

Lecture Notes in Empirical Finance (PhD)

Paul Söderlind¹

29 January 2018

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

Contents

1 Econometrics Cheat Sheet	6
1.1 Testing (Linear) Joint Hypotheses	6
1.2 Testing (Nonlinear) Joint Hypotheses: The Delta Method	7
1.3 The Variance of a Sample Average	10
1.4 OLS	14
1.5 MLE	15
1.6 GMM	17
1.7 Appendix: A Primer on Using Numerical Optimization Routines*	27
1.8 Appendix: Statistical Tables	29
1.9 Appendix: Data Sources	29
2 Basic Asset Pricing Theory	31
2.1 Three Pricing Principles	31
2.2 Stochastic Discount Factors	33
2.3 Beta Pricing Models	40
2.4 Risk Neutral Distributions	48
3 Simulating the Finite Sample Properties	55
3.1 Introduction	55
3.2 Monte Carlo Simulations	56
3.3 Bootstrapping	62
4 Return Distributions	68
4.1 Estimating and Testing Distributions	68
4.2 Estimating Risk-neutral Distributions from Options	79
4.3 Threshold Exceedance and Tail Distribution*	83
4.4 Exceedance Correlations*	87

4.5	Beyond (Linear) Correlations*	88
4.6	Copulas*	91
4.7	Joint Tail Distribution*	96
5	Predicting Asset Returns	127
5.1	A Little Financial Theory and Predictability	127
5.2	Autocorrelations	129
5.3	Multivariate (Auto-)correlations	138
5.4	Other Predictors	144
5.5	Spurious Regressions and In-Sample Overfitting	146
5.6	Model Selection	148
5.7	Forecast Averaging	154
5.8	Out-of-Sample Forecasting Performance	154
5.9	Evaluating Forecasting Performance	156
5.10	Appendix: Prices and Dividends	161
6	Predicting Asset Returns: Nonparametric Estimation	164
6.1	Basics of Kernel Regressions	164
6.2	Distribution of the Kernel Regression and Choice of Bandwidth	169
6.3	Local Linear Regressions	174
6.4	Applications of Kernel Regressions	177
7	Predicting and Modelling Volatility	181
7.1	Heteroskedasticity	181
7.2	ARCH Models	193
7.3	GARCH Models	197
7.4	Value at Risk	200
7.5	Non-Linear Extensions	202
7.6	GARCH Models with Exogenous Variables	204
7.7	Stochastic Volatility Models	205
7.8	(G)ARCH-M	207
7.9	Multivariate (G)ARCH	209
7.10	LAD and Quantile Regressions*	214
7.11	“A Closed-Form GARCH Option Valuation Model” by Heston and Nandi	221
7.12	“Fundamental Values and Asset Returns in Global Equity Markets,” by Bansal and Lundblad	228

7.13	Appendix: Using an FFT to Calculate the PDF from the Characteristic Function	232
7.14	Appendix: Some Proofs	234
8	Factor Models	236
8.1	CAPM Tests: Overview	236
8.2	Testing CAPM: Traditional LS Approach	237
8.3	Testing CAPM: GMM	240
8.4	Testing Multi-Factor Models (Factors are Excess Returns)	247
8.5	Testing Multi-Factor Models (General Factors)	253
8.6	Linear SDF Models	266
8.7	Conditional Factor Models	269
8.8	Conditional Models with “Regimes”	271
8.9	Fama-MacBeth*	273
8.10	Appendix: Details of CAPM Regression	276
8.11	Appendix: Details of SURE Systems	277
9	Consumption-Based Asset Pricing	281
9.1	Consumption-Based Asset Pricing	281
9.2	Asset Pricing Puzzles	284
9.3	The Cross-Section of Returns: Unconditional Models	288
9.4	The Cross-Section of Returns: Conditional Models	291
9.5	Ultimate Consumption	294
9.6	Long Run Risk	296
10	Financial Panel Data	305
10.1	Introduction to Panel Data	305
10.2	An Overview of Different Panel Data Models	306
10.3	Pooled OLS	307
10.4	The Within Estimator (“Fixed Effects Estimator”)	309
10.5	The First-Difference Estimator	312
10.6	Differences-in-Differences Estimator	312
10.7	Random Effects Model*	313
10.8	Fama-MacBeth	315
10.9	Calendar Time and Cross Sectional Regression	316
10.10	Panel Regressions, Driscoll-Kraay and Cluster Methods	320

