0, b_3 = 0$)

$$y_{nt} - \frac{1}{k}\sum_{s=0}^{k-1} y_{m,t+sm} = a_1 + b_1' x_t + u_{t+n} \tag{8.4}$$

$$f_t(k, k+m) - y_{m,t+k} = a_2 + b_2' x_t + u_{t+k+m} \tag{8.5}$$

$$\frac{1}{m}\ln(P_{n-m,t+m}/P_{nt}) - y_{mt} = a_3 + b_3' x_t + u_{t+n}. \tag{8.6}$$

These tests are based on the maintained hypothesis that the expectation errors (for instance, $y_{m,t+sm} - \mathrm{E}_t\, y_{m,t+sm}$) are unpredictable—as they would be if expectations are rational.

The intercepts in these regressions pick out constant term premia. Non-zero slopes would indicate that the changes of the term premia are predictable—which is at odds with the expectations hypothesis.

Notice that we use realized (instead of expected) values on the left hand side of the tests (8.4)–(8.6). This is valid—under the assumption that expectations can be well approximated by the properties of the sample data. To see that, consider the yield term premium in (8.1) and add/subtract the realized value of the average future short rate, $\sum_{s=0}^{k-1} y_{m,t+sm}/k$,

$$y_{nt} - \frac{1}{k}\sum_{s=0}^{k-1} \mathrm{E}_t\, y_{m,t+sm} = y_{nt} - \frac{1}{k}\sum_{s=0}^{k-1} y_{m,t+sm} + \varepsilon_{t+n}, \text{ where} \tag{8.7}$$

$$\varepsilon_{t+n} = \frac{1}{k}\sum_{s=0}^{k-1} y_{m,t+sm} - \frac{1}{k}\sum_{s=0}^{k-1} \mathrm{E}_t\, y_{m,t+sm}. \tag{8.8}$$

Use RHS of (8.7) in (8.1) to write

$$y_{nt} - \frac{1}{k}\sum_{s=0}^{k-1} y_{m,t+sm} = \varphi_t^y(n,m) - \varepsilon_{t+n} \tag{8.9}$$

Compare with (8.4) to notice that $a_1 + b_1' x_t$ captures the risk premium, $\varphi_t^y(n,m)$. Also notice that ε_{t+n} is the surprise, so it should not be forecastable by any information available in period t—provided expectations are rational. (This does not cause any econometric trouble since ε_{t+m} should be uncorrelated to all regressors—since they are know in t.)

8.2.2 A Single Factor for All Maturities?

Reference: Cochrane and Piazzesi (2005)

Cochrane and Piazzesi (2005) regress excess holding period return on forward rates, that is, (8.6) where x_t contain forward rates. They observe that the slope coefficients are very similar across different maturities of the bonds held (n). It seems as if the coefficients (b_3) for one maturity are the same as the coefficients for another maturity—apart from a common scaling factor. This means that if we construct a "forecasting factor"

$$ff_t = b_3' x_t \tag{8.10}$$

for one maturity (2-year bond, say), then the regressions

$$\frac{1}{m}\ln(P_{n-m,t+m}/P_{nt}) - y_{mt} = a_n + b_n ff_t \tag{8.11}$$

should work almost as well as using the full vector x_t.

Figure 8.4 and *Tables 8.1–8.2* illustrate some results.

8.2.3 Spread-Based Tests

Many classical tests of the expectations hypothesis have only used interest rates as predictors (x_t include only interest rates). In addition, since interest rates have long swings (are close to be non-stationary), the regressions have been expressed in terms of spreads.

To test that the yield term premium is zero (or at last constant), add and subtract y_{mt} (the current short m-period rate) from (8.4) and rearrange to get

$$\frac{1}{k}\sum_{s=0}^{k-1}(y_{m,t+sm} - y_{mt}) = (y_{nt} - y_{mt}) + \varepsilon_{t+n}, \tag{8.12}$$

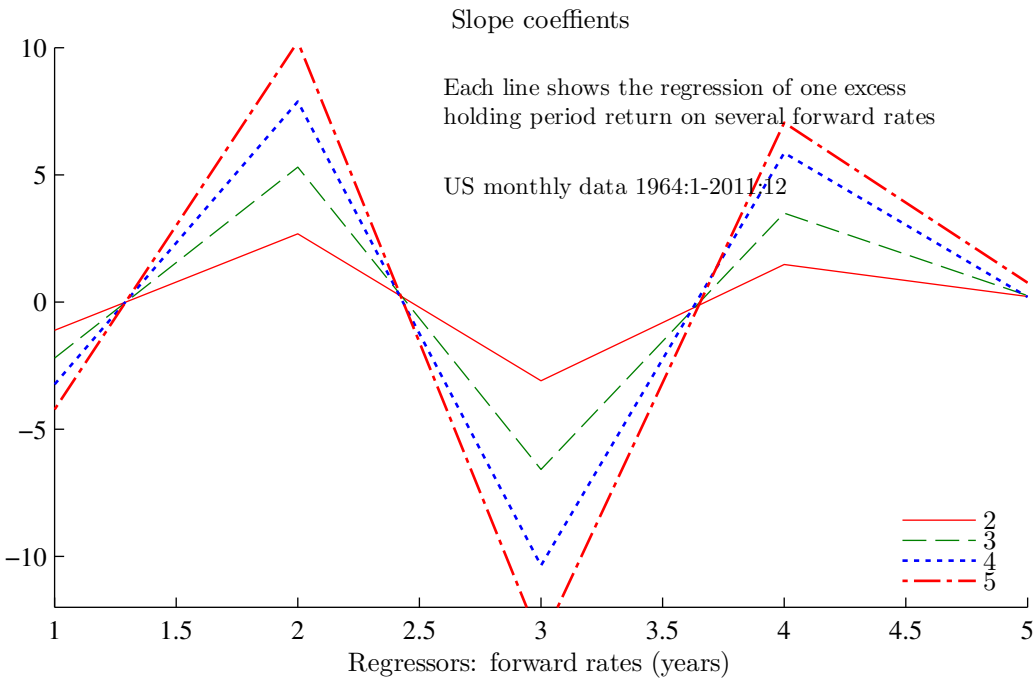


Figure 8.4: A single forecasting factor for bond excess hold period returns

	2	3	4	5
factor	1.00	1.88	2.69	3.46
	(6.48)	(6.66)	(6.82)	(6.98)
constant	−0.00	−0.00	−0.00	−0.00
	(−0.00)	(−0.52)	(−0.94)	(−1.33)
R2	0.14	0.15	0.16	0.17
obs	564.00	564.00	564.00	564.00

Table 8.1: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. U.S. data for 1964:1-2011:12.

which says that the term spread between a long and a short rate ($y_{nt} - y_{mt}$) equals the expected average future change of the short rate (relative to the current short rate).

	2	3	4	5
factor	1.00	1.88	2.69	3.46
	(3.89)	(4.05)	(4.21)	(4.36)
constant	−0.00	−0.00	−0.00	−0.00
	(−0.00)	(−0.25)	(−0.45)	(−0.64)
R2	0.14	0.15	0.16	0.17
obs	564.00	564.00	564.00	564.00

Table 8.2: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. U.S. data for 1964:1-2011:12. Bootstrapped standard errors, with blocks of 10 observations.

Example 8.3 *(Yield term premium, rolling over 3-month rates for a year)*

$$\frac{1}{4}\left[\left(y_{3m,t}-y_{3m,t}\right)+\left(y_{3m,t+3m}-y_{3m,t}\right)+\left(y_{3m,t+6m}-y_{3m,t}\right)+\left(y_{3m,t+9m}-y_{3m,t}\right)\right]=$$
$$y_{12m,t}-y_{3m,t}.$$

(8.12) can be tested by running the regression

$$\frac{1}{k}\sum_{s=0}^{k-1}\left(y_{m,t+sm}-y_{mt}\right)=\alpha+\beta\left(y_{nt}-y_{mt}\right)+\varepsilon_{t+n}, \tag{8.13}$$

where the expectations hypothesis (zero yield term premium) implies $\alpha=0$ and $\beta=1$. (Sometimes the intercept is disregarded). See Figure 8.5 for an empirical example.

Similarly, adding and subtracting y_{mt} to (8.5) and rearranging gives

$$y_{m,t+k}-y_{mt}=\alpha+\beta[f_t(k,k+m)-y_{mt}]+\varepsilon_{t+k+m}, \tag{8.14}$$

where the expectations hypothesis (zero forward term premium) implies $\alpha=0$ and $\beta=1$. This regression tests if the forward-spot spread is an unbiased predictor of the change of the spot rate.

Finally, use (8.3) to rearrange (8.6) as

$$y_{n-m,t+m}-y_{nt}=\alpha+\beta\frac{m}{n-m}\left(y_{nt}-y_{mt}\right)+\varepsilon_{t+n}, \tag{8.15}$$

the expectations hypothesis (zero holding premium) implies $\alpha=0$ and $\beta=1$. If the holding period (m) is short compared to the maturity (n), then this regression (almost)

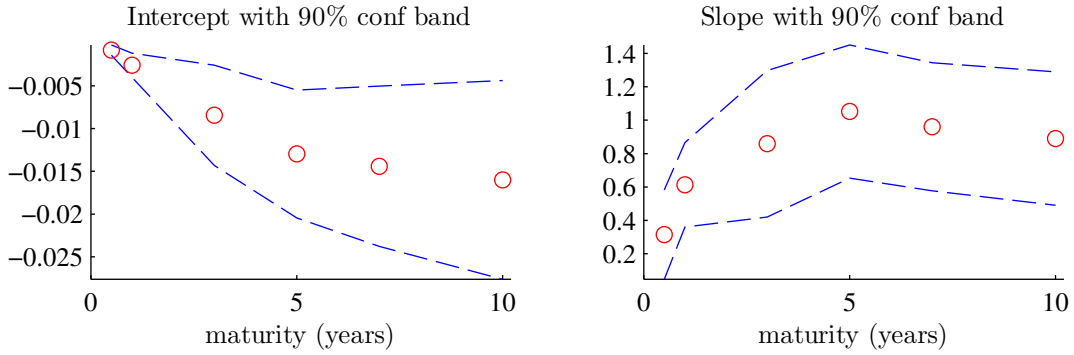


Figure 8.5: Testing the expectations hypothesis on US interest rates

tests if the current spread, scaled by $m/(n-m)$, is an unbiased predictor of the change in the long rate.

8.3 The Properties of Spread-Based EH Tests

Reference: Froot (1989)

The spread-based EH tests ((8.13), (8.14) and (8.15)), can be written

$$\Delta i_{t+1} = \alpha + \beta s_t + \varepsilon_{t+1}, \text{ where} \tag{8.16}$$

$$s_t = \mathrm{E}_t^m \Delta i_{t+1} + \varphi_t, \tag{8.17}$$

where $\mathrm{E}_t^m \Delta i_{t+1}$ is the market's expectations of the interest rate change and φ_t is the risk premium. In this expression, Δi_{t+1} is short hand notation for the dependent variable (which in all three cases is a change of an interest rate) and s_t denotes the regressor (which in all three cases is a term spread).

The regression coefficient in (8.16) is

$$\beta = 1 - \frac{\sigma(\sigma+\rho)}{1+\sigma^2+2\rho\sigma} + \gamma, \qquad \text{where} \tag{8.18}$$
$$\sigma = \frac{\text{Std}(\varphi)}{\text{Std}\left(\text{E}_t^m \Delta i_{t+1}\right)}, \rho = \text{Corr}\left(\text{E}_t^m \Delta i_{t+1}, \varphi\right), \text{ and}$$
$$\gamma = \frac{\text{Cov}\left[\left(\text{E}_t - \text{E}_t^m\right)\Delta i_{t+1}, \text{E}_t^m \Delta i_{t+1} + \varphi\right]}{\text{Var}\left(\text{E}_t^m \Delta i_{t+1} + \varphi\right)},$$

The second term in (8.18) captures the effect of the (time varying) risk premium and the third term (γ) captures any systematic expectations errors ($\left(\text{E}_t - \text{E}_t^m\right)\Delta i_{t+1}$).

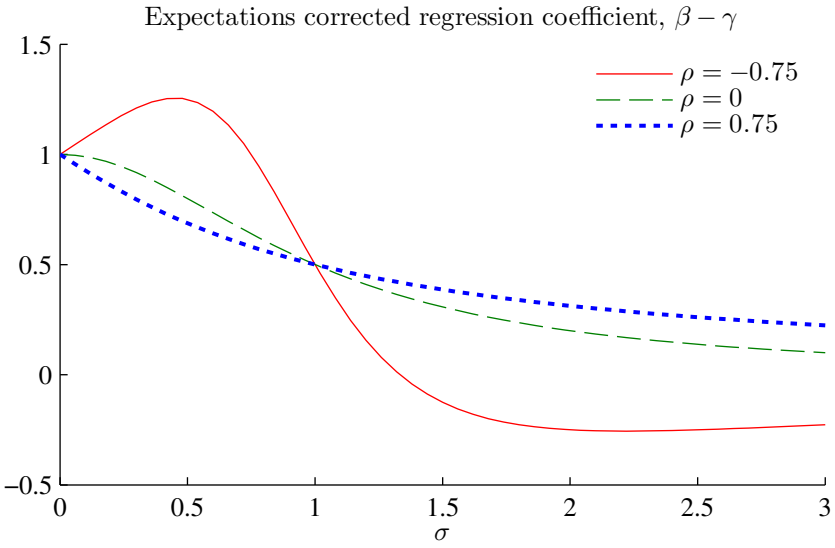


Figure 8.6: Regression coeffcient in EH test

Figure 8.6 shows how the expectations corrected regression coefficient ($\beta-\gamma$) depends on the relative volatility of the term premium and expected interest change (σ) and their correlation (ρ). A regression coefficient of unity could be due to either a constant term premium ($\sigma = 0$), or to a particular combination of relative volatility and correlation ($\rho = -\sigma$), which makes the forward spread an unbiased predictor.

When the correlation is zero, the regression coefficient decreases monotonically with σ, since an increasing fraction of the movements in the forward rate are then due to the risk premium. A coefficient below a half is only possible when the term premium is more

volatile than the expected interest rate change ($\sigma > 1$), and a coefficient below zero also requires a negative correlation ($\rho < 0$).

U.S. data often show β values between zero and one for very short maturities, around zero for maturities between 3 to 9 months, and often relatively close to one for longer maturities. Also, β tends to increase with the forecasting horizon (keeping the maturity constant), at least for horizons over a year.

The specification of the regression equation also matters, especially for long maturities: β is typically negative if the left hand side is the change in long rates, but much closer to one if it is an average of future short rates. The β estimates are typically much closer to one if the regression is expressed in levels rather than differences. Even if this is disregarded, the point estimates for long maturities differ a lot between studies. Clearly, if ρ is strongly negative, then even small changes in σ around one can lead large changes in the estimated β.

Froot (1989) uses a long sample of survey data on interest rate expectations. The results indicate that risk premia are important for the 3-month and 12-month maturities, but not for really long maturities. On the other hand, there seems to be significant systematic expectations errors ($\gamma < 0$) for the long maturities which explain the negative β estimates in ex post data. We cannot, of course, tell whether these expectation errors are due to a small sample (for instance, a "peso problem") or to truly irrational expectations.

Proof. (of (8.18)) Define

$$\begin{aligned} \Delta i_{t+1} &= \mathrm{E}_t\, \Delta i_{t+1} + u_{t+1} \\ \mathrm{E}_t\, \Delta i_{t+1} &= \mathrm{E}_t^m\, \Delta i_{t+1} + \eta_{t+1}. \end{aligned}$$

The regression coefficient is

$$\begin{aligned} \beta &= \frac{\mathrm{Cov}(s_t, \Delta i_{t+1})}{\mathrm{Var}(s_t)} \\ &= \frac{\mathrm{Cov}(\mathrm{E}_t^m\, \Delta i_{t+1} + \varphi_t, \mathrm{E}_t^m\, \Delta i_{t+1} + \eta_{t+1} + u_{t+1})}{\mathrm{Var}(\mathrm{E}_t^m\, \Delta i_{t+1} + \varphi_t)} \\ &= \frac{\mathrm{Var}(\mathrm{E}_t^m\, \Delta i_{t+1})}{\mathrm{Var}(\mathrm{E}_t^m\, \Delta i_{t+1} + \varphi_t)} + \frac{\mathrm{Cov}(\varphi_t, \mathrm{E}_t^m\, \Delta i_{t+1})}{\mathrm{Var}(\mathrm{E}_t^m\, \Delta i_{t+1} + \varphi_t)} + \frac{\mathrm{Cov}(\mathrm{E}_t^m\, \Delta i_{t+1} + \varphi_t, \eta_{t+1})}{\mathrm{Var}(\mathrm{E}_t^m\, \Delta i_{t+1} + \varphi_t)} \end{aligned}$$

The third term is γ. Write the first two terms as

$$\begin{aligned}
\frac{\sigma_{mm}+\sigma_{m\varphi}}{\sigma_{mm}+\sigma_{\varphi\varphi}+2\sigma_{m\varphi}} &= 1+\frac{\sigma_{mm}+\sigma_{m\varphi}-\left(\sigma_{mm}+\sigma_{\varphi\varphi}+2\sigma_{m\varphi}\right)}{\sigma_{mm}+\sigma_{\varphi\varphi}+2\sigma_{m\varphi}} \\
&= 1-\frac{\rho\sigma_m\sigma_\varphi+\sigma_\varphi^2}{\sigma_m^2+\sigma_\varphi^2+2\rho\sigma_m\sigma_\varphi} \\
&= 1-\frac{\left(\rho\sigma_m\sigma_\varphi+\sigma_\varphi^2\right)/\sigma_m^2}{\left(\sigma_m^2+\sigma_\varphi^2+2\rho\sigma_m\sigma_\varphi\right)/\sigma_m^2} \\
&= 1-\frac{\sigma\left(\sigma+\rho\right)}{1+\sigma^2+2\rho\sigma}
\end{aligned}$$

where the second line multiplies by σ_m^2/σ_m^2 and the third line uses the definition $\sigma = \sigma_\varphi/\sigma_m$. ■

Bibliography

Cochrane, J. H., and M. Piazzesi, 2005, "Bond risk premia," *American Economic Review*, 95, 138–160.

Froot, K. A., 1989, "New Hope for the Expectations Hypothesis of the Term Structure of Interest Rates," *The Journal of Finance*, 44, 283–304.

9 Yield Curve Models: MLE and GMM

Reference: Cochrane (2005) 19; Campbell, Lo, and MacKinlay (1997) 11, Backus, Foresi, and Telmer (1998); Singleton (2006) 12–13

9.1 Overview

On average, yield curves tend to be upward sloping (see Figure 9.2), but there is also considerable time variation on both the level and shape of the yield curves.