10.11 From CalTime to a Panel Regression	327
10.12 The Results in Hoechle, Schmid and Zimmermann	330
11 Expectations Hypothesis of Interest Rates	334
11.1 Term (Risk) Premia	334
11.2 Testing the Expectations Hypothesis of Interest Rates	337
11.3 Spread-Based Tests*	339
11 Yield Curve Models: MLE and GMM	344
11.1 Describing Yield Curves	344
11.2 Risk Premia on Fixed Income Markets	348
11.3 Summary of the Solutions of Some Affine Yield Curve Models	349
11.4 MLE of Affine Yield Curve Models	355
11.5 Summary of Some Empirical Findings	367
11.6 Appendix: Details on Yield Curve Models	371
12 Yield Curve Models: Nonparametric Estimation	374
12.1 Nonparametric Regression	374
12.2 Approximating Non-Linear Regression Functions	376
12.3 Appendix: Partial Linear Model	378

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

Chapter 1

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 Testing (Linear) Joint Hypotheses

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

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_{k \times k}), \quad (1.1)$$

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

$$\gamma_{q \times 1} = R\beta. \quad (1.2)$$

Under that null hypothesis (that $\gamma = \gamma_0$)

$$\begin{aligned} \sqrt{T}(R\beta - \gamma_0) &\xrightarrow{d} N(0, \Lambda_{q \times q}), \text{ where} \\ \Lambda &= RVR'. \end{aligned} \quad (1.3)$$

Example 1.1 (*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

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

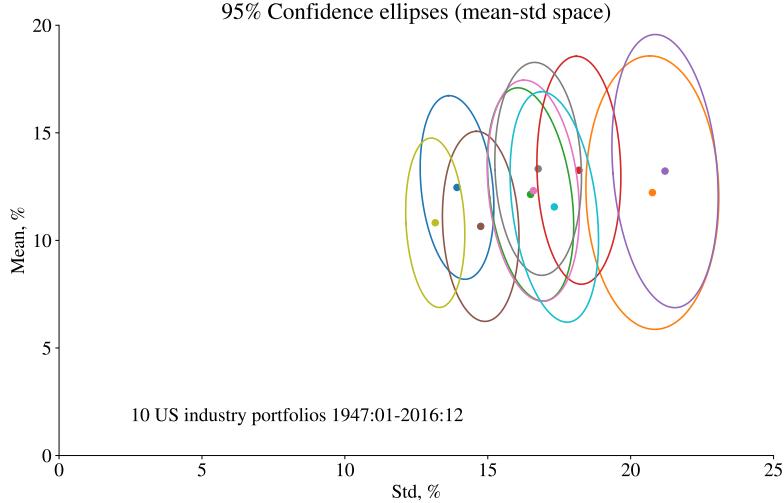


Figure 1.1: Confidence ellipses for estimating std and mean returns for a number of different portfolios

The test of the joint hypothesis is based on

$$T(R\beta - \gamma_0)' \Lambda^{-1} (R\beta - \gamma_0) \xrightarrow{d} \chi_q^2. \quad (1.4)$$

The appendix has tables for critical values of the χ_q^2 distribution (for different values of q).

Remark 1.2 (*Confidence ellipse**) A 95% confidence ellipse for γ_0 is defined as the set of those γ_0 vectors that satisfy $T(R\beta - \gamma_0)' \Lambda^{-1} (R\beta - \gamma_0) \leq c$ where c is the 95% critical value of a χ_q^2 distribution. Such ellipses are particularly often used when γ_0 is a vector with two elements. See Figure 1.1 for an illustration.