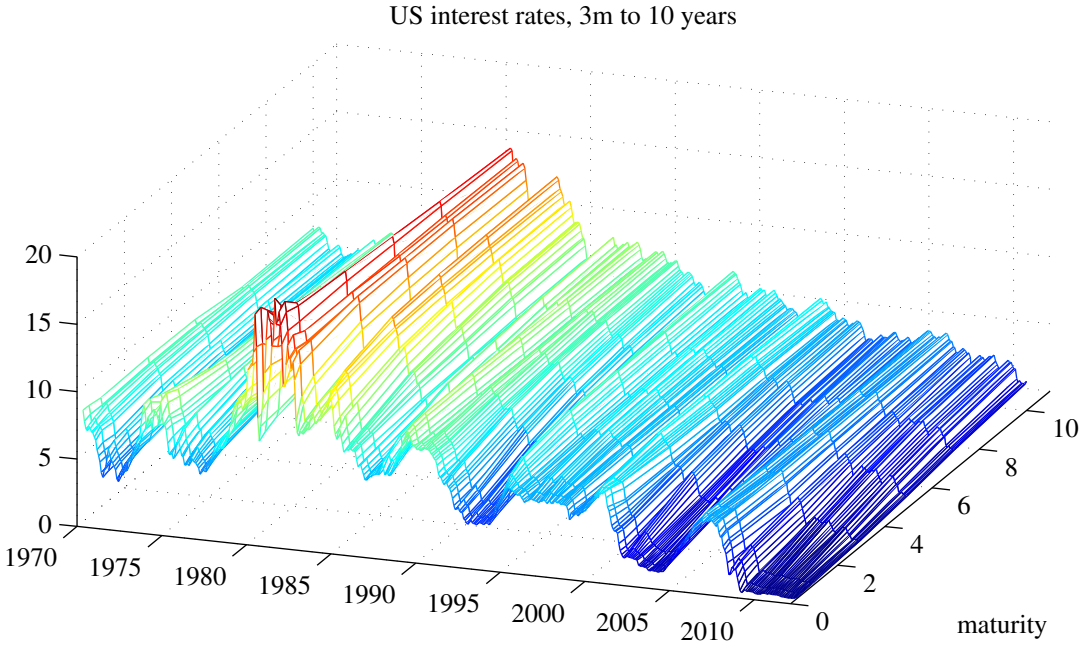


Figure 9.1: US yield curves

It is common to describe the movements in terms of three "factors": level, slope, and

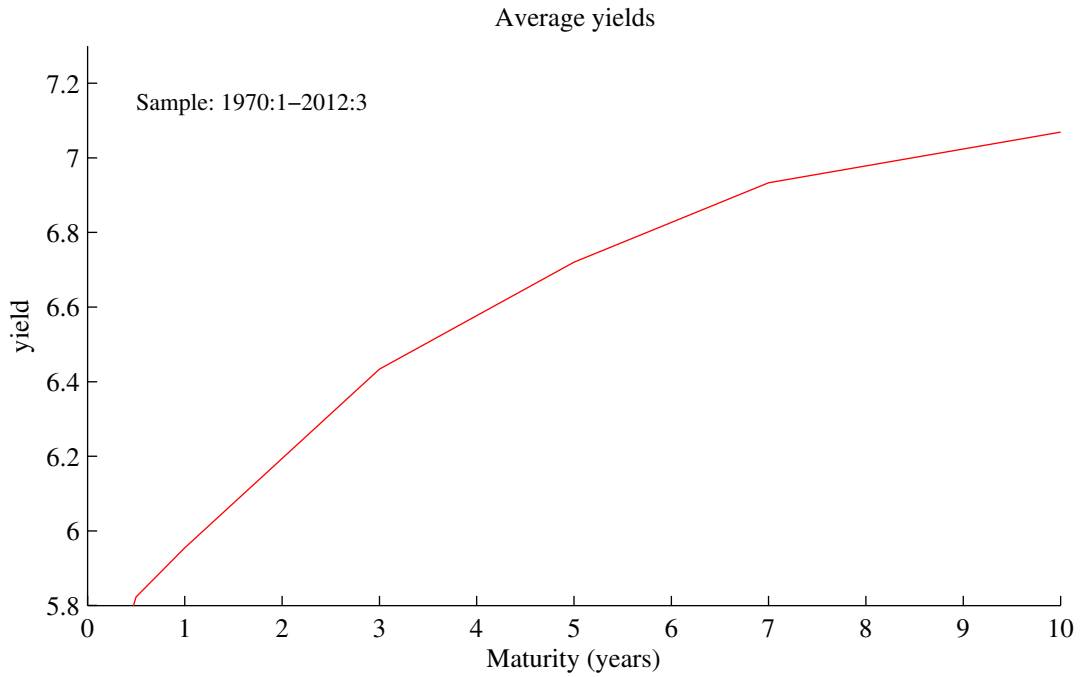


Figure 9.2: Average US yield curve

curvature. One way of measuring these factors is by defining

$$\begin{aligned} \text{Level}_t &= y_{10y} \\ \text{Slope}_t &= y_{10y} - y_{3m} \\ \text{Curvature}_t &= \left(y_{2y} - y_{3m}\right) - \left(y_{10y} - y_{2y}\right). \end{aligned} \tag{9.1}$$

This means that we measure the level by a long rate, the slope by the difference between a long and a short rate—and the curvature (or rather, concavity) by how much the medium/short spread exceeds the long/medium spread. For instance, if the yield curve is hump shaped (so y_{2y} is higher than both y_{3m} and y_{10y}), then the curvature measure is positive. In contrast, when the yield curve is U-shaped (so y_{2y} is lower than both y_{3m} and y_{10y}), then the curvature measure is negative. See Figure 9.3 for an example.

An alternative is to use principal component analysis. See Figure 9.4 for an example.

Remark 9.1 *(Principal component analysis) The first (sample) principal component of the zero (possibly demeaned) mean $N \times 1$ vector z_t is $w_1' z_t$ where w_1 is the eigenvec-*

tor associated with the largest eigenvalue of $\Sigma = \text{Cov}(z_t)$. *This value of* w_1 *solves the problem* $\max_w w'\Sigma w$ *subject to the normalization* $w'w = 1$. *This eigenvalue equals* $\text{Var}(w_1'z_t) = w_1'\Sigma w_1$. *The* j*th principal component solves the same problem, but under the additional restriction that* $w_i'w_j = 0$ *for all* $i < j$. *The solution is the eigenvector associated with the* j*th largest eigenvalue (which equals* $\text{Var}(w_j'z_t) = w_j'\Sigma w_j$*). This means that the first* K *principal components are those (normalized) linear combinations that account for as much of the variability as possible—and that the principal components are uncorrelated (*$\text{Cov}(w_i'z_t, w_j'z_t) = 0$*). Dividing an eigenvalue with the sum of eigenvalues gives a measure of the relative importance of that principal component (in terms of variance). If the rank of* Σ *is* K*, then only* K *eigenvalues are non-zero.*

Remark 9.2 *(Principal component analysis 2) Let* W *be* N *x* N *matrix with* w_i *as column* i. *We can the calculate the* N*x*1 *vector of principal components as* $pc_t = W'z_t$. *Since* $W^{-1} = W'$ *(the eigenvectors are orthogonal), we can invert as* $z_t = Wpc_t$. *The* w_i *vector (column* i *of* W*) therefore shows how the different elements in* z_t *change as the* i*th principal component changes.*

Interest rates are strongly related to business cycle conditions, so it often makes sense to include macro economic data in the modelling. See Figure 9.5 for how the term spreads are related to recessions: term spreads typically increase towards the end of recessions. The main reason is that long rates increase before short rates.

9.2 Risk Premia on Fixed Income Markets

There are many different types of risk premia on fixed income markets.

Nominal bonds are risky in real terms, and are therefore likely to carry *inflation risk premia*. Long bonds are risky because their market values fluctuate over time, so they probably have *term premia*. Corporate bonds and some government bonds (in particular, from developing countries) have *default risk premia*, depending on the risk for default. Interbank rates may be higher than T-bill of the same maturity for the same reason (see the TED spread, the spread between 3-month Libor and T-bill rates) and illiquid bonds may carry *liquidity premia* (see the spread between off-the run and on-the-run bonds).

Figures 9.6–9.9 provide some examples.

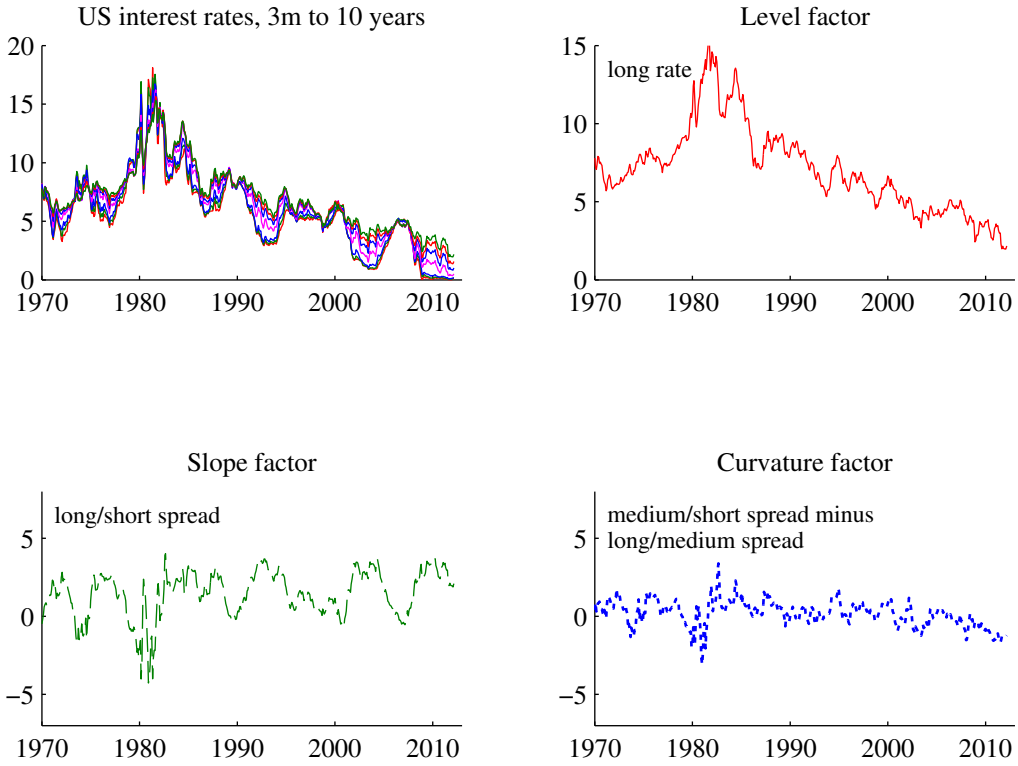


Figure 9.3: US yield curves: level, slope and curvature

9.3 Summary of the Solutions of Some Affine Yield Curve Models

An affine yield curve model implies that the yield on an n-period discount bond can be written

$$\begin{aligned} y_{nt} &= a_n + b_n' x_t, \text{ where} \\ a_n &= A_n/n \text{ and } b_n = B_n/n, \end{aligned} \tag{9.2}$$

where x_t is an $K \times 1$ vector of state variables. The A_n (a scalar) and the B_n (an $K \times 1$ vector) are discussed below.

The price of an n-period bond equals the cross-moment between the pricing kernel (M_{t+1}) and the value of the same bond next period (then an $n-1$-period bond)

$$P_{nt} = \mathrm{E}_t\, M_{t+1} P_{n-1,t+1}. \tag{9.3}$$

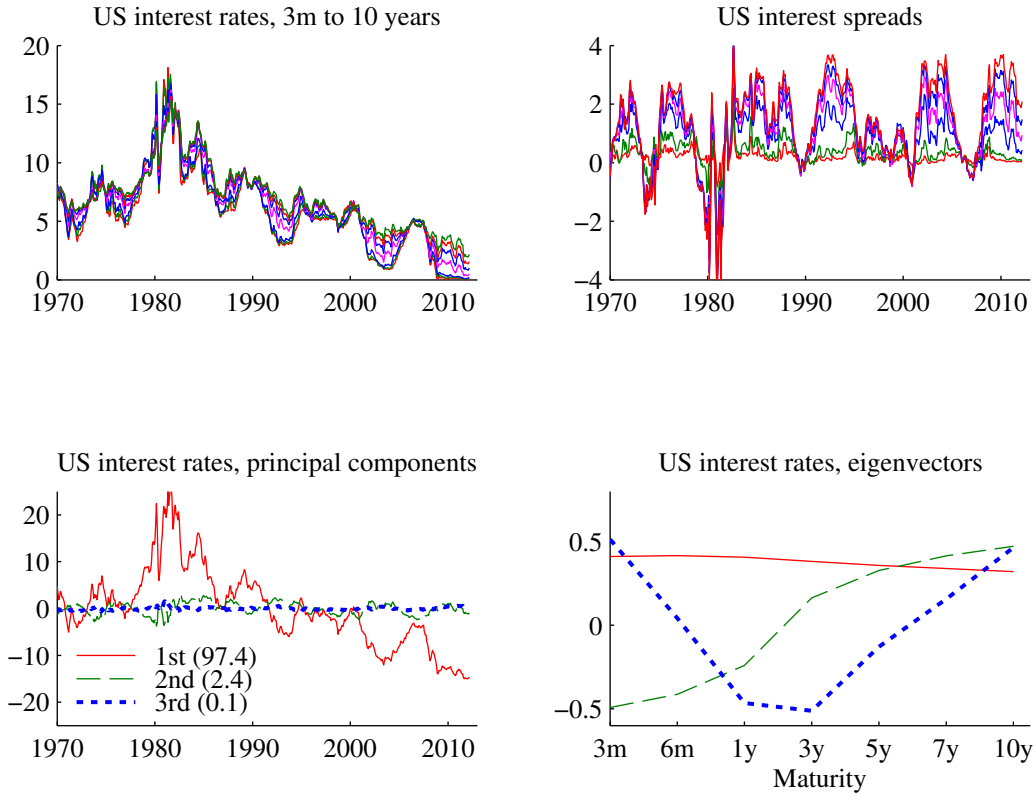


Figure 9.4: US yield curves and principal components

The *Vasicek model* assumes that the log SDF (m_{t+1}) is an affine function of a single AR(1) state variable

$$-m_{t+1} = x_t + \lambda\sigma\varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0,1) \text{ and} \tag{9.4}$$

$$x_{t+1} = (1-\rho)\,\mu + \rho x_t + \sigma\varepsilon_{t+1}. \tag{9.5}$$

To extend to a multifactor model, specify

$$-m_{t+1} = \mathbf{1}'x_t + \lambda' S\varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0, I) \text{ and} \tag{9.6}$$

$$x_{t+1} = (I-\Psi)\,\mu + \Psi x_t + S\varepsilon_{t+1}, \tag{9.7}$$

where S and Ψ are matrices while λ and μ are (column) vectors; and $\mathbf{1}$ is a vector of ones.

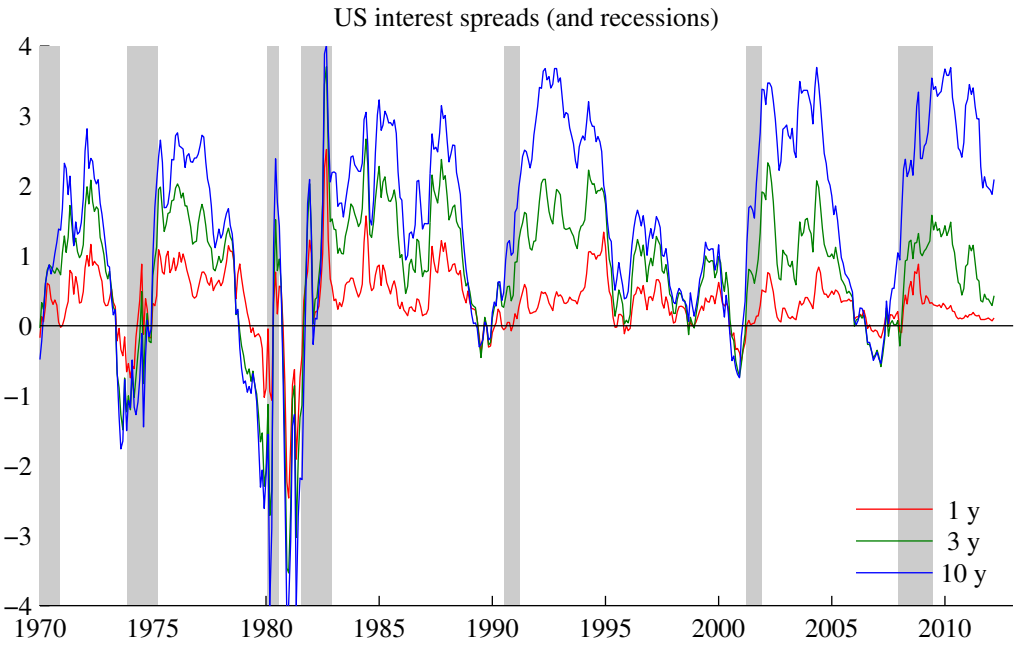


Figure 9.5: US term spreads (over a 3m T-bill)

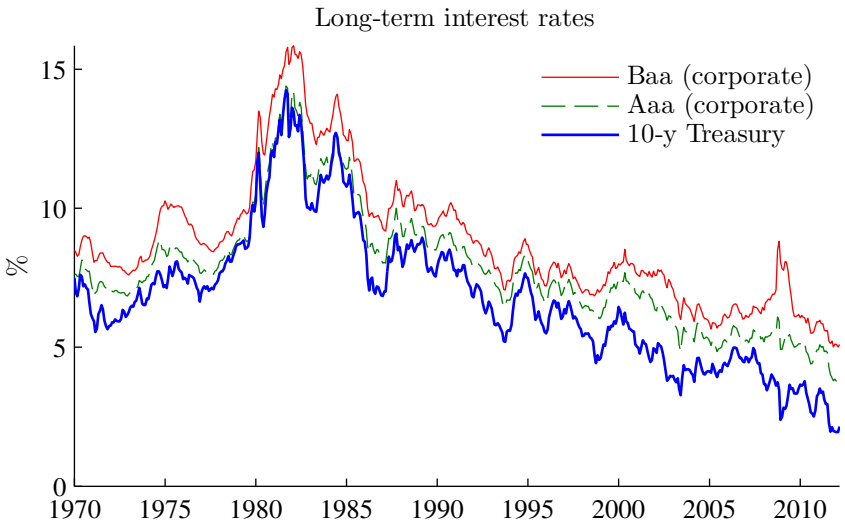


Figure 9.6: US interest rates

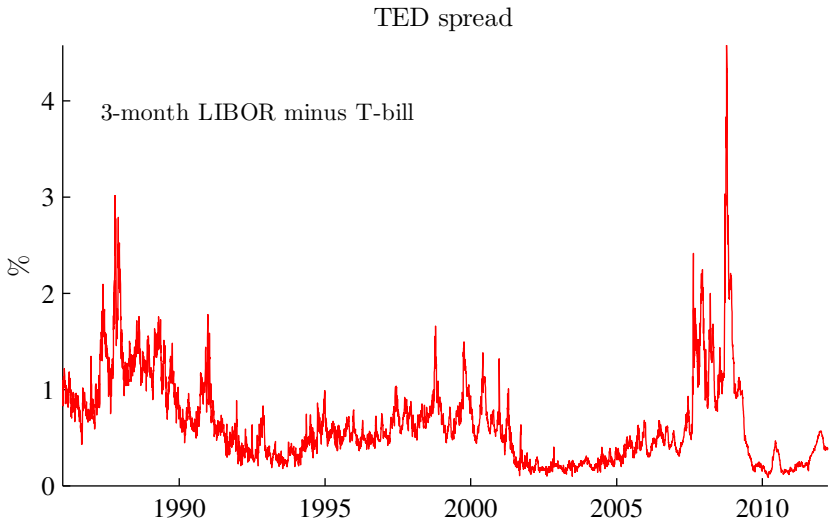


Figure 9.7: TED spread

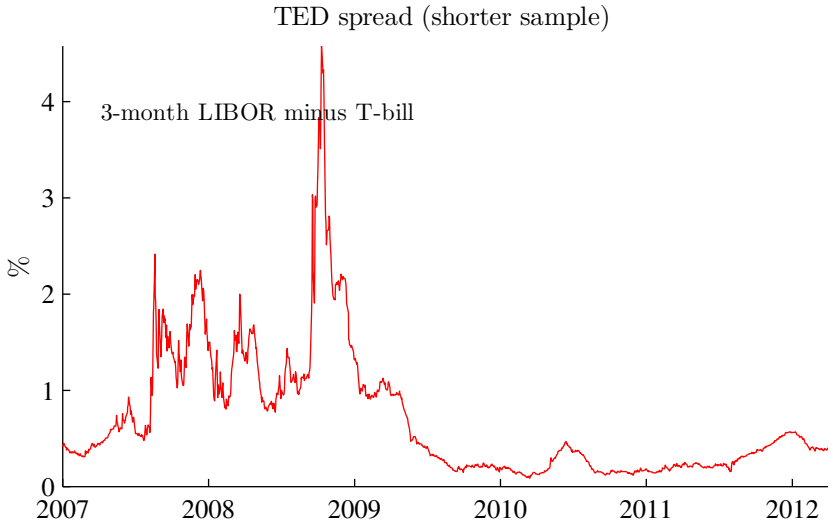


Figure 9.8: TED spread recently

For the single-factor Vasicek model the coefficients in (9.2) can be shown to be

$$B_n = 1 + B_{n-1}\rho \text{ and} \tag{9.8}$$

$$A_n = A_{n-1} + B_{n-1}(1-\rho)\mu - (\lambda + B_{n-1})^2 \sigma^2/2, \tag{9.9}$$

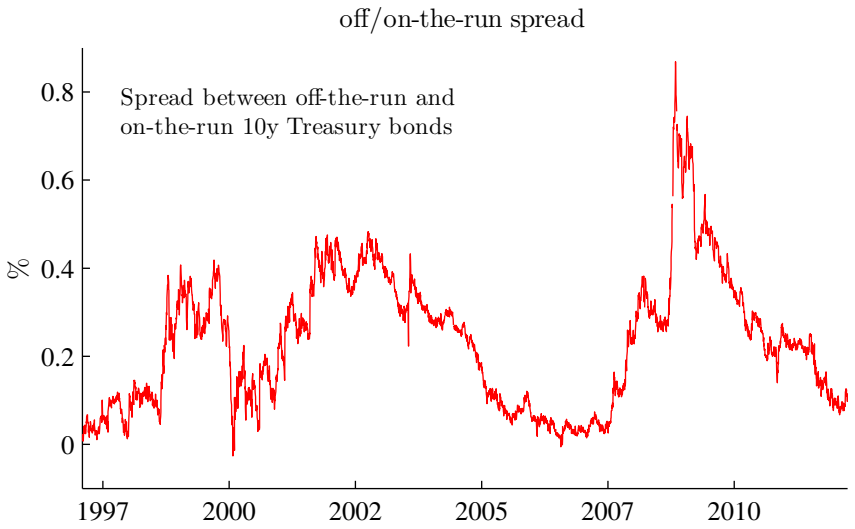


Figure 9.9: Off-the-run liquidity premium

where the recursion starts at $B_0 = 0$ and $A_0 = 0$. For the multivariate version we have

$$B_n = \mathbf{1} + \Psi' B_{n-1}, \text{ and} \tag{9.10}$$

$$A_n = A_{n-1} + B_{n-1}' (I - \Psi) \mu - \left(\lambda' + B_{n-1}'\right) S S' (\lambda + B_{n-1}) / 2, \tag{9.11}$$

where the recursion starts at $B_0 = 0$ and $A_0 = 0$. Clearly, A_n is a scalar and B_n is a $K \times 1$ vector.