1.2 Testing (Nonlinear) Joint Hypotheses: The Delta Method

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

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_{k \times k}), \quad (1.5)$$

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

$$\gamma_{q \times 1} = f(\beta), \quad (1.6)$$

where $f(\cdot)$ has continuous first derivatives. Under that null hypothesis (that $\gamma = \gamma_0$)

$$\begin{aligned} & \sqrt{T}(f(\hat{\beta}) - \gamma_0) \xrightarrow{d} N(0, \Lambda_{q \times q}), \text{ where} \\ & \Lambda = \frac{\partial f(\beta_0)}{\partial \beta'} V \frac{\partial f(\beta_0)'}{\partial \beta}, \text{ where} \\ & \frac{\partial f(\beta)}{\partial \beta'} = \begin{bmatrix} \frac{\partial f_1(\beta)}{\partial \beta_1} & \dots & \frac{\partial f_1(\beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q(\beta)}{\partial \beta_1} & \dots & \frac{\partial f_q(\beta)}{\partial \beta_k} \end{bmatrix}_{q \times k} \end{aligned} \quad (1.7)$$

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 (1.4).

Example 1.3 (*Testing a Sharpe ratio*) Stack the mean ($\mu = E x_t$) and second moment ($\mu_2 = E x_t^2$) as $\beta = [\mu, \mu_2]'$. The Sharpe ratio is calculated as a function of β

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

If $\hat{\beta}$ is distributed as in (1.5), then (1.7) is straightforward to apply.

Example 1.4 (*Linear function*) When $f(\beta) = R\beta$, then the Jacobian is $\frac{\partial f(\beta)}{\partial \beta'} = R$, so $\Lambda = RVR'$, just like in (1.3).

Example 1.5 (*Testing a correlation of x_t and y_t*) Suppose you have estimated the variances of (x_t, y_t) and also their covariance. Stack the parameters in the vector $\beta = [\sigma_{xx}, \sigma_{yy}, \sigma_{xy}]'$. The correlation and the Jacobian is then

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

Remark 1.6 (*Numerical derivatives*) A crude way to calculate a numerical derivative is a forward difference

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

A value of $h = 10^{-8} \max(|\beta_j|, 1)$ is often recommended. It is sometimes better to use a

central difference

$$\begin{bmatrix} \frac{\partial f_1(\beta)}{\partial \beta_j} \\ \vdots \\ \frac{\partial f_q(\beta)}{\partial \beta_j} \end{bmatrix} = \frac{f(\tilde{\beta}) - f(\beta^*)}{2h}, \text{ where } \tilde{\beta} = \beta^* = \beta \text{ except that } \begin{cases} \tilde{\beta}_j = \beta_j + h \\ \beta_j^* = \beta_j - h. \end{cases}$$

In this case, $h = 10^{-6} \max(|\beta_j|, 1)$ is often recommended.

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

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

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

1.2.2 Delta Method Example: 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 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.

1.2.3 Delta Method Example: Testing the Optimal Portfolio Weight

A mean-variance investor combines the riskfree asset with a mix (the tangency portfolio) of risky assets. The optimal portfolio weight on the latter is $w = E R_m^e / [k \text{Var}(R_m^e)]$, where R_m^e denotes the excess return of the tangency portfolio. If we have estimates and their covariance matrix of $E R_m^e$ and $\text{Var}(R_m^e)$, then it is straightforward to construct a confidence band around w . See Figure 1.3.

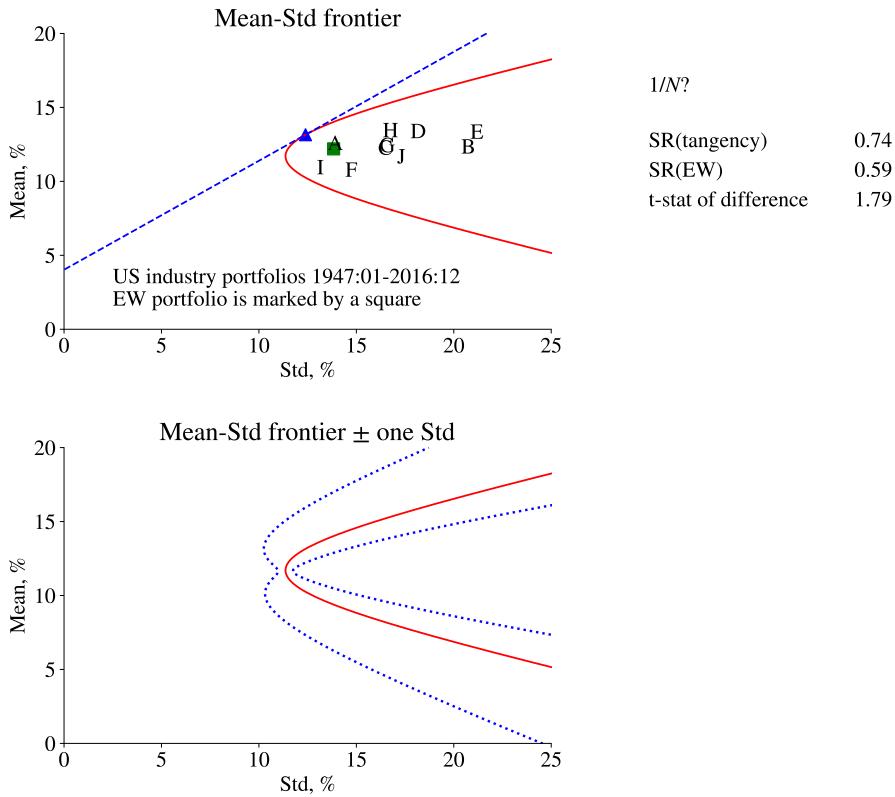


Figure 1.2: Mean-Variance frontier of US industry portfolios. Monthly returns are used in the calculations, but $100\sqrt{12}\text{Variance}$ is plotted against $100 * 12*\text{mean}$.

1.3 The Variance of a Sample Average

Many estimators (including OLS, MLE and 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. For a time series average, the [Newey and West \(1987\)](#) estimator is probably the most popular.

To illustrate the idea, consider a time series sample mean, \bar{x} , of a $K \times 1$ vector x_t

$$\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t. \quad (1.8)$$

If x_t is iid, then

$$\text{Cov}(\sqrt{T}\bar{x}) = \text{Cov}(x_t), \quad (1.9)$$

which is a $K \times K$ matrix. This clearly is the same as saying that $\text{Cov}(\bar{x}) = \text{Cov}(x_t)/T$.