See Figure 9.10 for an illustration.

The univariate *CIR model* (Cox, Ingersoll, and Ross (1985)) is

$$-m_{t+1} = x_t + \lambda \sqrt{x_t} \sigma \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0,1) \text{ and} \tag{9.12}$$

$$x_{t+1} = (1 - \rho)\mu + \rho x_t + \sqrt{x_t} \sigma \varepsilon_{t+1} \tag{9.13}$$

and its multivariate version is

$$-m_{t+1} = \mathbf{1}' x_t + \lambda' S \text{diag}(\sqrt{x_t}) \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0, I), \tag{9.14}$$

$$x_{t+1} = (I - \Psi) \mu + \Psi x_t + S \text{diag}(\sqrt{x_t}) \varepsilon_{t+1}. \tag{9.15}$$

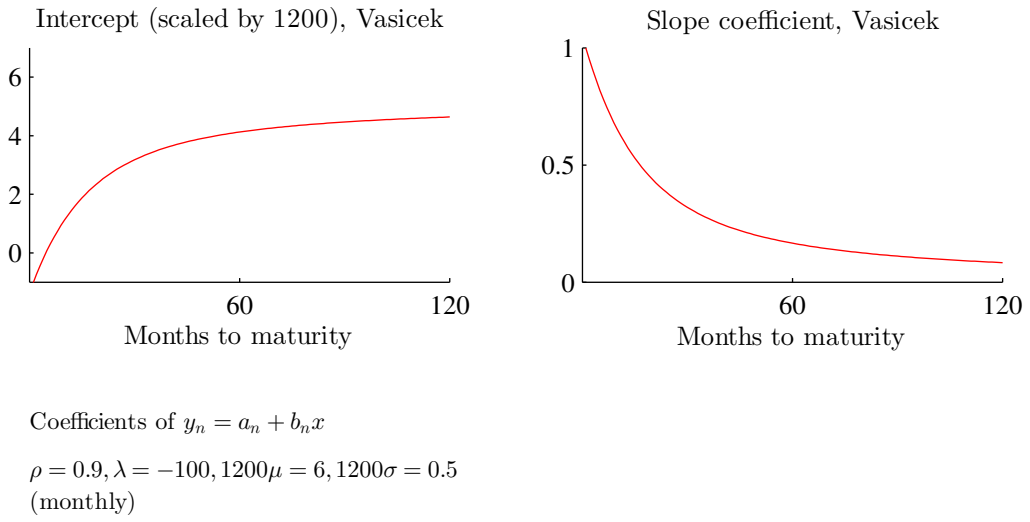


Figure 9.10: a_n and b_n in the Vasicek model

For these models, the coefficients are

$$B_n = 1 + B_{n-1}\rho - (\lambda + B_{n-1})^2 \sigma^2/2 \text{ and} \tag{9.16}$$

$$A_n = A_{n-1} + B_{n-1}(1-\rho)\mu, \tag{9.17}$$

and

$$B_n = \mathbf{1} + \Psi' B_{n-1} - \left[\left(\lambda' S + B_{n-1}' S\right) \odot \left(\lambda' S + B_{n-1}' S\right)\right]'/2, \text{ and} \tag{9.18}$$

$$A_n = A_{n-1} + B_{n-1}'(I - \Psi)\mu, \tag{9.19}$$

where the recursion starts at $B_0 = 0$ and $A_0 = 0$. In (9.18), $\odot$ denotes elementwise multiplication (the Hadamard product).

A model with *affine market price of risk* defines the log SDF in terms of the short rate (y_{1t}) and an innovation to the SDF (χ_{t+1}) as

$$\begin{aligned} y_{1t} &= a_1 + b_1' x_t, \\ -m_{t+1} &= y_{1t} - \chi_{t+1}, \\ \chi_{t+1} &= -\theta_t'\theta_t/2 - \theta_t'\varepsilon_{t+1}, \text{ with } \varepsilon_{t+1} \sim N(0, I). \end{aligned} \tag{9.20}$$

The $K \times 1$ vector of market prices of risk (θ_t) is affine in the state vector

$$\theta_t = \theta^0 + \theta^1 x_t, \tag{9.21}$$

where θ^0 is a $K \times 1$ vector of parameters and θ^1 is $K \times K$ matrix of parameters. Finally, the state vector dynamics is the same as in the multivariate Vasicek model (9.7). For this model, the coefficients are

$$B_n' = B_{n-1}' \left(\Psi - S\theta^1\right) + b_1' \tag{9.22}$$

$$A_n = A_{n-1} + B_{n-1}' \left[(I - \Psi)\,\mu - S\theta^0\right] - B_{n-1}' SS' B_{n-1}/2 + a_1. \tag{9.23}$$

where the recursion starts at $B_0 = 0$ and $A_0 = 0$ (or $B_1 = b_1$ and $A_1 = a_1$).

9.4 MLE of Affine Yield Curve Models

The maximum likelihood approach typically "backs out" the unobservable factors from the yields—by either assuming that some of the yields are observed without any errors or by applying a filtering approach.

9.4.1 Backing out Factors from Yields without Errors

We assume that K yields (as many as there are factors) are observed without any errors—these can be used in place of the state vector. Put the perfectly observed yields in the vector y_{ot} and stack the factor model for these yields—and do the same for the J yields (times maturity) with errors ("unobserved"), y_{ut},

$$y_{ot} = a_o + b_o' x_t \text{ so } x_t = b_o'^{-1} \left(y_{ot} - a_o\right), \text{ and} \tag{9.24}$$

$$y_{ut} = a_u + b_u' x_t + \epsilon_t \tag{9.25}$$

where ϵ_t are the measurement errors. The vector a_o and matrix b_o stacks the a_n and b_n for the perfectly observed yields; a_u and b_u for the yields that are observed with measurement errors (u of "unobserved", although that is something of a misnomer). Clearly, the a vectors and b matrices depend on the parameters of the model, and need to be calculated (recursively) for the maturities included in the estimation.

The measurement errors are not easy to interpret: they may include a bit of pure

measurement errors, but they are also likely to pick up model specification errors. It is therefore difficult to know which distribution they have, and whether they are correlated across maturities and time. The perhaps most common (ad hoc) approach is to assume that the errors are iid normally distributed with a diagonal covariance matrix. To the extent that is a false assumption, the MLE approach should perhaps be better thought of as a quasi-MLE.

The estimation clearly relies on assuming rational expectations: the perceived dynamics (which govern who the market values different bonds) is estimated from the actual dynamics of the data. In a sense, the models themselves do not assume rational expectations: we could equally well think of the state dynamics as reflecting what the market participants believed in. However, in the econometrics we estimate this by using the actual dynamics in the historical sample.

Remark 9.3 *(Log likelihood based on normal distribution) The log pdf of an $q \times 1$ vector $z_t \sim N(\mu_t, \Sigma_t)$ is*

$$\ln \text{pdf}(z_t) = -\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_t| - \frac{1}{2}(z_t - \mu_t)'\Sigma_t^{-1}(z_t - \mu_t).$$

Example 9.4 *(Backing out factors) Suppose there are two factor and that y_{1t} and y_{12t} are assumed to be observed without errors and y_{6t} with a measurement error, then (9.24)–(9.25) are*

$$\begin{aligned}
\begin{bmatrix} y_{1t} \\ y_{12t} \end{bmatrix} &= \underbrace{\begin{bmatrix} a_1 \\ a_{12} \end{bmatrix}}_{a_o} + \underbrace{\begin{bmatrix} b_1' \\ b_{12}' \end{bmatrix}}_{b_o'} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \\
&= \begin{bmatrix} a_1 \\ a_{12} \end{bmatrix} + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{12,1} & b_{12,2} \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix}, \text{ and} \\
y_{6t} &= \underbrace{a_6}_{a_u} + \underbrace{b_6'}_{b_u'} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \epsilon_{6t} \\
&= a_6 + \begin{bmatrix} b_{6,1} & b_{6,2} \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \epsilon_{6t}.
\end{aligned}$$

Remark 9.5 *(Discrete time models and how to quote interest rates) In a discrete time model, it is often convenient to define the period length according to which maturities*

we want to analyze. For instance, with data on 1-month, 3-month, and 4 year rates, it is convenient to let the period length be one month. The (continuously compounded) interest rate data are then scaled by 1/12.

Remark 9.6 *(Data on coupon bonds) The estimation of yield curve models is typically done on data for spot interest rates (yields on zero coupon bonds). The reason is that coupon bond prices (and yield to maturities) are not exponentially affine in the state vector. To see that, notice that a bond that pays coupons in period 1 and 2 has the price* $P_2^c = cP_1 + (1+c)P_2 = c\exp(-A_1 - B_1'x_t) + (1+c)\exp(-A_2 - B_2'x_t)$. *However, this is not difficult to handle. For instance, the likelihood function could be expressed in terms of the log bond prices divided by the maturity (a quick approximate "yield"), or perhaps in terms of the yield to maturity.*

Remark 9.7 *(Filtering out the state vector) If we are unwilling to assume that we have enough yields without observation errors, then the "backing out" approach does not work. Instead, the estimation problem is embedded into a Kalman filter that treats the states are unobservable. In this case, the state dynamics is combined with measurement equations (expressing the yields as affine functions of the states plus errors). The Kalman filter is a convenient way to construct the likelihood function (when errors are normally distributed). See de Jong (2000) for an example.*

Remark 9.8 *(GMM estimation) Instead of using MLE, the model can also be estimated by GMM. The moment conditions could be the unconditional volatilities, autocorrelations and covariances of the yields. Alternatively, they could be conditional moments (conditional on the current state vector), which are transformed into moment conditions by multiplying by some instruments (for instance, the current state vector). See, for instance, Chan, Karolyi, Longstaff, and Sanders (1992) for an early example—which is discussed in Section 9.5.4.*

9.4.2 Adding Explicit Factors*

Assume that we have data on K_F factors, F_t. We then only have to assume that $K_y = K - K_F$ yields are observed without errors. Instead of (9.24) we then have

$$\underbrace{\begin{bmatrix} y_{ot} \\ F_t \end{bmatrix}}_{\tilde{y}_{ot}} = \underbrace{\begin{bmatrix} a_o \\ \mathbf{0}_{K_F \times 1} \end{bmatrix}}_{\tilde{a}_0} + \underbrace{\begin{bmatrix} b_o' \\ \begin{bmatrix} \mathbf{0}_{K_F \times K_y} & I_{K_F} \end{bmatrix} \end{bmatrix}}_{\tilde{b}_0} x_t \text{ so } x_t = \tilde{b}_o'^{-1} (\tilde{y}_{ot} - \tilde{a}_o) . \tag{9.26}$$

Clearly, the last K_F elements of x_t are identical to F_t.

Example 9.9 *(Some explicit and some implicit factors) Suppose there are three factors and that y_{1t} and y_{12t} are assumed to be observed without errors and F_t is a (scalar) explicit factor. Then (9.26) is*

$$\begin{aligned} \begin{bmatrix} y_{1t} \\ y_{12t} \\ F_t \end{bmatrix} &= \begin{bmatrix} a_1 \\ a_{12} \\ 0 \end{bmatrix} + \begin{bmatrix} b_1' \\ b_{12}' \\ [0,0,1] \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix} \\ &= \begin{bmatrix} a_1 \\ a_{12} \\ 0 \end{bmatrix} + \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{12,1} & b_{12,2} & b_{12,3} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix} \end{aligned}$$

Clearly, $x_{3t} = F_t$.

9.4.3 A Pure Time Series Approach

Reference: Chan, Karolyi, Longstaff, and Sanders (1992), Dahlquist (1996)

In a single-factor model, we could invert the relation between (say) a short interest rate and the factor (assuming no measurement errors)—and then estimate the model parameters from the time series of this yield. The data on the other maturities are then not used. This can, in principle, also be used to estimate a multi-factor model, although it may then be difficult to identify the parameters.

The approach is to maximize the likelihood function

$$\ln \mathcal{L}_o = \sum_{t=1}^{T} \ln L_{ot}, \text{ with } \ln L_{ot} = \ln \text{pdf}(y_{ot} | y_{o,t-1}). \tag{9.27}$$

Notice that the relation between x_t and y_{ot} in (9.24) is continuous and invertible, so a

density function of x_t immediately gives the density function of y_{ot}. In particular, with a multivariate normal distribution $x_t|x_{t-1} \sim N\left[\mathrm{E}_{t-1}\, x_t, \mathrm{Cov}_{t-1}\,(x_t)\right]$ we have

$$y_{ot}|y_{o,t-1} \sim N\left[\underbrace{a_o + b_o' \,\mathrm{E}_{t-1}\, x_t}_{\mathrm{E}_{t-1}\, y_{ot}}, \underbrace{b_o' \,\mathrm{Cov}_{t-1}\,(x_t)\, b_o}_{\mathrm{Var}_{t-1}(y_{ot})}\right], \text{ with} \tag{9.28}$$

$$x_t = b_o'^{-1}\,(y_{ot} - a_o)\,.$$

To calculate this expression, we must use the relevant expressions for the conditional mean and covariance.

See *Figure 9.11* for an illustration.

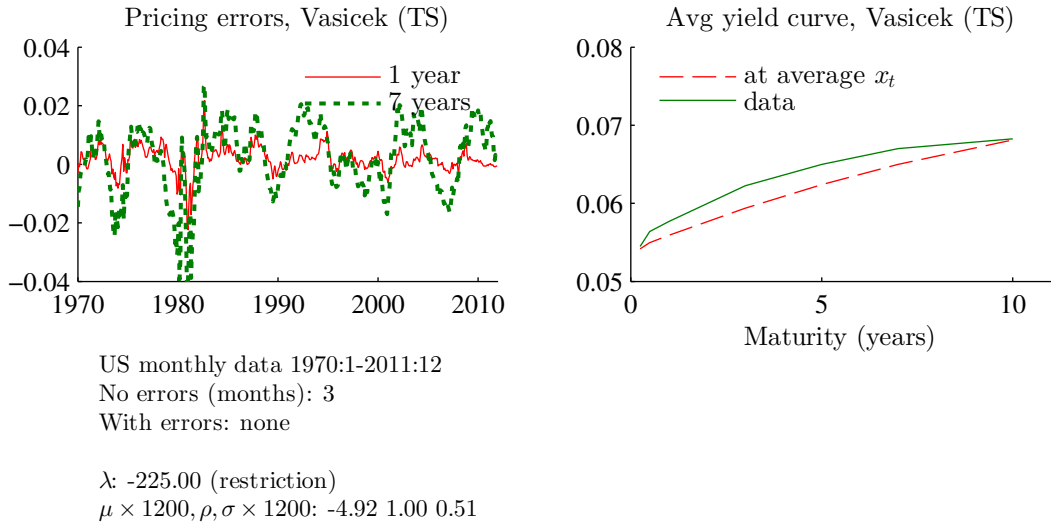


Figure 9.11: Estimation of Vasicek model, time-series approach

Example 9.10 *(Time series estimation of the Vasicek model) In the Vasicek model,*

$$-m_{t+1} = x_t + \lambda\sigma\varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0,1) \text{ and}$$
$$x_{t+1} = (1-\rho)\,\mu + \rho x_t + \sigma\varepsilon_{t+1}$$

we have the 1-period interest rate

$$y_{1t} = -\lambda^2\sigma^2/2 + x_t.$$

The distribution of x_t conditional on x_{t-1} is

$$x_t|x_{t-1} \sim N\left[(1-\rho)\,\mu + \rho x_t, \sigma^2\right].$$

Similarly, the distribution of y_{1t} conditional on $y_{1,t-1}$ is

$$\begin{aligned} y_{1t}|y_{1,t-1} &\sim N\left\{a_1 + b_1[(1-\rho)\,\mu + \rho x_t], b_1\sigma^2 b_1\right\} \text{ with} \\ a_1 &= -\lambda^2\sigma^2/2, b_1 = 1,\ \mathrm{E}_{t-1}\, x_t = (1-\rho)\mu + \rho x_{t-1}. \end{aligned}$$

Inverting the short rate equation (compare with (9.24)) gives

$$x_t = y_{1t} + \lambda^2\sigma^2/2.$$

Combining gives

$$y_{1t}|y_{1,t-1} \sim N\left[(1-\rho)(\mu - \lambda^2\sigma^2/2) + \rho y_{1,t-1}, \sigma^2\right].$$

This can also be written as an AR(1)

$$y_{1t} = (1-\rho)(\mu - \lambda^2\sigma^2/2) + \rho y_{1,t-1} + \sigma\varepsilon_t.$$

Clearly, we can estimate an intercept, ρ, and σ^2 from this relation (with LS or ML), so it is not possible to identify μ and λ separately. We can therefore set λ to an arbitrary value. For instance, we can use λ to fit the average of a long interest rate. The other parameters are estimated to fit the time series behaviour of the short rate only.

Remark 9.11 *(Slope of yield curve when $\rho = 1$) When $\rho = 1$, then the slope of the yield curve is*

$$y_{nt} - y_{1t} = -\left[\left(1 - 3n + 2n^2\right)/6 + (n-1)\,\lambda\right]\sigma^2/2.$$

(To show this, notice that $b_n = 1$ for all n when $\rho = 1$.) As a starting point for calibrating λ, we could therefore use

$$\lambda_{guess} = \frac{-1}{n-1}\left[\frac{\bar{y}_{nt} - \bar{y}_{1t}}{\sigma^2/2} + \frac{1 - 3n + 2n^2}{6}\right],$$

where $\bar{y}_{nt}$ and $\bar{y}_{1t}$ are the sample means of a long and short rate.

Example 9.12 *(Time series estimation of the CIR model) In the CIR model,*

$$-m_{t+1} = x_t + \lambda\sqrt{x_t}\sigma\varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0,1) \text{ and}$$
$$x_{t+1} = (1-\rho)\mu + \rho x_t + \sqrt{x_t}\sigma\varepsilon_{t+1}$$

we have the short rate

$$y_{1t} = (1 - \lambda^2\sigma^2/2)x_t.$$

The conditional distribution is then

$$y_{1t}|y_{1,t-1} \sim N\left[(1-\lambda^2\sigma^2/2)(1-\rho)\mu + \rho y_{1,t-1}, y_{1,t-1}(1-\lambda^2\sigma^2/2)\sigma^2\right], \text{ that is,}$$
$$y_{1t} = (1-\lambda^2\sigma^2/2)(1-\rho)\mu + \rho y_{1,t-1} + \sqrt{y_{1,t-1}}\sqrt{(1-\lambda^2\sigma^2/2)}\sigma\varepsilon_{t+1},$$

which is a heteroskedastic AR(1)—where the variance of the residual is proportional to $\sqrt{y_{1,t-1}}$. *Once again, not all parameters are identified, so a normalization is necessary, for instance, pick* λ *to fit a long rate. In practice, it may be important to either restrict the parameters so the implied* x_t *is positive (so the variance is), or to replace* x_t *by* $max(x_t, 1e-7)$ *or so in the definition of the variance.*

Example 9.13 *(Empirical results from the Vasicek model, time series estimation) Figure 9.11 reports results from a time series estimation of the Vasicek model: only a (relatively) short interest rate is used. The estimation uses monthly observations of monthly interest rates (that is the usual interest rates/1200). The restriction* $\lambda = -200$ *is imposed (as* λ *is not separately identified by the data), since this allows us to also fit the average 10-year interest rate. The upward sloping (average) yield curve illustrates the kind of risk premia that this model can generate.*

Remark 9.14 *(Likelihood function with explicit factors) In case we have some explicit factors like in (9.26), then (9.24) must be modified as*

$$\tilde{y}_{ot}|\tilde{y}_{o,t-1} \sim N\left[\tilde{a}_o + \tilde{b}_o' \mathrm{E}_{t-1}\, x_t, \tilde{b}_o' \mathrm{Cov}_{t-1}\,(x_t)\,\tilde{b}_o\right], \text{ with } x_t = \tilde{b}_o'^{-1}\,(\tilde{y}_{ot} - \tilde{a}_o)\,.$$

9.4.4 A Pure Cross-Sectional Approach

Reference: Brown and Schaefer (1994)

In this approach, we estimate the parameters by using the cross-sectional information (yields for different maturities).

The approach is to maximize the likelihood function

$$\ln \mathcal{L}_u = \sum_{t=1}^{T} \ln L_{ut}, \text{ with } \ln L_{ut} = \ln \text{pdf}\left(y_{ut}|y_{ot}\right) \tag{9.29}$$

It is common to assume that the measurement errors are iid normal with a zero mean and a diagonal covariance with variances ω_i^2 (often pre-assigned, not estimated)

$$\begin{aligned} y_{ut}|y_{ot} &\sim N\left[\underbrace{a_u + b_u' x_t}_{\text{E}(y_{ut}|y_{ot})}, \underbrace{\text{diag}(\omega_i^2)}_{\text{Var}(y_{ut}|y_{ot})}\right], \text{ with} \\ x_t &= b_o'^{-1}\left(y_{ot} - a_o\right). \end{aligned} \tag{9.30}$$

Under this assumption (normal distribution with a diagonal covariance matrix), maximizing the likelihood function amounts to minimizing the weighted squared errors of the yields

$$\arg\max \ln \mathcal{L}_u = \arg\min \sum_{t=1}^{T} \sum_{n \in u} \left(\frac{y_{nt} - \hat{y}_{nt}}{\omega_i}\right)^2, \tag{9.31}$$

where $\hat{y}_{nt}$ are the fitted yields, and the sum is over all "unobserved" yields. In some applied work, the model is reestimated on every date. This is clearly not model consistent—since the model (and the expectations embedded in the long rates) is based on constant parameters.

See *Figure 9.12* for an illustration.

Example 9.15 *(Cross-sectional likelihood for the Vasicek model) In the Vasicek model in Example 9.10, the two-period rate is*

$$y_{2t} = (1-\rho)\,\mu/2 + (1+\rho)x_t/2 - \left[\lambda^2 + (1+\lambda)^2\right]\sigma^2/4.$$

The pdf of y_{2t}, conditional on y_{1t}, is therefore

$$\begin{aligned} &y_{2t}|y_{1t} \sim N(a_2 + b_2 x_t, \omega^2), \text{ with } x_t = y_{1t} + \lambda^2\sigma^2/2, \text{ where} \\ &\quad b_2 = (1+\rho)/2 \text{ and } a_2 = (1-\rho)\,\mu/2 - \left[\lambda^2 + (1+\lambda)^2\right]\sigma^2/4. \end{aligned}$$

Clearly, with only one interest rate (y_{2t}) we can only estimate one parameter, so we need a larger cross section. However, even with a larger cross-section there are serious identification problems. The ρ parameter is well identified from how the entire yield curve

typically move in tandem with y_{ot}. However, μ, σ^2, and λ can all be tweaked to generate a sloping yield curve. For instance, a very high mean μ will make it look as if we are (even on average) below the mean, so the yield curve will be upward sloping. Similarly, both a very negative value of λ (essentially the negative of the price of risk) and a high volatility (risk), will give large risk premia—especially for longer maturities. In practice, it seems as if only one of the parameters μ, σ^2, and λ is well identified in the cross-sectional approach.

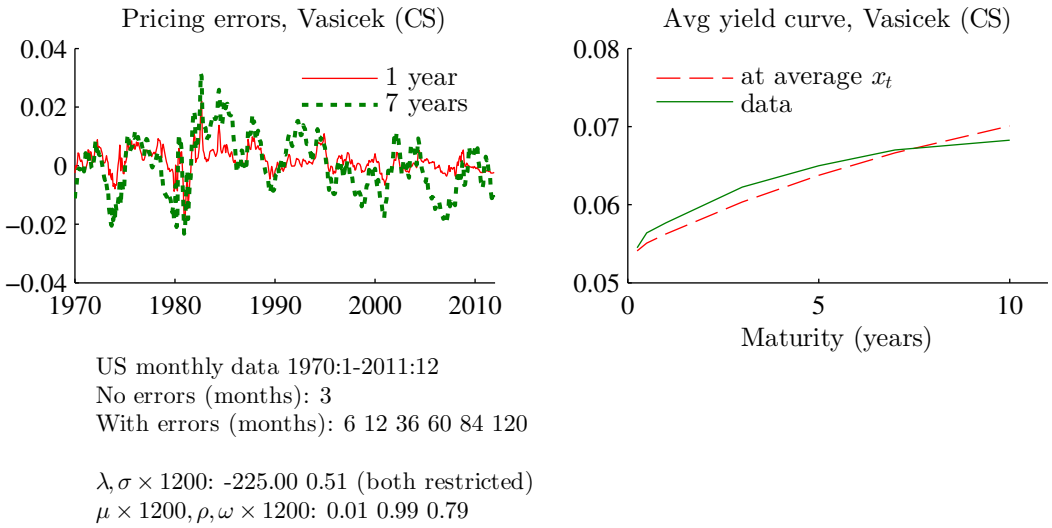


Figure 9.12: Estimation of Vasicek model, cross-sectional approach

Example 9.16 *(Empirical results from the Vasicek model, cross-sectional estimation) Figure 9.12 reports results from a cross-sectional estimation of the Vasicek model, where it is assumed that the variances of the observation errors (ω_i^2) are the same across yields. The estimation uses monthly observations of monthly interest rates (that is the usual interest rates/1200). The values of μ and σ^2 are restricted to the values obtained in the time series estimations, so only ρ and λ are estimated. Choosing other values for μ and σ^2 gives different estimates of λ, but still the same yield curve (at least on average).*

9.4.5 Combined Time Series and Cross-Sectional Approach

Reference: Duffee (2002)

The approach here combines the time series and cross-sectional methods—in order to fit the whole model on the whole sample (all maturities, all observations). This is the full maximum likelihood, since it uses all available information.

The log likelihood function is

$$\ln \mathcal{L} = \sum_{t=1}^{T} \ln L_t, \text{ with } \ln L_t = \ln \text{pdf}\left(y_{ut}, y_{ot} | y_{o,t-1}\right). \tag{9.32}$$

Notice that the joint density of (y_{ut}, y_{ot}), conditional on $y_{o,t-1}$ can be split up as

$$\text{pdf}\left(y_{ut}, y_{ot} | y_{ot-1}\right) = \text{pdf}\left(y_{ut} | y_{ot}\right) \text{pdf}(y_{ot} | y_{o,t-1}), \tag{9.33}$$

since $y_{o,t-1}$ does not affect the distribution of y_{ut} conditional on y_{ot}. Taking logs gives

$$\ln L_t = \ln \text{pdf}\left(y_{ut} | y_{ot}\right) + \ln \text{pdf}(y_{ot} | y_{o,t-1}). \tag{9.34}$$

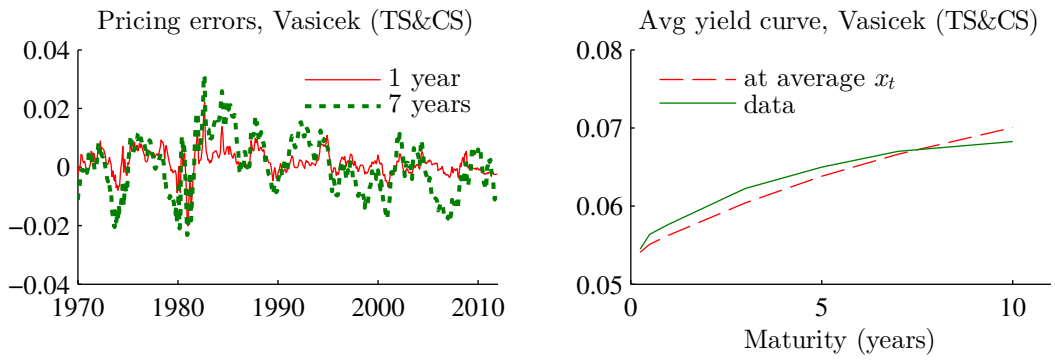
The first term is the same as in the cross-sectional estimation and the second is the same as in the time series estimation. The log likelihood (9.32) is therefore just the sum of the log likelihoods from the pure cross-sectional and the pure time series estimations

$$\ln \mathcal{L} = \sum_{t=1}^{T} \ln L_{ut} + \ln L_{ot}. \tag{9.35}$$

See *Figures 9.13–9.17* for illustrations. Notice that the variances of the observation errors (ω_i^2) are important for the relative "weight" of the contribution from the time series and cross-sectional parts.

Example 9.17 *(MLE of the Vasicek Model) Consider the Vasicek model, where we observe y_{1t} without errors and y_{2t} with measurement errors. The likelihood function is then the sum of the log pdfs in Examples 9.10 and 9.15, except that the cross-sectional part must be include the variance of the observation errors (ω^2) which is assumed to be equal across maturities.*

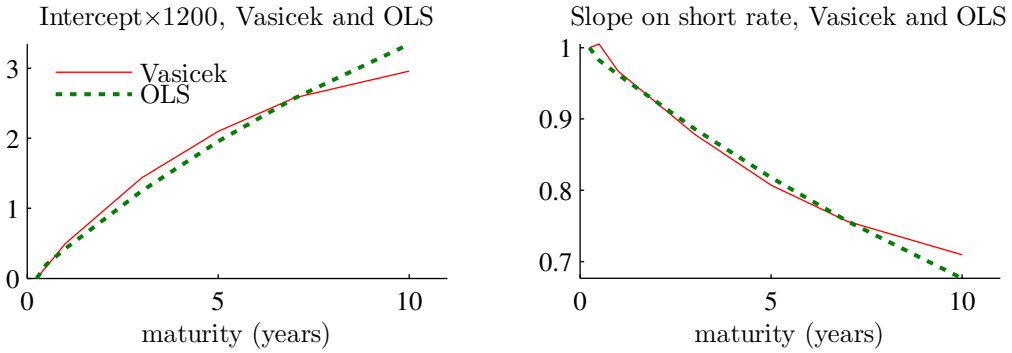
Example 9.18 *(Empirical results from the Vasicek model, combined time series and cross-sectional estimation) Figure 9.13 reports results from a combined time series and cross-sectional estimation of the Vasicek model. The estimation uses monthly observations of monthly interest rates (that is the usual interest rates/1200). All model parameters*



US monthly data 1970:1-2011:12
No errors (months): 3
With errors (months): 6 12 36 60 84 120

$\lambda, \mu \times 1200, \rho, \sigma \times 1200, \omega \times 1200$: -241.59 10.10 0.99 0.53 0.79

Figure 9.13: Estimation of Vasicek model, combined time series and cross-sectional approach



The figures show the relation $y_n = \alpha_n + \beta_n y_1$
US monthly data 1970:1-2011:12

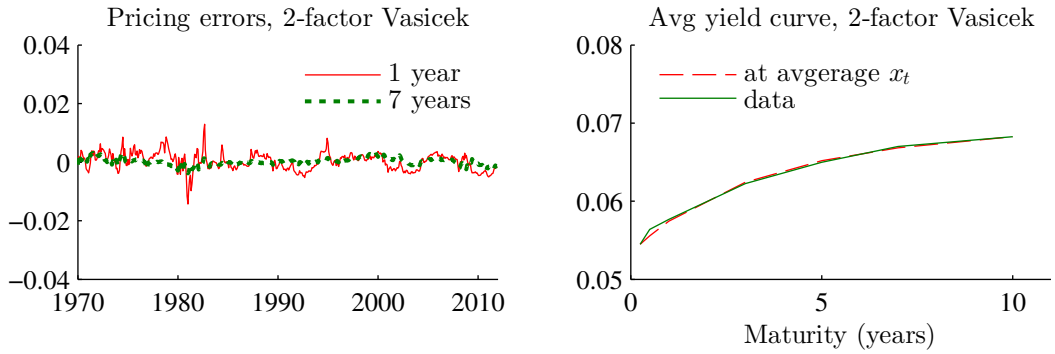
The Vasicek model is estimated with ML (TS&CS),
while OLS is a time-series regression for each maturity

Figure 9.14: Loadings in a one-factor model: LS and Vasicek

$(\lambda, \mu, \rho, \sigma^2)$ are estimated, along with the variance of the measurement errors. (All measurement errors are assumed to have the same variances, ω.) Figure 9.14 reports the loadings on the constant and the short rate according to the Vasicek model and (unrestricted) OLS. The patterns are fairly similar, suggesting that the cross-equation (-maturity) re-

strictions imposed by the Vasicek model are not at great odds with data.

Remark 9.19 *(Imposing a unit root) If a factor appears to have a unit root, it may be easier to impose this on the estimation. This factor then causes parallel shifts of the yield curve—and makes the yields being cointegrated. Imposing the unit root leads the estimation being effectively based on the changes of the factor, so standard econometric techniques can be applied. See Figure 9.16 for an example.*

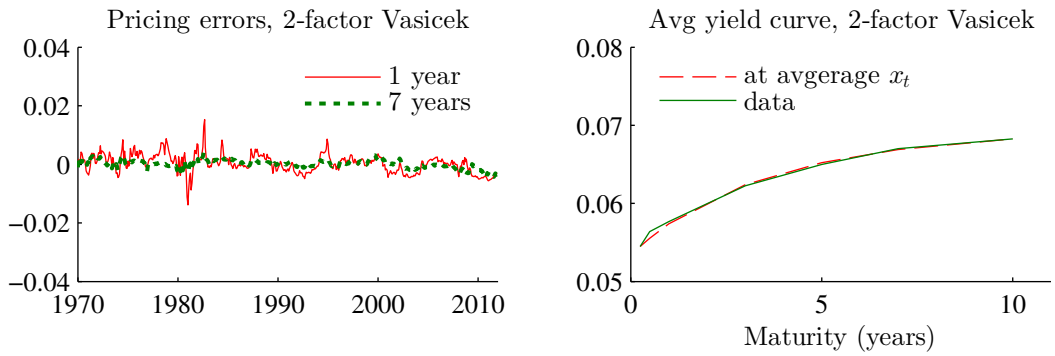


US monthly data 1970:1-2011:12
No errors (months): 3 120
With errors (months): 6 12 36 60 84

$\gamma_1 = \gamma_2 = 1$ (restriction)
λ_1 and λ_2: -190.08 -278.67
$\mu_1 = 0$ (restriction) and $\mu_2 \times 1200$: 13.93
ρ_1 and ρ_2: 1.00 0.95
S_1 and $S_2(\times 1200)$: 0.39 0.55
$\omega(\times 1200)$: 0.24

Figure 9.15: Estimation of 2-factor Vasicek model, time-series&cross-section approach

Example 9.20 *(Empirical results from a two-factor Vasicek model) Figure 9.15 reports results from a two-factor Vasicek model. The estimation uses monthly observations of monthly interest rates (that is the usual interest rates/1200). We can only identify the mean of the SDF, not whether if it is due to factor one or two. Hence, I restrict* $\mu_2 = 0$. *The results indicate that there is one very persistent factor (affecting the yield curve level), and another slightly less persistent factor (affecting the yield curve slope). The "price of risk" is larger (*λ_i *more negative) for the more persistent factor. This means that the risk premia will scale almost linearly with the maturity. As a practical matter, it turned out*



US monthly data 1970:1-2011:12
No errors (months): 3 120
With errors (months): 6 12 36 60 84

$\gamma_1 = \gamma_2 = 1$ (restriction)
λ_1 and λ_2: -82.40 -316.45
$\mu_1 = 0$ (restriction) and $\mu_2 \times 1200$: -15.35
ρ_1 and ρ_2: 1.00 0.96
S_1 and $S_2(\times 1200)$: 0.31 0.54
$\omega(\times 1200)$: 0.27

Notice: ρ_1 is restricted to unity

Figure 9.16: Estimation of 2-factor Vasicek model, time-series&cross-section approach, $\rho_1 = 1$ is imposed

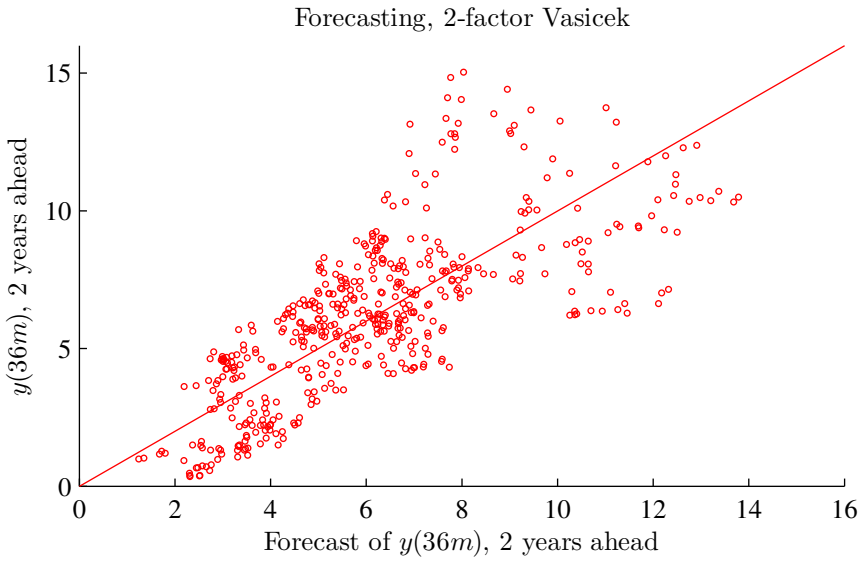


Figure 9.17: Forecasting properties of estimated of 2-factor Vasicek model

that a derivative-free method (fminsearch in MatLab) worked much better than standard optimization routines. The pricing errors are clearly smaller than in a one-factor Vasicek model. Figure 9.17 illustrates the forecasting performance of the model by showing scatter plots of predicted yields and future realized yields. An unbiased forecasting model should have the points scattered (randomly) around a 45 degree line. There are indications that the really high forecasts (above 10%, say) are biased: they are almost always followed be realized rates below 10%. A standard interpretations would be that the model underestimates risk premia (overestimates expected future rates) when the current rates are high. I prefer to think of this as a shift in monetary policy regime: all the really high forecasts are done during the Volcker deflation—which was surprisingly successful in bringing down inflation. Hence, yields never became that high again. The experience from the optimization suggests that the objective function has some flat parts.

9.5 Summary of Some Empirical Findings

9.5.1 Term Premia and Interest Rate Forecasts in Affine Models by Duffee (2002)

Reference: Duffee (2002)

This paper estimates several affine and "essentially affine" models on monthly data 1952–1994 on US zero-coupon interest rates, using a combined time series and cross-sectional approach. The data for 1995–1998 are used for evaluating the out-of-sample forecasts of the model. The likelihood function is constructed by assuming normally distributed errors, but this is interpreted as a quasi maximum likelihood approach. All the estimated models have three factors. A fairly involved optimization routine is needed in order to keep the parameters such that variances are always positive.

The models are used to forecast yields (3, 6, and 12 months) ahead, and then evaluated against the actual yields. It is found that a simple random walk beats the affine models in forecasting the yields. The forecast errors tend to be negatively correlated with the slope of the term structure: with a steep slope of the yield curve, the affine models produce too high forecasts. (The models are closer to the expectations hypothesis than data is.) The essentially affine model produce much better forecasts. (The essentially affine models extend the affine models by allowing the market price of risk to be linear functions of the state vector.)

9.5.2 "A Joint Econometric Model of Macroeconomic and Term Structure Dynamics" by Hördahl et al (2005)

Reference: Hördahl, Tristiani, and Vestin (2006), Ang and Piazzesi (2003)

This paper estimates both an affine yield curve model and a macroeconomic model on monthly German data 1975–1998.

To identify the model, the authors put a number of restrictions on the θ_1 matrix. In particular, the lagged variables in x_t are assumed to have no effect on θ_t.

The key distinguishing feature of this paper is that a macro model (for inflation, output, and the policy for the short interest rate) is estimated jointly with the yield curve model. (In contrast, Ang and Piazzesi (2003) estimate the macro model separately.) In this case, the unobservable factors include variables that affect both yields and the macro variables (for instance, the time-varying inflation target). Conversely, the observable data includes not only yields, but also macro variables (output, inflation). It is found, among other things, that the time-varying inflation target has a crucial effect on yields and that bond risk premia are affected both by policy shocks (both to the short-run policy rule and to the inflation target), as well as the business cycle shocks.