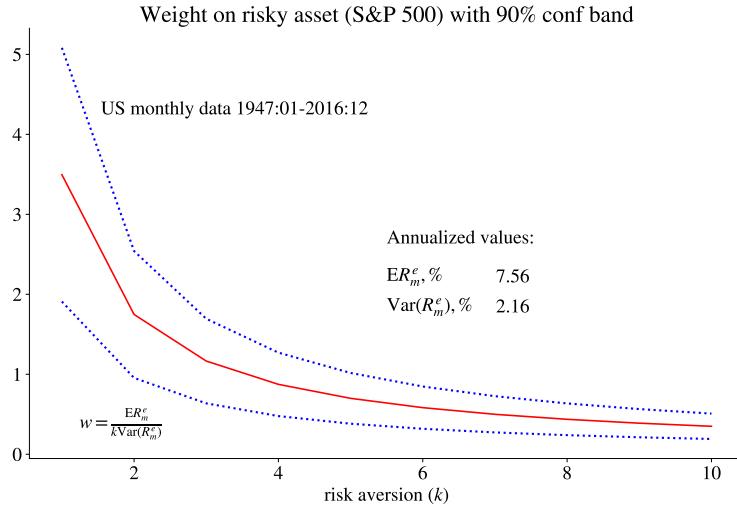


Figure 1.3: Portfolio choice for different risk aversions, with confidence band

Instead, if x_t is autocorrelated, then

$$\text{Cov}(\sqrt{T}\bar{x}) = \sum_{s=-(T-1)}^{T-1} \left(1 - \frac{|s|}{T}\right) \Gamma(s), \text{ where } \Gamma(s) = \text{Cov}(x_t, x_{t-s}), \quad (1.10)$$

where $\Gamma(s)$ is a $K \times K$ matrix, where cell (i, j) is the covariance between element i of x_t and element j of x_{t-s} .

Example 1.7 ($\Gamma(s)$ for a vector with two variables) If $x_t = [x_{1t}, x_{2t}]'$ where x_{1t} is one variable and x_{2t} is another, then

$$\Gamma(s) = \begin{bmatrix} \text{Cov}(x_{1,t}, x_{1,t-s}) & \text{Cov}(x_{1,t}, x_{2,t-s}) \\ \text{Cov}(x_{2,t}, x_{1,t-s}) & \text{Cov}(x_{2,t}, x_{2,t-s}) \end{bmatrix}.$$

Proof. (of (1.10)) Notice that for $T = 3$, we have

$$\begin{aligned} \text{Var}(x_1 + x_2 + x_3) &= \underbrace{\text{Cov}(x_1, x_3)}_{\Gamma(-2)} + \underbrace{\text{Cov}(x_1, x_2) + \text{Cov}(x_2, x_3)}_{2\Gamma(-1)} + \\ &\quad \underbrace{\text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3)}_{3\Gamma(0)} + \underbrace{\text{Cov}(x_2, x_1) + \text{Cov}(x_3, x_2)}_{2\Gamma(1)} + \underbrace{\text{Cov}(x_3, x_1)}_{\Gamma(2)}. \end{aligned}$$

The general pattern is

$$\text{Var} \left(\sum_{t=1}^T x_t \right) = \sum_{s=-(T-1)}^{T-1} (T - |s|) \Gamma(s).$$

Divide both sides by T to get (1.10). ■

Example 1.8 (*Variance of sample mean of AR(1)*.) Let $x_t = \rho x_{t-1} + u_t$, where $\text{Var}(u_t) = \sigma^2$. Let $\Gamma(s)$ denote the s th autocovariance and notice that $\Gamma(s) = \rho^{|s|}\sigma^2 / (1 - \rho^2)$. The asymptotic (as $T \rightarrow \infty$ so $|s|/T \rightarrow 0$ in (1.10)) variance can be written

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

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 $\sigma^2 / (1 - \rho^2)$, 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 1.4 for an illustration. Notice that $\text{Var}(\sqrt{T}\bar{x}) / \text{Var}(x_t) = \text{Var}(\bar{x}) / [\text{Var}(x_t)/T]$, so the ratio also shows the relation between the true variance of \bar{x} and the classical estimator of it (based of the iid assumption).

The Newey-West estimator of the variance-covariance matrix of $\sqrt{T}\bar{x}$ is

$$\widehat{\text{Cov}}(\sqrt{T}\bar{x}) = \sum_{s=-n}^n \left(1 - \frac{|s|}{n+1} \right) \widehat{\text{Cov}}(x_t, x_{t-s}), \quad (1.11)$$

where n is a finite “bandwidth” parameter. The “weights,” $1 - |s|/(n+1)$, are clearly tent-shaped: 1 at the zero lag—and lower as the lags become longer. Figure 1.5 illustrates the weights (the term in parentheses in (1.11)) for different choices of the bandwidth (n). This is similar to (1.10), but the weights decrease quicker (assuming $n < T-1$). This suggests that n should be somewhat larger than last lag with significant autocorrelation. Alternatively, a common rule of thumb is $n = \text{floor}(0.75T^{1/3})$, where $\text{floor}()$ means rounding down to nearest integer (sometimes $n = \text{floor}(4(T/100)^{2/9})$ is used instead).