9.5.3 The Diebold-Li Approach

Diebold and Li (2006) use the Nelson-Siegel model for an m-period interest rate as

$$y(m) = \beta_0 1 + \beta_1 \frac{1 - \exp(-m/\tau_1)}{m/\tau_1} + \beta_2 \left[\frac{1 - \exp(-m/\tau_1)}{m/\tau_1} - \exp\left(-\frac{m}{\tau_1}\right) \right], \quad (9.36)$$

and set $\tau_1 = 1/(12 \times 0.0609)$. Their approach is as follows. For a given trading date, construct the factors (the terms multiplying the beta coefficients) for each bond. Then, run a regression of the cross-section of yields on these factors—to estimate the beta coefficients. Repeat this for every trading day—and plot the three time series of the coefficients.

See Figure 9.18 for an example. The results are very similar to the factors calculated directly from yields (cf. Figure 9.3).

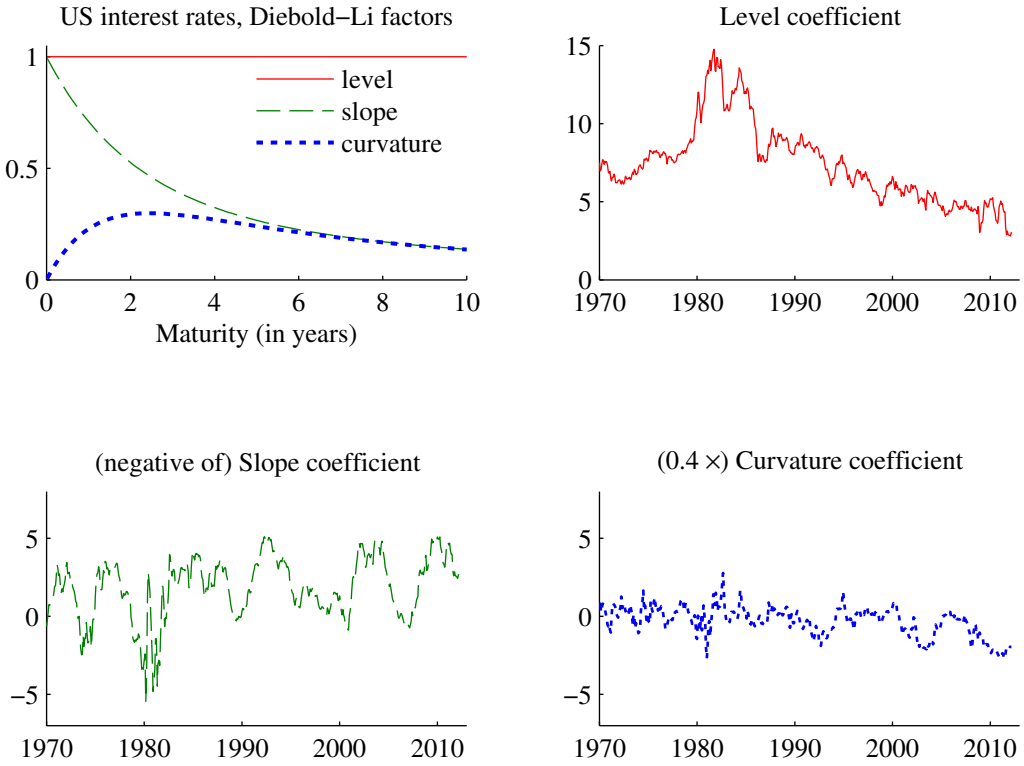


Figure 9.18: US yield curves: level, slope and curvature, Diebold-Li approach

9.5.4 "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate" by Chan et al (1992)

Reference: Chan, Karolyi, Longstaff, and Sanders (1992) (CKLS), Dahlquist (1996)

This paper focuses on the dynamics of the short rate process. The models that CKLS study have the following dynamics (under the natural/physical distribution) of the one-period interest rate, y_{1t}

$$y_{1,t+1} - y_{1t} = \alpha + \beta y_{1t} + \varepsilon_{t+1}, \text{ where} \tag{9.37}$$
$$\mathrm{E}_t\, \varepsilon_{t+1} = 0 \text{ and } \mathrm{E}_t\, \varepsilon_{t+1}^2 = \mathrm{Var}_t(\varepsilon_{t+1}) = \sigma^2 y_{1t}^{2\gamma}.$$

This formulation nests several well-known models: $\gamma = 0$ gives a Vasicek model and $\gamma = 1/2$ a CIR model (which are the only cases which will deliver a single-factor affine

model). It is an approximation of the diffusion process

$$dr_t = (\beta_0 + \beta_1 r_t)dt + \sigma r_t^{\gamma} dW_t, \tag{9.38}$$

where W_t is a Wiener process. (For an introduction to the issue of being more careful with estimating a continuous time model on discrete data, see Campbell, Lo, and MacKinlay (1997) 9.3 and Harvey (1989) 9.) In some cases, like the homoskedastic AR(1), there is no approximation error because of the discrete sampling. In other cases, there is an error.)

CKLS estimate the model (9.37) with GMM using the following moment conditions

$$g_t(\alpha, \beta, \gamma, \sigma^2) = \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1}^2 - \sigma^2 y_{1t}^{2\gamma} \end{bmatrix} \otimes \begin{bmatrix} 1 \\ y_{1t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1} y_{1t} \\ \varepsilon_{t+1}^2 - \sigma^2 y_{1t}^{2\gamma} \\ (\varepsilon_{t+1}^2 - \sigma^2 y_{1t}^{2\gamma}) y_{1t} \end{bmatrix}, \tag{9.39}$$

so there are four moment conditions and four parameters (α, β, σ^2, and γ). The choice of the instruments (1 and y_{1t}) is somewhat arbitrary since any variables in the information set in t would do.

CKLS estimate this model in various forms (imposing different restrictions on the parameters) on monthly data on one-month T-bill rates for 1964–1989. They find that both $\hat{\alpha}$ and $\hat{\beta}$ are close to zero (in the unrestricted model $\hat{\beta} < 0$ and almost significantly different from zero—indicating mean-reversion). They also find that $\hat{\gamma} > 1$ and significantly so. This is problematic for the affine one-factor models, since they require $\gamma = 0$ or $\gamma = 1/2$. A word of caution: the estimated parameter values suggest that the interest rate is non-stationary, so the properties of GMM are not really known. In particular, the estimator is probably not asymptotically normally distributed—and the model could easily generate extreme interest rates.

See *Figures 9.19–9.20* for an illustration.

Example 9.21 *(Re-estimating the Chan et al model) Some results obtained from re-estimating the model on a longer data set are found in Figure 9.19. In this figure, $\alpha = \beta = 0$ is imposed, but the results are very similar if this is relaxed. One of the first thing to note is that the loss function is very flat in the $\gamma \times \sigma$ space—the parameters are not pinned down very precisely by the model/data. Another way to see this is to note that the moments in (9.39) are very strongly correlated: moment 1 and 2 have a very strong correlation, and*

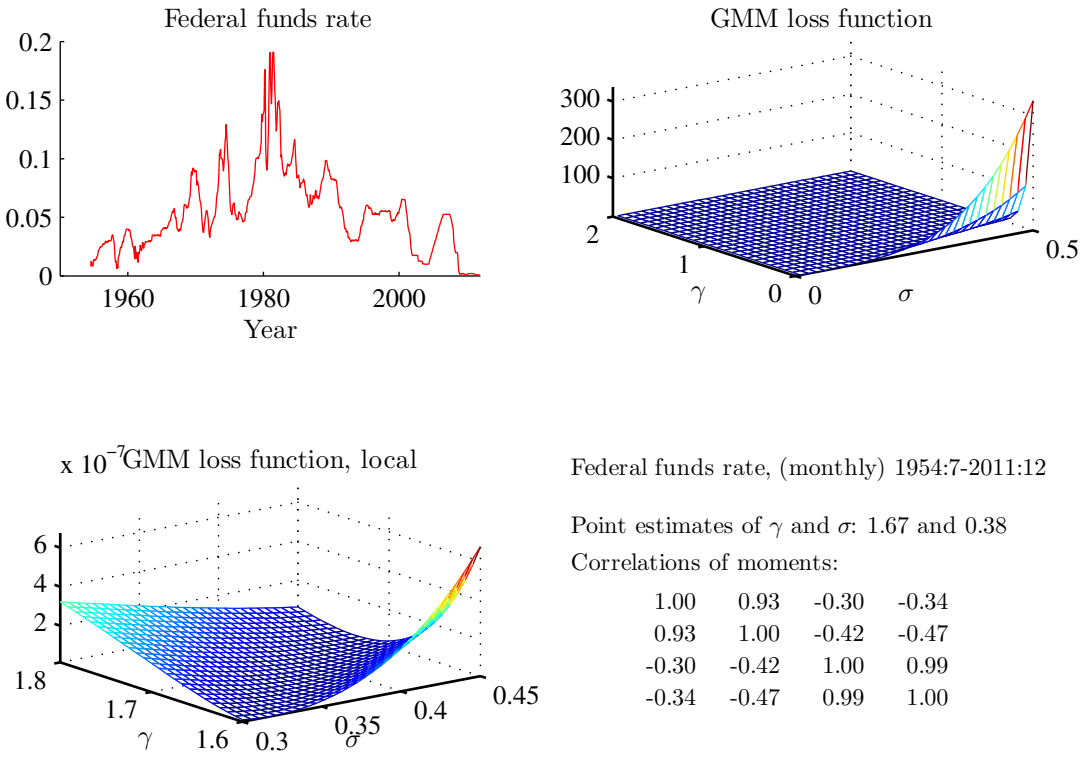
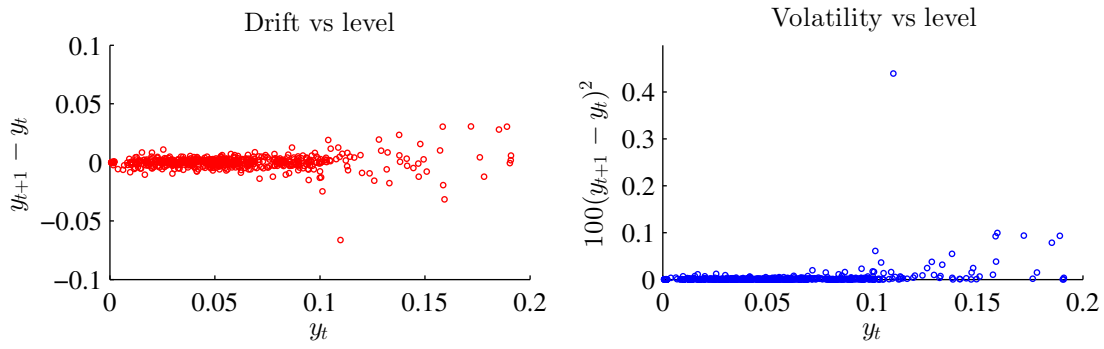


Figure 9.19: Federal funds rate, monthly data, $\alpha = \beta = 0$ imposed

this is even worse for moments 3 and 4. The latter two moment conditions are what identifies σ^2 from γ, so it is a serious problem for the estimation. The reason for these strong correlations is probably that the interest rate series is very persistent so, for instance, ε_{t+1} and $\varepsilon_{t+1} y_{1t}$ look very similar (as y_{1t} tends to be fairly constant due to the persistence). Figure 9.20, which shows cross plots of the interest rate level and the change and volatility in the interest rate, suggests that some of the results might be driven by outliers. There is, for instance, a big volatility outlier in May 1980 and most of the data points with high interest rate and high volatility are probably from the Volcker deflation in the early 1980s. It is unclear if that particular episode can be modelled as belonging to the same regime as the rest of the sample (in particular since the Fed let the interest rate fluctuate a lot more than before). Maybe this episode needs a special treatment.



Federal funds rate, sample: 1954:7-2011:12

Figure 9.20: Federal funds rate, monthly data

Bibliography

Ang, A., and M. Piazzesi, 2003, "A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables," *Journal of Monetary Economics*, 60, 745–787.

Backus, D., S. Foresi, and C. Telmer, 1998, "Discrete-time models of bond pricing," Working Paper 6736, NBER.

Brown, R. H., and S. M. Schaefer, 1994, "The term structure of real interest rates and the Cox, Ingersoll, and Ross model," *Journal of Financial Economics*, 35, 3–42.

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

Chan, K. C., G. A. Karolyi, F. A. Longstaff, and A. B. Sanders, 1992, "An empirical comparison of alternative models of the short-term interest rate," *Journal of Finance*, 47, 1209–1227.

Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.

Cox, J. C., J. E. Ingersoll, and S. A. Ross, 1985, "A theory of the term structure of interest rates," *Econometrica*, 53, 385–407.

Dahlquist, M., 1996, "On alternative interest rate processes," *Journal of Banking and Finance*, 20, 1093–1119.

de Jong, F., 2000, "Time series and cross-section information in affine term-structure models," *Journal of Business and Economic Statistics*, 18, 300–314.

Diebold, F. X., and C. Li, 2006, "Forecasting the term structure of government yields," *Journal of Econometrics*, 130, 337–364.

Duffee, G. R., 2002, "Term premia and interest rate forecasts in affine models," *Journal of Finance*, 57, 405–443.

Harvey, A. C., 1989, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.

Hördahl, P., O. Tristiani, and D. Vestin, 2006, "A joint econometric model of macroeconomic and term structure dynamics," *Journal of Econometrics*, 131, 405–444.

Singleton, K. J., 2006, *Empirical dynamic asset pricing*, Princeton University Press.

10 Yield Curve Models: Nonparametric Estimation

10.1 Nonparametric Regression

Reference: Campbell, Lo, and MacKinlay (1997) 12.3; Härdle (1990); Pagan and Ullah (1999); Mittelhammer, Judge, and Miller (2000) 21

10.1.1 Introduction

Nonparametric 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

$$y_t = b(x_t) + \varepsilon_t, \tag{10.1}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t|x_t) = 0$. The function $b()$ is unknown and possibly non-linear.

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 $b(x_t)$ at a particular x_t value ($x_t = 1.9$, 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 (y) observations where x_t is close to x. The general form of this type of estimator is

$$\hat{b}(x) = \frac{\sum_{t=1}^{T} w_t(x) y_t}{\sum_{t=1}^{T} w_t(x)}, \tag{10.2}$$

where $w_t(x)/\Sigma_{t=1}^{T} w_t(x)$ is the weight on observation t, which his non-negative and (weakly) increasing in the the distance of x_t from x. Note that the denominator makes the weights sum to unity. The basic assumption behind (10.2) is that the $b(x)$ function is smooth so local (around x) averaging makes sense.

Remark 10.1 *(Local constant estimator*) Notice that (10.2) solves the problem* $\min \sum_{t=1}^{T} w_t(x)(y_t - \alpha_x)^2$ *for each value of* x*. (The result is* $\hat{b}(x) = \alpha_x$*.) This is (for each value of* x*) like a weighted regression of* x_t *on a constant. This immediately suggests that the method could be extended to solving a problem like* $\min \sum_{t=1}^{T} w_t(x)[y_t - \alpha_x - b_x(x_t - x)]^2$*, which defines the local linear estimator.*

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 (often used in macroeconomics when $x_t = t$, and is then called the Hodrick-Prescott filter).

Remark 10.2 *(Easy way to calculate the "nearest neighbor" estimator, univariate case) Create a matrix* Z *where row* t *is* (y_t, x_t)*. Sort the rows of* Z *according to the second column* (x)*. Calculate an equally weighted centered moving average of the first column* (y)*.*

10.1.2 Kernel Regression

A *Kernel regression* uses a pdf as the weight function, $w_t(x) = K[(x_t - x)/h]$, where the choice of h (also called bandwidth) allows us to easily vary the relative weights of different observations.

The perhaps simplest choice is a uniform density function for x_t over $x - h/2$ to $x + h/2$ (and zero outside this interval). In this case, the weighting function is

$$w_t(x) = \frac{1}{h}\delta\left(\left|\frac{x_t - x}{h}\right| \leq 1/2\right), \text{ where } \delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases} \tag{10.3}$$

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 a density function that puts more weight to very local information, but also tapers off more smoothly. The pdf of $N(x, h^2)$ is often used for $K()$. This weighting function is positive, so all observations get a positive weight, but the weights are highest for observations close to x and then taper off in a bell-shaped way. A low value of h means that the weights taper off fast.

See Figure 10.1 for an example.

With the $N(x, h^2)$ kernel, we get the following weights at a point x

$$w_t(x) = \frac{\exp\left[-\left(\frac{x_t - x}{h}\right)^2/2\right]}{h\sqrt{2\pi}}. \tag{10.4}$$

Remark 10.3 *(Kernel as a pdf of $N(x, h^2)$) If $K(z)$ is the pdf of an $N(0,1)$ variable, then $K\left[(x_t - x)/h\right]/h$ is the same as using an $N(x, h^2)$ pdf of x_t. Clearly, the $1/h$ term would cancel in (10.2).*

Effectively, we can think of these weights as being calculated from an $N(0,1)$ density function, but where we use $(x_t - x)/h$ as the argument.

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 the maximum values observed in the sample. 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 Figure 10.2 for an illustration.

Example 10.4 *(Kernel regression) 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 (10.4) is*

$$\begin{aligned}\sum_{t=1}^{T} w_t(x) y_t &= \left(e^{-(1.5-1.9)^2/2} \times 5 + e^{-(2-1.9)^2/2} \times 4 + e^{-(2.5-1.9)^2/2} \times 3.5\right)/\sqrt{2\pi} \\ &\approx (0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5)/\sqrt{2\pi} \\ &= 11.52/\sqrt{2\pi}.\end{aligned}$$

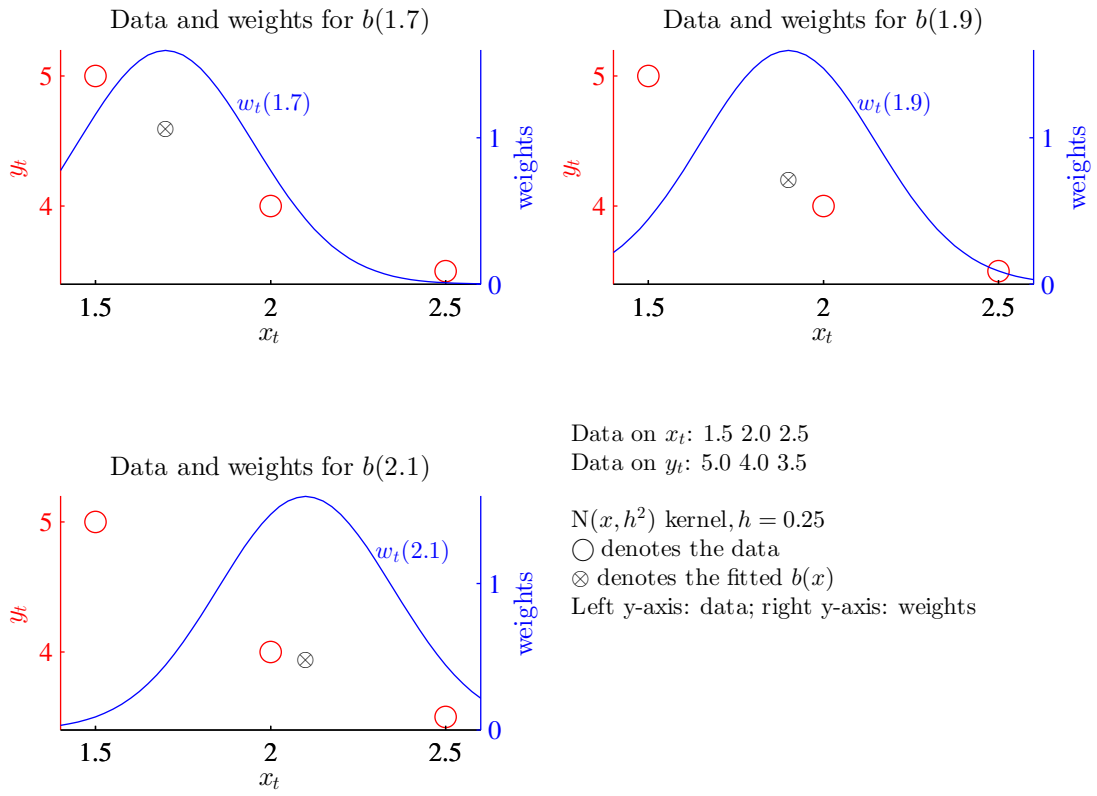


Figure 10.1: Example of kernel regression with three data points

The denominator is

$$\sum_{t=1}^{T} w_t(x) = \left(e^{-(1.5-1.9)^2/2} + e^{-(2-1.9)^2/2} + e^{-(2.5-1.9)^2/2}\right)/\sqrt{2\pi}$$
$$\approx 2.75/\sqrt{2\pi}.$$

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. It can be shown (see Härdle (1990) 3.1 and Pagan and Ullah (1999) 3.3–4) that under the assumption that x_t is iid, the mean squared error, variance and bias

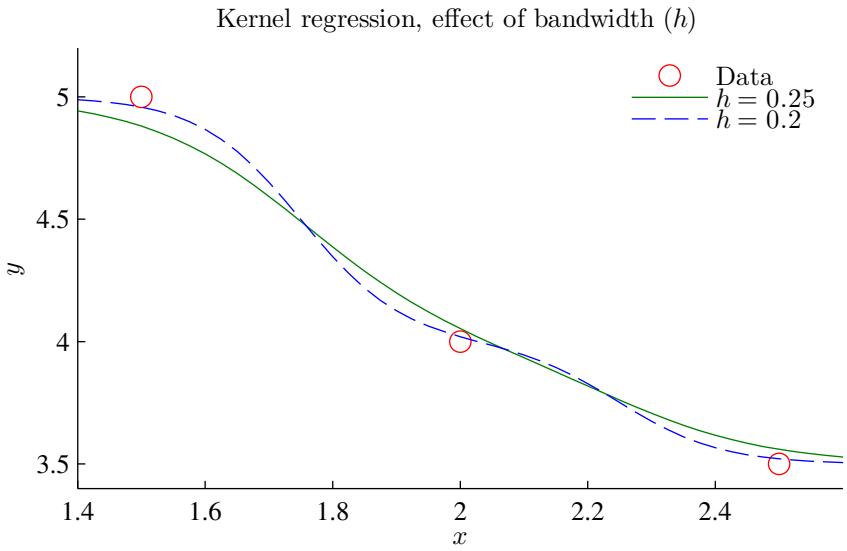


Figure 10.2: Example of kernel regression with three data points

of the estimator at the value x are approximately (for general kernel functions)

$$\begin{aligned}
\text{MSE}(x) &= \text{Var}\left[\hat{b}(x)\right] + \left\{\text{Bias}[\hat{b}(x)]\right\}^2, \text{ with} \\
\text{Var}\left[\hat{b}(x)\right] &= \frac{1}{Th}\frac{\sigma^2(x)}{f(x)} \times \int_{-\infty}^{\infty} K(u)^2 du \\
\text{Bias}[\hat{b}(x)] &= h^2 \times \left[\frac{1}{2}\frac{d^2b(x)}{dx^2} + \frac{df(x)}{dx}\frac{1}{f(x)}\frac{db(x)}{dx}\right] \times \int_{-\infty}^{\infty} K(u)u^2 du. \qquad (10.5)
\end{aligned}$$

In these expressions, $\sigma^2(x)$ is the variance of the residuals in (10.1), $f(x)$ the marginal density of x and $K(u)$ the kernel (pdf) used as a weighting function for $u = (x_t - x)/h$. The remaining terms are functions of either the true regression function.

With a gaussian kernel these expressions can be simplified to

$$\begin{aligned}
\text{Var}\left[\hat{b}(x)\right] &= \frac{1}{Th}\frac{\sigma^2(x)}{f(x)} \times \frac{1}{2\sqrt{\pi}} \\
\text{Bias}[\hat{b}(x)] &= h^2 \times \left[\frac{1}{2}\frac{d^2b(x)}{dx^2} + \frac{df(x)}{dx}\frac{1}{f(x)}\frac{db(x)}{dx}\right]. \qquad (10.6)
\end{aligned}$$

Proof. (of (10.6)) We know that

$$\int_{-\infty}^{\infty} K(u)^2 du = \frac{1}{2\sqrt{\pi}} \text{ and } \int_{-\infty}^{\infty} K(u)u^2 du = 1,$$

if $K(u)$ is the density function of a standard normal distribution. (We are effectively using the $N(0,1)$ pdf for the variable $(x_t - x)/h$.) Use in (10.5). ■

A smaller h increases the variance (we effectively use fewer data points to estimate $b(x)$) but decreases the bias of the estimator (it becomes more local to x). If h decreases less than proportionally with the sample size (so hT in the denominator of the first term increases with T), then the variance goes to zero and the estimator is consistent (since the bias in the second term decreases as h does).

The variance is a function of the variance of the residuals and the "peakedness" of the kernel, but not of the $b(x)$ function. The more concentrated the kernel is ($\int K(u)^2 du$ large) around x (for a given h), the less information is used in forming the average around x, and the uncertainty is therefore larger—which is similar to using a small h. A low density of the regressors ($f(x)$ low) means that we have little data at x which drives up the uncertainty of the estimator.

The bias increases (in magnitude) with the curvature of the $b(x)$ function (that is, $(d^2b(x)/dx^2)^2$). This makes sense, since rapid changes of the slope of $b(x)$ make it hard to get $b(x)$ right by averaging at nearby x values. It also increases with the variance of the kernel since a large kernel variance is similar to a large h.

It is clear that the choice of h has a major importance on the estimation results. A lower value of h means a more "local" averaging, which has the potential of picking up sharp changes in the regression function—at the cost of being more affected by randomness.

See Figures 10.3–10.4 for an example.

A good (but computationally intensive) approach to choose h is by the leave-one-out *cross-validation* technique. This approach would, for instance, choose h to minimize the expected (or average) prediction error

$$\text{EPE}(h) = \sum_{t=1}^{T} \left[y_t - \hat{b}_{-t}(x_t, h) \right]^2 / T, \tag{10.7}$$

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

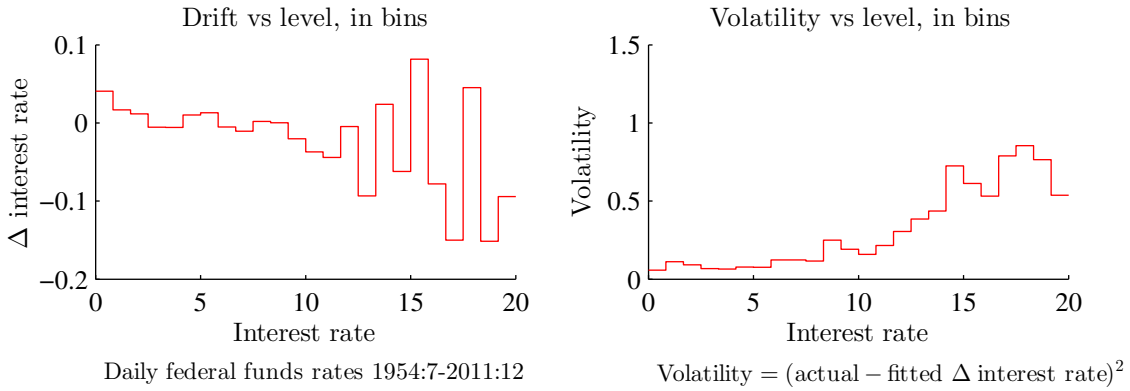


Figure 10.3: Crude non-parametric estimation

is out-of-sample. To calculate (10.7) we clearly need to make T estimations (for each x_t)—and then repeat this for different values of h to find the minimum.

See Figure 10.5 for an example.

Remark 10.5 *(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 (10.7).
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)

Remark 10.6 *(Speed and fast Fourier transforms) The calculation of the kernel estimator can often be speeded up by the use of a fast Fourier transform.*

If the observations are independent, then it can be shown (see Härdle (1990) 4.2, Pagan and Ullah (1999) 3.3–6, and also (10.6)) that, with a Gaussian kernel, the estimator at point x is asymptotically normally distributed

$$\sqrt{Th}\left[\hat{b}(x) - \mathrm{E}\,\hat{b}(x)\right] \to^d N\left[0, \frac{1}{2\sqrt{\pi}}\frac{\sigma^2(x)}{f(x)}\right], \tag{10.8}$$

where $\sigma^2(x)$ is the variance of the residuals in (10.1) and $f(x)$ the marginal density of x. (A similar expression holds for other choices of the kernel.) This expression assumes

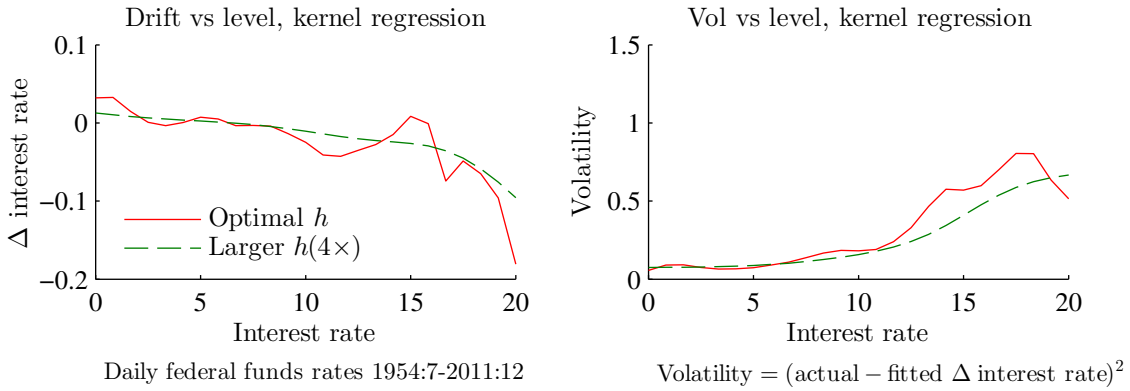


Figure 10.4: Kernel regression, importance of bandwidth

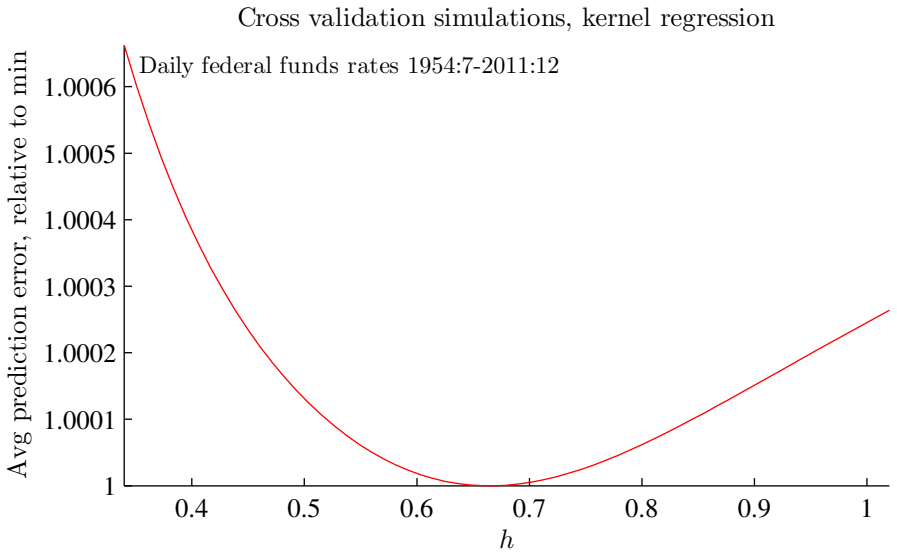


Figure 10.5: Cross-validation

that the asymptotic bias is zero, which is guaranteed if h is decreased (as T increases) slightly faster than $T^{-1/5}$. To estimate the density of x, we can apply a standard method, for instance using a Gaussian kernel and the bandwidth (for the density estimate only) of $1.06\,\text{Std}(x_t)T^{-1/5}$.

To estimate $\sigma^2(x)$ in (10.8), we use a non-parametric regression of the squared fitted residuals on x_t

$$\hat{\varepsilon}_t^2 = \sigma^2(x_t), \text{ where } \hat{\varepsilon}_t = y_t - \hat{b}(x_t), \tag{10.9}$$

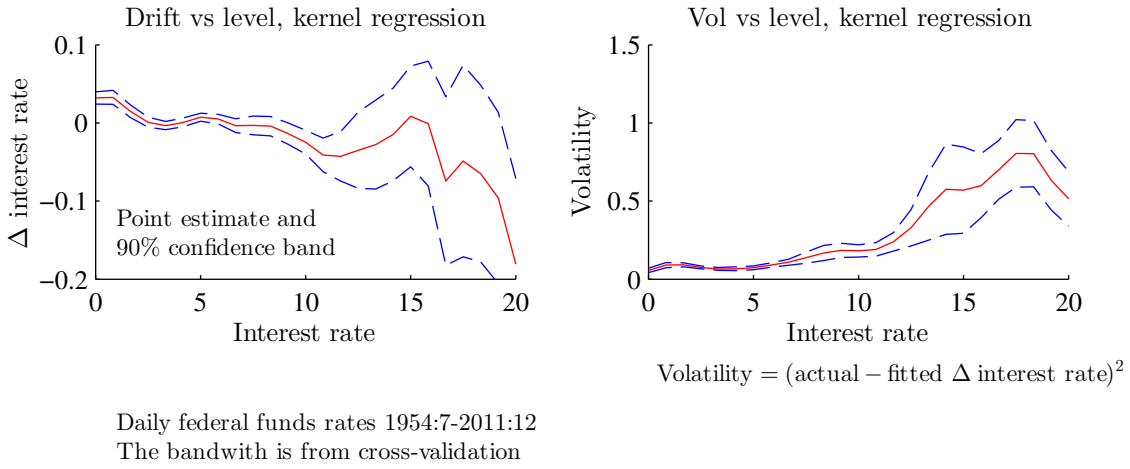


Figure 10.6: Kernel regression, confidence band

where $\hat{b}(x_t)$ are the fitted values from the non-parametric regression (10.1). Notice that the estimation of $\sigma^2(x)$ is quite computationally intensive since it requires estimating $\hat{b}(x)$ at every point $x = x_t$ in the sample. To draw confidence bands, it is typically assumed that the asymptotic bias is zero ($\mathrm{E}\,\hat{b}(x) = b(x)$).

See Figure 10.6 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.)

10.1.3 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \tag{10.10}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t | x_t, z_t) = 0$. 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 of observations in every bin.

In any case, the most common way to implement the kernel regressor is to let

$$\hat{b}(x,z) = \frac{\sum_{t=1}^{T} w_t(x) w_t(z) y_t}{\sum_{t=1}^{T} w_t(x) w_t(z)}, \tag{10.11}$$

where $w_t(x)$ and $w_t(z)$ are two kernels like in (10.4) 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.

10.1.4 "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices," by Ait-Sahalia and Lo (1998)

Reference: Ait-Sahalia and Lo (1998)

There seem 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").

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 nonparametric 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).

This paper estimates nonparametric option price functions and calculates the implicit risk-neutral distribution as the second partial derivative of this function with respect to the strike price.

1. First, the call option price, H_{it}, is estimated as a multivariate kernel regression

$$H_{it} = b(S_t, X, \tau, r_{\tau t}, \delta_{\tau t}) + \varepsilon_{it}, \tag{10.12}$$

   where S_t is the price of the underlying asset, X is the strike price, τ is time to expiry, $r_{\tau t}$ is the interest rate between t and $t + \tau$, and $\delta_{\tau t}$ is the dividend yield

(if any) between t and $t + \tau$. It is very hard to estimate a five-dimensional kernel regression, so various ways of reducing the dimensionality are tried. For instance, by making $b()$ a function of the forward price, $S_t[\tau \exp(r_{\tau t} - \delta_{\tau t})]$, instead of S_t, $r_{\tau t}$, and $\delta_{\tau t}$ separably.

2. Second, the implicit risk-neutral pdf of the future asset price is calculated as $\partial^2 b(S_t, X, \tau, r_{\tau t}, \delta_{\tau t})/\partial X^2$, properly scaled so it integrates to unity.

3. This approach is used on 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 nonparametric 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. (See their Fig 7.)

4. They also argue that there is little evidence of instability in the implicit pdf over their sample.

10.1.5 "Testing Continuous-Time Models of the Spot Interest Rate," by Ait-Sahalia (1996)

Reference: Ait-Sahalia (1996)

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

$$r_{t+1} = \alpha + \rho r_t + \varepsilon_{t+1}, \text{ where } \mathrm{E}_t\, \varepsilon_{t+1} = 0 \text{ and } \mathrm{E}_t\, \varepsilon_{t+1}^2 = \sigma^2 r_t^{2\gamma} \tag{10.13}$$

$$r_{t+1} - r_t = \alpha + \underbrace{\beta}_{\rho - 1} r_t + \varepsilon_{t+1} \tag{10.14}$$

nests several well-known models. It is an approximation of the diffusion process

$$dr_t = (\beta_0 + \beta_1 r_t)dt + \sigma r_t^{\gamma} dW_t, \tag{10.15}$$

where W_t is a Wiener process. Recall that affine one-factor models require $\gamma = 0$ (the Vasicek model) or $\gamma = 0.5$ (Cox-Ingersoll-Ross).

This paper tests several models of the short interest rate by using a nonparametric technique.

1. The first step of the analysis is to estimate the unconditional distribution of the short interest rate by a kernel density estimator. The estimated pdf at the value r is denoted $\hat{\pi}_0(r)$.

2. The second step is to estimate the parameters in a short rate model (for instance, Vasicek's model) by making the unconditional distribution implied by the model parameters (denoted $\pi(\theta, r)$ where θ is a vector of the model parameters and r a value of the short rate) as close as possible to the nonparametric estimate obtained in step 1. This is done by choosing the model parameters as

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{T} \sum_{t=1}^{T} [\pi(\theta, r_t) - \hat{\pi}_0(r)]^2. \tag{10.16}$$

3. The model is tested by using a scaled version of the minimized value of the right hand side of (10.16) as a test statistic (it has an asymptotic normal distribution).

4. It is found that most standard models are rejected (daily data on 7-day Eurodollar deposit rate, June 1973 to February 1995, 5,500 observations), mostly because actual mean reversion is much more non-linear in the interest rate level than suggested by most models (the mean reversion seems to kick in only for extreme interest rates and to be virtually non-existent for moderate rates).

5. For a critique of this approach (biased estimator...), see Chapman and Pearson (2000)

Remark 10.7 *The very non-linear mean reversion in Figures 10.3–10.4* seems to be the key reason for why *Ait-Sahalia (1996) rejects most short rate models.*

10.2 Approximating Non-Linear Regression Functions

10.2.1 Partial Linear Model

A possible way out of the curse of dimensionality of the multivariate kernel regression is to specify a partially linear model

$$y_t = z_t'\beta + b(x_t) + \varepsilon_t, \tag{10.17}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t|x_t, z_t) = 0$. This model is linear in z_t, but possibly non-linear in x_t since the function $b(x_t)$ is unknown.

To construct an estimator, start by taking expectations of (10.17) conditional on x_t

$$\mathrm{E}(y_t|x_t) = \mathrm{E}(z_t|x_t)'\beta + b(x_t). \tag{10.18}$$

Subtract from (10.17) to get

$$y_t - \mathrm{E}(y_t|x_t) = [z_t - \mathrm{E}(z_t|x_t)]'\beta + \varepsilon_t. \tag{10.19}$$

The *double residual method* (see Pagan and Ullah (1999) 5.2) has several steps. *First*, estimate $\mathrm{E}(y_t|x_t)$ by a kernel regression of y_t on x_t ($\hat{b}_y(x)$), and $\mathrm{E}(z_t|x_t)$ by a similar kernel regression of z_t on x_t ($\hat{b}_z(x)$). *Second*, use these estimates in (10.19)

$$y_t - \hat{b}_y(x_t) = [z_t - \hat{b}_z(x_t)]'\beta + \varepsilon_t \tag{10.20}$$

and estimate β by least squares. *Third*, use these estimates in (10.18) to estimate $b(x_t)$ as

$$\hat{b}(x_t) = \hat{b}_y(x_t) - \hat{b}_z(x_t)'\hat{\beta}. \tag{10.21}$$

It can be shown that (under the assumption that y_t, z_t and x_t are iid)

$$\sqrt{T}(\hat{\beta} - \beta) \to^d N\left[0, \mathrm{Var}(\varepsilon_t)\,\mathrm{Cov}(z_t|x_t)^{-1}\right]. \tag{10.22}$$

We can consistently estimate $\mathrm{Var}(\varepsilon_t)$ by the sample variance of the fitted residuals in (10.17)—plugging in the estimated β and $b(x_t)$: and we can also consistently estimate $\mathrm{Cov}(z_t|x_t)$ by the sample variance of $z_t - \hat{b}_z(x_t)$. Clearly, this result is based on the idea that we asymptotically know the non-parametric parts of the problem (which relies on the consistency of their estimators).

10.2.2 Basis Expansion

Reference: Hastie, Tibshirani, and Friedman (2001); Ranaldo and Söderlind (2010) (for an application of the method to exchange rates)

The label "non-parametrics" is something of a misnomer since these models typically have very many "parameters". For instance, the kernel regression is an attempt to estimate a specific slope coefficient at almost each value of the regressor. Not surprisingly, this becomes virtually impossible if the data set is small and/or there are several regressors.

An alternative approach is to estimate an approximation of the function $b(x_t)$ in

$$y_t = b(x_t) + \varepsilon_t. \tag{10.23}$$

This can be done by using piecewise polynomials or splines. In the simplest case, this amounts to just a piecewise linear (but continuous) function. For instance, if x_t is a scalar and we want three segments (pieces), then we could use the following building blocks

$$\begin{bmatrix} x_t \\ \max(x_t - \xi_1, 0) \\ \max(x_t - \xi_2, 0) \end{bmatrix} \tag{10.24}$$

and approximate as

$$b(x_t) = \beta_1 x_t + \beta_2 \max(x_t - \xi_1, 0) + \beta_3 \max(x_t - \xi_2, 0). \tag{10.25}$$

This can also be written

$$b(x_t) = \begin{bmatrix} \beta_1 x_t & \text{if } x_t < \xi_1 \\ \beta_1 x_t + \beta_2 (x_t - \xi_1) & \text{if } \xi_1 \leq x_t < \xi_2 \\ \beta_1 x_t + \beta_2 (x_t - \xi_1) + \beta_3 (x_t - \xi_2) & \text{if } \xi_2 \leq x_t \end{bmatrix}. \tag{10.26}$$

This function has the slope β_1 for $x_t < \xi_1$, the slope $\beta_1 + \beta_2$ between ξ_1 and ξ_2, and $\beta_1 + \beta_2 + \beta_3$ above ξ_2. It is no more sophisticated than using dummy variables (for the different segments), except that the current approach is a convenient way to guarantee that the function is continuous (this can be achieved also with dummies provided there are dummies for the intercept and a we impose restrictions on the slopes and intercepts). Figure 10.7 gives an illustration. It is straightforward to extend this to more segments.

However, the main difference to the typical use of dummy variables is that the "knots"

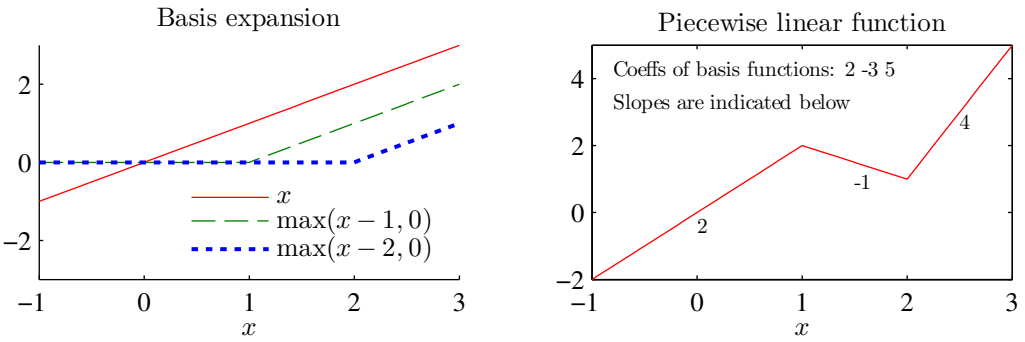


Figure 10.7: Example of piecewise linear function, created by basis expansion

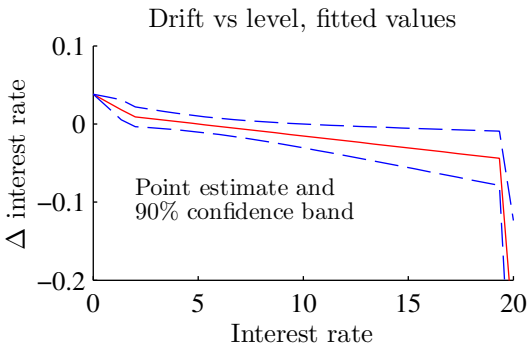
(here ξ_1 and ξ_2) are typically estimated along with the slopes (here β_1, β_2 and β_3). This can, for instance, be done by non-linear least squares.

Remark 10.8 *(NLS estimation) The parameter vector (ξ, β) is easily estimated by Non-Linear least squares (NLS) by concentrating the loss function: optimize (numerically) over ξ and let (for each value of ξ) the parameters in β be the OLS coefficients on the vector of regressors z_t (as in (10.24)).*

Let V be the covariance of the parameters collected in the vector θ (here $\xi_1, \xi_2, \beta_1, \beta_2, \beta_3$). For instance, we can use the t-stat for β_2 to test if the slope of the second segment ($\beta_1+\beta_2$) is different from the slope of the first segment (β_1).

To get the variance of $b(x_t)$ at a given point x_t, we can apply the delta method. To do that, we need the Jacobian of the $b(x_t)$ function with respect to θ. In applying the delta method we are assuming that $b(x_t)$ has continuos first derivatives—which is clearly not the case for the max function. However, we could replace the max function with an approximation like $\max(z, 0) \approx z/[1 + \exp(-2kz)]$ and then let k become very small—and we get virtually the same result. In any case, apart from at the knot points (where $x_t = \xi_1$ or $x_t = \xi_2$) we have the following derivatives

$$\frac{\partial b(x_t)}{\partial \theta} = \begin{bmatrix} \partial b\left(x_t\right)/\partial \xi_1 \\ \partial b\left(x_t\right)/\partial \xi_2 \\ \partial b\left(x_t\right)/\partial \beta_1 \\ \partial b\left(x_t\right)/\partial \beta_2 \\ \partial b\left(x_t\right)/\partial \beta_3 \end{bmatrix} = \begin{bmatrix} -\beta_2 I(x_t - \xi_1 \geq 0) \\ -\beta_3 I(x_t - \xi_2 \geq 0) \\ x_t \\ \max(x_t - \xi_1, 0) \\ \max(x_t - \xi_2, 0) \end{bmatrix}, \tag{10.27}$$



Daily federal funds rates 1954:7-2011:12
piecewise linear regession

	coeff	Std
knot 1	1.93	0.87
knot 2	19.49	0.30
slope seg 1	-0.01	0.01
extra slope seg 2	0.01	0.01
extra slope seg 3	-0.43	0.18
const	0.04	0.01

Figure 10.8: Federal funds rate, piecewise linear model

where $I(q) = 1$ if q is true and 0 otherwise. The variance of $\hat{b}(x_t)$ is then

$$\text{Var}[\hat{b}(x_t)] = \frac{\partial b(x_t)}{\partial \theta'} V \frac{\partial b(x_t)}{\partial \theta}. \tag{10.28}$$

Remark 10.9 *(The derivatives of $b(x_t)$) From (10.26) we have the following derivatives*

$$\begin{bmatrix} \partial b(x_t)/\partial \xi_1 \\ \partial b(x_t)/\partial \xi_2 \\ \partial b(x_t)/\partial \beta_1 \\ \partial b(x_t)/\partial \beta_2 \\ \partial b(x_t)/\partial \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ x_t \\ 0 \\ 0 \end{bmatrix} \text{ if } x_t < \xi_1, \begin{bmatrix} -\beta_2 \\ 0 \\ x_t \\ x_t - \xi_1 \\ 0 \end{bmatrix} \text{ if } \xi_1 \leq x_t < \xi_2, \begin{bmatrix} -\beta_2 \\ -\beta_3 \\ x_t \\ x_t - \xi_1 \\ x_t - \xi_2 \end{bmatrix} \text{ if } \xi_2 \leq x_t.$$

It is also straightforward to extend this several regressors—at least as long as we assume additivity of the regressors. For instance, with two variables (x_t and z_t)

$$b(x_t, z_t) = b_x(x_t) + b_z(z_t), \tag{10.29}$$

where both $b_x(x_t)$ and $b_z(z_t)$ are piecewise functions of the sort discussed in (10.26). Estimation is just as before, except that we have different knots for different variables. Estimating $\text{Var}[\hat{b}_x(x_t)]$ and $\text{Var}[\hat{b}_z(z_t)]$ follows the same approach as in (10.28).

See Figure 10.8 for an illustration.

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.

Chapman, D., and N. D. Pearson, 2000, "Is the short rate drift actually nonlinear?," *Journal of Finance*, 55, 355–388.

Härdle, W., 1990, *Applied nonparametric regression*, Cambridge University Press, Cambridge.

Hastie, T., R. Tibshirani, and J. Friedman, 2001, *The elements of statistical learning: data mining, inference and prediction*, Springer Verlag.

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.

Ranaldo, A., and P. Söderlind, 2010, "Safe haven currencies," *Review of Finance*, 10, 385–407.

11 Alphas /Betas and Investor Characteristics

11.1 Basic Setup

The task is to evaluate if alphas or betas of individual investors (or funds) are related to investor (fund) characteristics, for instance, age or trading activity. The data set is panel with observations for T periods and N investors. (In many settings, the panel is unbalanced, but, to keep things reasonably simple, that is disregarded in the discussion below.)

11.2 Calendar Time and Cross Sectional Regression

The *calendar time* (CalTime) approach is to first define M discrete investor groups (for instance, age 18–30, 31–40, etc) and calculate their respective average excess returns ($\bar{y}_{jt}$ for group j)

$$\bar{y}_{jt} = \frac{1}{N_j}\sum_{i\in\text{Group}\,j} y_{it}, \tag{11.1}$$

where N_j is the number of individuals in group j.

Then, we run a factor model

$$\bar{y}_{jt} = x_t'\beta_j + v_{jt}, \text{ for } j = 1, 2, \ldots, M \tag{11.2}$$

where x_t typically includes a constant and various return factors (for instance, excess returns on equity and bonds). By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the "alpha") is higher for the Mth group than for the for first group.

Example 11.1 *(CalTime with two investor groups) With two investor groups, estimate the*

following SURE system

$$\bar{y}_{1t} = x_t'\beta_1 + v_{1t},$$
$$\bar{y}_{2t} = x_t'\beta_2 + v_{2t}.$$

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

$$y_{it} = x_t'\beta_i + \varepsilon_{it}, \text{ for } i = 1, 2, \ldots, N \tag{11.3}$$

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

$$\hat{\beta}_{pi} = z_i'c_p + w_{pi}. \tag{11.4}$$

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.

11.3 Panel Regressions, Driscoll-Kraay and Cluster Methods

References: Hoechle (2011) and Driscoll and Kraay (1998)

11.3.1 OLS

Consider the regression model

$$y_{it} = x_{it}'\beta + \varepsilon_{it}, \tag{11.5}$$

where x_{it} is an $K \times 1$ vector. Notice that the coefficients are the same across individuals (and time). Define the matrices

$$\Sigma_{xx} = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} x_{it} x_{it}' \text{ (an } K \times K \text{ matrix)} \tag{11.6}$$

$$\Sigma_{xy} = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} x_{it} y_{it} \text{ (a } K \times 1 \text{ vector).} \tag{11.7}$$

The LS estimator (stacking all TN observations) is then

$$\hat{\beta} = \Sigma_{xx}^{-1} \Sigma_{xy}. \tag{11.8}$$

11.3.2 GMM

The sample moment conditions for the LS estimator are

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} h_{it} = \mathbf{0}_{K \times 1}, \text{ where } h_{it} = x_{it} \varepsilon_{it} = x_{it}(y_{it} - x_{it}' \beta). \tag{11.9}$$

Remark 11.2 *(Distribution of GMM estimates) Under fairly weak assumption, the exactly identified GMM estimator* $\sqrt{TN}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, D_0^{-1} S_0 D_0^{-1})$*, where* D_0 *is the Jacobian of the average moment conditions and* S_0 *is the covariance matrix of* $\sqrt{TN}$ *times the average moment conditions.*

Remark 11.3 *(Distribution of* $\hat{\beta} - \beta_0$*) As long as* TN *is finite, we can (with some abuse of notation) consider the distribution of* $\hat{\beta} - \beta$ *instead of* $\sqrt{TN}(\hat{\beta} - \beta_0)$ *to write*

$$\hat{\beta} - \beta_0 \sim N(0, D_0^{-1} S D_0^{-1}),$$

where $S = S_0/(TN)$ *which is the same as the covariance matrix of the average moment conditions (11.9).*

To apply these remarks, first notice that the Jacobian D_0 corresponds to (the probability limit of) the Σ_{xx} matrix in (11.6). Second, notice that

$$\text{Cov(average moment conditions)} = \text{Cov}\left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} h_{it} \right) \tag{11.10}$$

looks differently depending on the assumptions of cross correlations.

In particular, if h_{it} has no correlation across time (effectively, $\frac{1}{N}\sum_{i=1}^{N} h_{it}$ is not auto-correlated), then we can simplify as

$$\text{Cov(average moment conditions)} = \frac{1}{T^2}\sum_{t=1}^{T} \text{Cov}\left(\frac{1}{N}\sum_{i=1}^{N} h_{it}\right). \tag{11.11}$$

We would then design an estimator that would consistently estimate this covariance matrix by using the time dimension.

Example 11.4 *(DK on $T = 2$ and $N = 4$) As an example, suppose $K = 1$, $T = 2$ and $N = 4$. Then, (11.10) can be written*

$$\text{Cov}\left[\frac{1}{2\times 4}\left(h_{1t} + h_{2t} + h_{3t} + h_{4t}\right) + \frac{1}{2\times 4}\left(h_{1,t+1} + h_{2,t+1} + h_{3,t+1} + h_{4,t+1}\right)\right].$$

If there is no correlation across time periods, then this becomes

$$\frac{1}{2^2}\text{Cov}\left[\frac{1}{4}\left(h_{1t} + h_{2t} + h_{3t} + h_{4t}\right)\right] + \frac{1}{2^2}\text{Cov}\left[\frac{1}{4}\left(h_{1,t+1} + h_{2,t+1} + h_{3,t+1} + h_{4,t+1}\right)\right],$$

which has the same form as (11.11).

11.3.3 Driscoll-Kraay

The Driscoll and Kraay (1998) (DK) covariance matrix is

$$\text{Cov}(\hat{\beta}) = \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}, \tag{11.12}$$

where

$$S = \frac{1}{T^2}\sum_{t=1}^{T} h_t h_t', \text{ with } h_t = \frac{1}{N}\sum_{i=1}^{N} h_{it}, h_{it} = x_{it}\varepsilon_{it}, \tag{11.13}$$

where h_{it} is the LS moment condition for individual i. Clearly, h_{it} and h_t are $K \times 1$, so S is $K \times K$. Since we use the covariance matrix of the moment conditions, heteroskedasticity is accounted for.

Notice that h_t is the cross-sectional average moment condition (in t) and that S is an

estimator of the covariance matrix of those average moment conditions

$$S = \widehat{\text{Cov}}\left(\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it}\right).$$

To calculate this estimator, (11.13) uses the time dimension (and hence requires a reasonably long time series).

Remark 11.5 *(Relation to the notation in Hoechle (2011)) Hoechle writes* $\text{Cov}(\hat{\beta}) = (X'X)^{-1}\hat{S}_T(X'X)^{-1}$, *where* $\hat{S}_T = \sum_{t=1}^{T}\hat{h}_t\hat{h}_t'$, *with* $\hat{h}_t = \sum_{i=1}^{N} h_{it}$. *Clearly, my* $\Sigma_{xx} = X'X/(TN)$ *and my* $S = \hat{S}_T/(T^2N^2)$. *Combining gives* $\text{Cov}(\hat{\beta}) = (\Sigma_{xx}TN)^{-1}\left(ST^2N^2\right)(\Sigma_{xx}TN)^{-1}$, *which simplifies to (11.12).*

Example 11.6 *(DK on* $N = 4$*) As an example, suppose* $K = 1$ *and* $N = 4$. *Then, (11.13) gives the cross-sectional average in period t*

$$h_t = \frac{1}{4}\left(h_{1t} + h_{2t} + h_{3t} + h_{4t}\right),$$

and the covariance matrix

$$\begin{aligned}
S &= \frac{1}{T^2}\sum_{t=1}^{T} h_t h_t' \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\left[\frac{1}{4}\left(h_{1t} + h_{2t} + h_{3t} + h_{4t}\right)\right]^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{16}(h_{1t}^2 + h_{2t}^2 + h_{3t}^2 + h_{4t}^2, \\
&+ 2h_{1t}h_{2t} + 2h_{1t}h_{3t} + 2h_{1t}h_{4t} + 2h_{2t}h_{3t} + 2h_{2t}h_{4t} + 2h_{3t}h_{4t})
\end{aligned}$$

so we can write

$$\begin{aligned}
S = \frac{1}{T\times 16}\Bigg[&\sum_{i=1}^{4}\widehat{\text{Var}}(h_{it}) \\
&+ 2\widehat{\text{Cov}}(h_{1t}, h_{2t}) + 2\widehat{\text{Cov}}(h_{1t}, h_{3t}) + 2\widehat{\text{Cov}}(h_{1t}, h_{4t}) \\
&+ 2\widehat{\text{Cov}}(h_{2t}, h_{3t}) + 2\widehat{\text{Cov}}(h_{2t}, h_{4t}) \\
&+ 2\widehat{\text{Cov}}(h_{3t}, h_{4t})\Bigg].
\end{aligned}$$

Notice that S is the (estimate of) the variance of the cross-sectional average, $\text{Var}(h_t) = \text{Var}[(h_{1t} + h_{2t} + h_{3t} + h_{4t})/4]$.

A *cluster method* puts restrictions on the covariance terms (of h_{it}) that are allowed to enter the estimate S. In practice, all terms across clusters are left out. This can be implemented by changing the S matrix. In particular, instead of interacting all i with each other, we only allow for interaction within each of the G clusters ($g = 1, ..., G$)

$$S = \sum_{g=1}^{G} \frac{1}{T^2} \sum_{t=1}^{T} h_t^g \left(h_t^g\right)', \text{ where } h_t^g = \frac{1}{N} \sum_{i \in \text{cluster } g} h_{it}. \tag{11.14}$$

(Remark: the cluster sums should be divided by N, not the number of individuals in the cluster.)

Example 11.7 *(Cluster method on $N = 4$, changing Example 11.6 directly) Reconsider Example 11.6, but assume that individuals* 1 *and* 2 *form cluster* 1 *and that individuals* 3 *and* 4 *form cluster* 2*—and disregard correlations across clusters. This means setting the covariances across clusters to zero,*

$$\begin{aligned} S = \frac{1}{T^2} \sum_{t=1}^{T} \frac{1}{16} (&h_{1t}^2 + h_{2t}^2 + h_{3t}^2 + h_{4t}^2, \\ &2h_{1t}h_{2t} + \underbrace{2h_{1t}h_{3t}}_{0} + \underbrace{2h_{1t}h_{4t}}_{0} + \underbrace{2h_{2t}h_{3t}}_{0} + \underbrace{2h_{2t}h_{4t}}_{0} + 2h_{3t}h_{4t}) \end{aligned}$$

so we can write

$$S = \frac{1}{T \times 16} \left[\sum_{i=1}^{4} \widehat{\text{Var}}(h_{it}) + 2\widehat{\text{Cov}}(h_{1t}, h_{2t}) + 2\widehat{\text{Cov}}(h_{3t}, h_{4t}) \right].$$

Example 11.8 *(Cluster method on $N = 4$) From (11.14) we have the cluster (group) averages*

$$h_t^1 = \frac{1}{4} (h_{1t} + h_{2t}) \text{ and } h_t^2 = \frac{1}{4} (h_{3t} + h_{4t}).$$

Assuming only one regressor (to keep it simple), the time averages, $\frac{1}{T} \sum_{t=1}^{T} h_t^g \left(h_t^g\right)'$, *are*

then (for cluster 1 *and then* 2*)*

$$\frac{1}{T}\sum_{t=1}^{T} h_t^1 \left(h_t^1\right)' = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{4}\left(h_{1t}+h_{2t}\right)\right]^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{16}\left(h_{1t}^2+h_{2t}^2+2h_{1t}h_{2t}\right), \text{ and}$$

$$\frac{1}{T}\sum_{t=1}^{T} h_t^2 \left(h_t^2\right)' = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{16}\left(h_{3t}^2+h_{4t}^2+2h_{3t}h_{4t}\right).$$

Finally, summing across these time averages gives the same expression as in Example 11.7. The following 4×4 *matrix illustrates which cells that are included (assumption: no dependence across time)*

i	$\underline{1}$	$\underline{2}$	$\underline{3}$	$\underline{4}$
$\underline{1}$	h_{1t}^2	$h_{1t}h_{2t}$	0	0
$\underline{2}$	$h_{1t}h_{2t}$	h_{2t}^2	0	0
$\underline{3}$	0	0	h_{3t}^2	$h_{3t}h_{4t}$
$\underline{4}$	0	0	$h_{3t}h_{4t}$	h_{4t}^2

In comparison, the iid case only sums up the principal diagonal, while the DK method fills the entire matrix.

Instead, we get *White's covariance matrix* by excluding all cross terms. This can be accomplished by defining

$$S = \frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N^2}\sum_{i=1}^{N} h_{it}h_{it}'. \tag{11.15}$$

Example 11.9 *(White's method on* $N=4$*) With only one regressor (11.15) gives*

$$\begin{aligned} S &= \frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{16}\left(h_{1t}^2+h_{2t}^2+h_{3t}^2+h_{4t}^2\right) \\ &= \frac{1}{T\times 16}\sum_{i=1}^{4}\widehat{\text{Var}}(h_{it}) \end{aligned}$$

Finally, the traditional LS covariance matrix assumes that $\text{E}\, h_{it}h_{it}' = \Sigma_{xx}\times \text{E}\,\varepsilon_{it}^2$, so we get

$$\text{Cov}_{LS}(\hat{\beta}) = \Sigma_{xx}^{-1}s^2/TN, \text{ where } s^2 = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\varepsilon_{it}^2. \tag{11.16}$$

Remark 11.10 *(Why the cluster method fails when there is a missing "time fixed effect"—and one of the regressors indicates the cluster membership) To keep this remark short, assume* $y_{it} = 0q_{it} + \varepsilon_{it}$*, where* q_{it} *indicates the cluster membership of individual* i *(constant over time). In addition, assume that all individual residuals are entirely due to an (excluded) time fixed effect,* $\varepsilon_{it} = w_t$*. Let* $N = 4$ *where* $i = (1, 2)$ *belong to the first cluster (*$q_i = -1$*) and* $i = (3, 4)$ *belong to the second cluster (*$q_i = 1$*). (Using the values* $q_i = \pm 1$ *gives* q_i *a zero mean, which is convenient.) It is straightforward to demonstrate that the estimated (OLS) coefficient in any sample must be zero: there is in fact no uncertainty about it. The individual moments in period* t *are then* $h_{it} = q_{it} \times w_t$

$$\begin{bmatrix} h_{1t} \\ h_{2t} \\ h_{3t} \\ h_{4t} \end{bmatrix} = \begin{bmatrix} -w_t \\ -w_t \\ w_t \\ w_t \end{bmatrix}.$$

The matrix in Example 11.8 is then

i	$\underline{1}$	$\underline{2}$	$\underline{3}$	$\underline{4}$
$\underline{1}$	w_t^2	w_t^2	0	0
$\underline{2}$	w_t^2	w_t^2	0	0
$\underline{3}$	0	0	w_t^2	w_t^2
$\underline{4}$	0	0	w_t^2	w_t^2

These elements sum up to a positive number—which is wrong since $\sum_{i=1}^{N} h_{it} = 0$ *by definition, so its variance should also be zero. In contrast, the DK method adds the off-diagonal elements which are all equal to* $-w_t^2$*, so summing the whole matrix indeed gives zero. If we replace the* q_{it} *regressor with something else (eg a constant), then we do* not *get this result.*

To see what happens if the q_i *variable does not coincide with the definitions of the clusters change the regressor to* $q_i = (-1, 1, -1, 1)$ *for the four individuals. We then get* $(h_{1t}, h_{2t}, h_{3t}, h_{4t}) = (-w_t, w_t, -w_t, w_t)$*. If the definition of the clusters (for the covari-*

ance matrix) are unchanged, then the matrix in Example 11.8 becomes

$$\begin{array}{cccccc} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & w_t^2 & -w_t^2 & 0 & 0 \\ \underline{2} & -w_t^2 & w_t^2 & 0 & 0 \\ \underline{3} & 0 & 0 & w_t^2 & -w_t^2 \\ \underline{4} & 0 & 0 & -w_t^2 & w_t^2 \end{array}$$

which sum to zero: the cluster covariance estimator works fine. The DK method also works since it adds the off-diagonal elements which are

$$\begin{array}{cccccc} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & & & w_t^2 & -w_t^2 \\ \underline{2} & & & -w_t^2 & w_t^2 \\ \underline{3} & w_t^2 & -w_t^2 & & \\ \underline{4} & -w_t^2 & w_t^2 & & \end{array}$$

which also sum to zero. This suggests that the cluster covariance matrix goes wrong only when the cluster definition (for the covariance matrix) is strongly related to the q_i regressor.

11.4 From CalTime To a Panel Regression

The CalTime estimates can be replicated by using the individual data in the panel. For instance, with two investor groups we could estimate the following two regressions

$$y_{it} = x_t'\beta_1 + u_{it}^{(1)} \text{ for } i \in \text{group 1} \tag{11.17}$$

$$y_{it} = x_t'\beta_2 + u_{it}^{(2)} \text{ for } i \in \text{group 2.} \tag{11.18}$$

More interestingly, these regression equations can be combined into one panel regression (and still give the same estimates) by the help of dummy variables. Let $z_{ji} = 1$ if individual i is a member of group j and zero otherwise. Stacking all the data, we have

(still with two investor groups)

$$\begin{aligned} y_{it} &= (z_{1i}x_t)'\beta_1 + (z_{2i}x_t)'\beta_2 + u_{it} \\ &= \left(\begin{bmatrix} z_{1i}x_t \\ z_{2i}x_t \end{bmatrix}\right)' \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u_{it} \\ &= (z_i \otimes x_t)'\beta + u_{it}, \text{ where } z_i = \begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix}. \end{aligned} \tag{11.19}$$

This is estimated with LS by stacking all NT observations.

Since the CalTime approach (11.2) and the panel approach (11.19) give the same coefficients, it is clear that the errors in the former are just group averages of the errors in the latter

$$v_{jt} = \frac{1}{N_j}\sum_{i\in\text{Group } j} u_{it}^{(j)}. \tag{11.20}$$

We know that

$$\text{Var}(v_{jt}) = \frac{1}{N_j}(\overline{\sigma}_{ii} - \overline{\sigma}_{ih}) + \overline{\sigma}_{ih}, \tag{11.21}$$

where $\overline{\sigma}_{ii}$ is the average $\text{Var}(u_{it}^{(j)})$ and $\overline{\sigma}_{ih}$ is the average $\text{Cov}(u_{it}^{(j)}, u_{ht}^{(j)})$. With a large cross-section, only the covariance matters. A good covariance estimator for the panel approach will therefore have to handle the covariance with a group—and perhaps also the covariance across groups. This suggests that the panel regression needs to handle the cross-correlations (for instance, by using the cluster or DK covariance estimators).

11.5 The Results in Hoechle, Schmid and Zimmermann

Hoechle, Schmid, and Zimmermann (2009) (HSZ) suggest the following regression on all data ($t = 1, \ldots, T$ and also $i = 1, \ldots, N$)

$$y_{it} = (z_{it} \otimes x_t)'d + v_{it} \tag{11.22}$$

$$= ([1, z_{1it}, \ldots, z_{mit}] \otimes [1, x_{1t}, \ldots, x_{kt}])'d + v_{it}, \tag{11.23}$$

where y_{it} is the return of investor i in period t, z_{qit} measures characteristics q of investor i in period t and where x_{pt} is the pth pricing factor. In many cases z_{jit} is time-invariant and could even be just a dummy: $z_{jit} = 1$ if investor i belongs to investor group j (for instance being 18–30 years old). In other cases, z_{jit} is still time invariant and con-

tains information about the number of fund switches as well as other possible drivers of performance like gender. The x_t vector contains the pricing factors. In case the characteristics $z_{1it}, \ldots, z_{mit}$ sum to unity (for a given individual i and time t), the constant in $[1, z_{1it}, \ldots, z_{mit}]$ is dropped.

This model is estimated with LS (stacking all NT observations), but the standard errors are calculated according to Driscoll and Kraay (1998) (DK)—which accounts for cross-sectional correlations, for instance, correlations between the residuals of different investors (say, v_{1t} and v_{7t}).

HSZ prove the following two propositions.

Proposition 11.11 *If the z_{it} vector in (11.22) consists of dummy variables indicating exclusive and constant group membership ($z_{1it} = 1$ means that investor i belongs to group 1, so $z_{jit} = 0$ for $j = 2, ..., m$), then the LS estimates and DK standard errors of (11.22) are the same as LS estimates and Newey-West standard errors of the CalTime approach (11.2). (See HSZ for a proof.)*

Proposition 11.12 *(When z_{it} is a measure of investor characteristics, eg number of fund switches) The LS estimates and DK standard errors of (11.22) are the same as the LS estimates of CrossReg approach (11.4), but where the standard errors account for the cross-sectional correlations, while those in the CrossReg approach do not. (See HSZ for a proof.)*

Example 11.13 *(One investor characteristic and one pricing factor). In this case (11.22) is*

$$\begin{aligned} y_{it} &= \begin{bmatrix} 1 \\ x_{1t} \\ z_{it} \\ z_{it}x_{1t} \end{bmatrix}' d + v_{it}, \\ &= d_0 + d_1 x_{1t} + d_2 z_{it} + d_3 z_{it} x_{1t} + v_{it}. \end{aligned}$$

In case we are interested in how the investor characteristics (z_{it}) affect the alpha (intercept), then d_2 is the key coefficient.

11.6 Monte Carlo Experiment

11.6.1 Basic Setup

This section reports results from a simple Monte Carlo experiment. We use the model

$$y_{it} = \alpha + \beta f_t + \delta g_i + \varepsilon_{it}, \tag{11.24}$$

where y_{it} is the return of individual i in period t, f_t a benchmark return and g_i is the (demeaned) number of the cluster $(-2, -1, 0, 1, 2)$ that the individual belongs to. This is a simplified version of the regressions we run in the paper. In particular, δ measures how the performance depends on the number of fund switches.

The experiment uses 3000 artificial samples with $t = 1, \ldots, 2000$ and $i = 1, \ldots, 1665$. Each individual is a member of one of five equally sized groups (333 individuals in each group). The benchmark return f_t is iid normally distributed with a zero mean and a standard deviation equal to $15/\sqrt{250}$, while ε_{it} is a also normally distributed with a zero mean and a standard deviation of one (different cross-sectional correlations are shown in the table). In generating the data, the true values of α and δ are zero, while β is one—and these are also the hypotheses tested below. To keep the simulations easy to interpret, there is no autocorrelation or heteroskedasticity.

Results for three different GMM-based methods are reported: Driscoll and Kraay (1998), a cluster method and White's method. To keep the notation short, let the regression model be $y_{it} = x_{it}'b + \varepsilon_{it}$, where x_{it} is a $K \times 1$ vector of regressors. The (least squares) moment conditions are

$$\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it} = \mathbf{0}_{K\times 1}, \text{ where } h_{it} = x_{it}\varepsilon_{it}. \tag{11.25}$$

Standard GMM results show that the variance-covariance matrix of the coefficients is

$$\text{Cov}(\hat{b}) = \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}, \text{ where } \Sigma_{xx} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it}x_{it}', \tag{11.26}$$

and S is covariance matrix of the moment conditions.

The three methods differ with respect to how the S matrix is estimated

$$
\begin{aligned}
S_{DK} &= \frac{1}{T^2N^2}\sum_{t=1}^{T} h_t h_t', \text{ where } h_t = \sum_{i=1}^{N} h_{it}, \\
S_{Cl} &= \frac{1}{T^2N^2}\sum_{t=1}^{T}\sum_{j=1}^{M} h_t^j (h_t^j)', \text{ where } h_t^j = \sum_{i \in \text{ cluster } j} h_{it}, \\
S_{Wh} &= \frac{1}{T^2N^2}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it} h_{it}'.
\end{aligned} \tag{11.27}
$$

To see the difference, consider a simple example with $N = 4$ and where $i = (1, 2)$ belong to the first cluster and $i = (3, 4)$ belong to the second cluster. The following matrix shows the outer product of the moment conditions of all individuals. White's estimator sums up the cells on the principal diagonal, the cluster method adds the underlined cells, and the DK method adds also the remaining cells

$$
\begin{bmatrix}
i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\
\underline{1} & h_{1t}h_{1t}' & \underline{h_{1t}h_{2t}'} & h_{1t}h_{3t}' & h_{1t}h_{4t}' \\
\underline{2} & \underline{h_{2t}h_{1t}'} & h_{2t}h_{2t}' & h_{2t}h_{3t}' & h_{2t}h_{4t}' \\
\underline{3} & h_{3t}h_{1t}' & h_{3t}h_{2t}' & h_{3t}h_{3t}' & \underline{h_{3t}h_{4t}'} \\
\underline{4} & h_{4t}h_{1t}' & h_{4t}h_{2t}' & \underline{h_{4t}h_{3t}'} & h_{4t}h_{4t}'
\end{bmatrix} \tag{11.28}
$$

11.6.2 MC Covariance Structure

To generate data with correlated (in the cross-section) residuals, let the residual of individual i (belonging to group j) in period t be

$$
\varepsilon_{it} = u_{it} + v_{jt} + w_t, \tag{11.29}
$$

where $u_{it} \sim N(0, \sigma_u^2)$, $v_{jt} \sim N(0, \sigma_v^2)$ and $w_t \sim N(0, \sigma_w^2)$—and the three components are uncorrelated. This implies that

$$
\begin{aligned}
\text{Var}(\varepsilon_{it}) &= \sigma_u^2 + \sigma_v^2 + \sigma_w^2, \\
\text{Cov}(\varepsilon_{it}, \varepsilon_{kt}) &= \begin{bmatrix} \sigma_v^2 + \sigma_w^2 & \text{if individuals } i \text{ and } k \text{ belong to the same group} \\ \sigma_w^2 & \text{otherwise.} \end{bmatrix}
\end{aligned} \tag{11.30}
$$

Clearly, when $\sigma_w^2 = 0$ then the correlation across groups is zero, but there may be correlation within a group. If both $\sigma_v^2 = 0$ and $\sigma_w^2 = 0$, then there is no correlation at all across individuals. For CalTime portfolios (one per activity group), we expect the u_{it} to average out, so a group portfolio has the variance $\sigma_v^2 + \sigma_w^2$ and the covariance of two different group portfolios is σ_w^2.

The Monte Carlo simulations consider different values of the variances—to illustrate the effect of the correlation structure.

11.6.3 Results from the Monte Carlo Simulations

Table 11.1 reports the fraction of times the absolute value of a t-statistics for a true null hypothesis is higher than 1.96. The table has three panels for different correlation patterns the residuals (ε_{it}): no correlation between individuals, correlations only within the pre-specified clusters and correlation across all individuals.

In the *upper panel*, where the residuals are iid, all three methods have rejection rates around 5% (the nominal size).

In the *middle panel*, the residuals are correlated within each of the five clusters, but there is no correlation between individuals that belong to the different clusters. In this case, but the DK and the cluster method have the right rejection rates, while White's method gives much too high rejection rates (around 85%). The reason is that White's method disregards correlation between individuals—and in this way underestimates the uncertainty about the point estimates. It is also worth noticing that the good performance of the cluster method depends on pre-specifying the correct clustering. Further simulations (not tabulated) shows that with a completely random cluster specification (unknown to the econometrician), gives almost the same results as White's method.

The *lower panel* has no cluster correlations, but all individuals are now equally correlated (similar to a fixed time effect). For the intercept (α) and the slope coefficient on the common factor (β), the DK method still performs well, while the cluster and White's methods give too many rejects: the latter two methods underestimate the uncertainty since some correlations across individuals are disregarded. Things are more complicated for the slope coefficient of the cluster number (δ). Once again, DK performs well, but both the cluster and White's methods lead to too few rejections. The reason is the interaction of the common component in the residual with the cross-sectional dispersion of the group number (g_i).

	White	Cluster	Driscoll-Kraay
A. No cross-sectional correlation			
α	0.049	0.049	0.050
β	0.044	0.045	0.045
γ	0.050	0.051	0.050
B. Within-cluster correlations			
α	0.853	0.053	0.054
β	0.850	0.047	0.048
γ	0.859	0.049	0.050
C. Within- and between-cluster correlations			
α	0.935	0.377	0.052
β	0.934	0.364	0.046
γ	0.015	0.000	0.050

Table 11.1: **Simulated size of different covariance estimators** This table presents the fraction of rejections of true null hypotheses for three different estimators of the covariance matrix: White's (1980) method, a cluster method, and Driscoll and Kraay's (1998) method. The model of individual i in period t and who belongs to cluster j is $r_{it} = \alpha + \beta f_t + \gamma g_i + \varepsilon_{it}$, where f_t is a common regressor (iid normally distributed) and g_i is the demeaned number of the cluster that the individual belongs to. The simulations use 3000 repetitions of samples with $t = 1, \ldots, 2000$ and $i = 1, \ldots, 1665$. Each individual belongs to one of five different clusters. The error term is constructed as $\varepsilon_{it} = u_{it} + v_{jt} + w_t$, where u_{it} is an individual (iid) shock, v_{jt} is a shock common to all individuals who belong to cluster j, and w_t is a shock common to all individuals. All shocks are normally distributed. In Panel A the variances of (u_{it}, v_{jt}, w_t) are (1,0,0), so the shocks are iid; in Panel B the variances are (0.67,0.33,0), so there is a 33% correlation within a cluster but no correlation between different clusters; in Panel C the variances are (0.67,0,0.33), so there is no cluster-specific shock and all shocks are equally correlated, effectively having a 33% correlation within a cluster and between clusters.

To understand this last result, consider a stylised case where $y_{it} = \delta g_i + \varepsilon_{it}$ where $\delta = 0$ and $\varepsilon_{it} = w_t$ so all residuals are due to an (excluded) time fixed effect. In this case, the matrix above becomes

$$\begin{bmatrix} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & w_t^2 & \underline{w_t^2} & -w_t^2 & -w_t^2 \\ \underline{2} & \underline{w_t^2} & w_t^2 & -w_t^2 & -w_t^2 \\ \underline{3} & -w_t^2 & -w_t^2 & w_t^2 & \underline{w_t^2} \\ \underline{4} & -w_t^2 & -w_t^2 & \underline{w_t^2} & w_t^2 \end{bmatrix} \tag{11.31}$$

(This follows from $g_i = (-1, -1, 1, 1)$ and since $h_{it} = g_i \times w_t$ we get $(h_{1t}, h_{2t}, h_{3t}, h_{4t}) = (-w_t, -w_t, w_t, w_t)$.) Both White's and the cluster method sums up only positive cells, so S is a strictly positive number. (For this the cluster method, this result relies on the assumption that the clusters used in estimating S correspond to the values of the regressor, g_i.) However, that is wrong since it is straightforward to demonstrate that the estimated coefficient in any sample must be zero. This is seen by noticing that $\sum_{i=1}^{N} h_{it} = 0$ at a zero slope coefficient holds for all t, so there is in fact no uncertainty about the slope coefficient. In contrast, the DK method adds the off-diagonal elements which are all equal to $-w_t^2$, giving the correct result $S = 0$.

11.7 An Empirical Illustration

See 11.2 for results on a ten-year panel of some 60,000 Swedish pension savers (Dahlquist, Martinez and Söderlind, 2011).

Bibliography

Driscoll, J., and A. Kraay, 1998, "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *Review of Economics and Statistics*, 80, 549–560.

Hoechle, D., 2011, "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence," *The Stata Journal* forhcoming.

Hoechle, D., M. M. Schmid, and H. Zimmermann, 2009, "A Generalization of the Cal-

endar Time Portfolio Approach and the Performance of Private Investors," Working paper, University of Basel.

Table 11.2: **Investor activity, performance, and characteristics**

	I	II	III	IV
Constant	–0.828 (2.841)	–1.384 (3.284)	–0.651 (2.819)	–1.274 (3.253)
Default fund	0.406 (1.347)	0.387 (1.348)	0.230 (1.316)	0.217 (1.320)
1 change	0.117 (0.463)	0.125 (0.468)		
2– 5 changes	0.962 (0.934)	0.965 (0.934)		
6–20 changes	2.678 (1.621)	2.665 (1.623)		
21–50 changes	4.265 (2.074)	4.215 (2.078)		
51– changes	7.114 (2.529)	7.124 (2.535)		
Number of changes			0.113 (0.048)	0.112 (0.048)
Age		0.008 (0.011)		0.008 (0.011)
Gender		0.306 (0.101)		0.308 (0.101)
Income		–0.007 (0.033)		0.009 (0.036)
R-squared (in %)	55.0	55.1	55.0	55.1

The table presents the results of pooled regressions of an individual's daily excess return on return factors, and measures of individuals' fund changes and other characteristics. The return factors are the excess returns of the Swedish stock market, the Swedish bond market, and the world stock market, and they are allowed to across the individuals' characteristics. For brevity, the coefficients on these return factors are not presented in the table. The measure of fund changes is either a dummy variable for an activity category (see Table **??**) or a variable counting the number of fund changes. Other characteristics are the individuals' age in 2000, gender, or pension rights in 2000, which is a proxy for income. The constant term and coefficients on the dummy variables are expressed in % per year. The income variable is scaled down by 1,000. Standard errors, robust to conditional heteroscedasticity and spatial cross-sectional correlations as in Driscoll and Kraay (1998), are reported in parentheses. The sample consists of 62,640 individuals followed daily over the 2000 to 2010 period.