Example 1.9 (*Newey-West estimator*) With $n = 1$ in (1.11) the Newey-West estimator

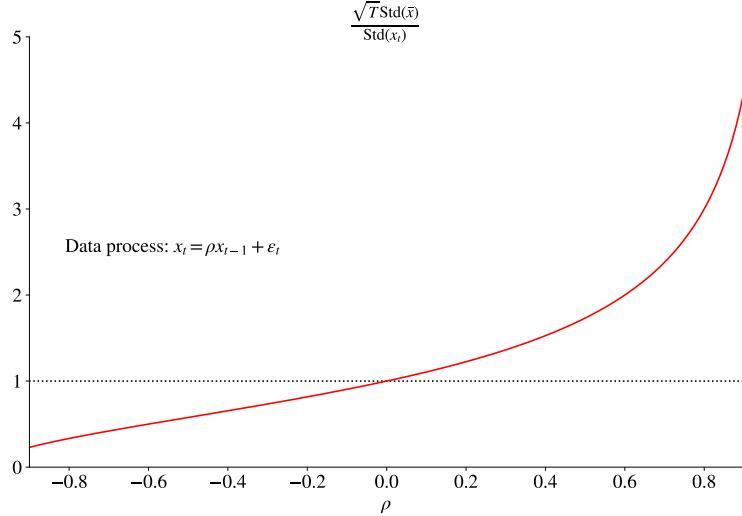


Figure 1.4: Variance of $\sqrt{T} \times$ sample average of an AR(1) series

becomes

$$\widehat{\text{Cov}}(\sqrt{T}\bar{x}) = \frac{1}{2}\widehat{\text{Cov}}(x_t, x_{t+1}) + \widehat{\text{Cov}}(x_t, x_t) + \frac{1}{2}\widehat{\text{Cov}}(x_t, x_{t-1}).$$

Remark 1.10 (VARHAC*) The VARHAC estimator of the covariance matrix (see Andrews and Monahan (1992)) is as follows. First, fit a VAR(p) to x_t

$$x_t = A_0 + \sum_{i=1}^p A_i x_{t-i} + \varepsilon_t$$

and calculate $D = I - \sum_{i=1}^p A_i$. Then, use $\text{Cov}(\sqrt{T}\bar{x}) = D^{-1}S^*D^{-1}$, where S^* is Newey-West estimate of $\text{Cov}(\sqrt{T}\bar{\varepsilon})$. As an example, let x_t be a scalar that follows an AR(1) process, $x_t = \rho x_{t-1} + \varepsilon_t$. If ε_t is iid, then $\text{Cov}(\sqrt{T}\bar{\varepsilon}) = \sigma^2$ where σ^2 is the variance of ε_t . $D = 1 - \rho$, so $\text{Cov}(\sqrt{T}\bar{x}) = \sigma^2/(1 - \rho)^2$ which is the same as the variance in Example 1.8 (since $(1 - \rho^2)/(1 + \rho) = 1 - \rho$).

Remark 1.11 (Cross-sectional averages) The insight that correlations matter for an average applies also to a cross-sectional average. The only difference is that it is harder to motivate why the variances should be the same across observations. As an example, consider the cross-sectional average return (in period t) across n assets, $\bar{R}_t = \sum_{i=1}^n R_{i,t}/n$. It is clear that $\text{Var}(\bar{R}_t) = \mathbf{1}'\Sigma\mathbf{1}/n^2$, where $\mathbf{1}$ is an $n \times 1$ vector of ones and Σ is the covariance matrix of the n assets. This is just the sum of all elements, divided by n^2 , which is very similar to (1.10), although we are here studying a cross-section, not a time series.

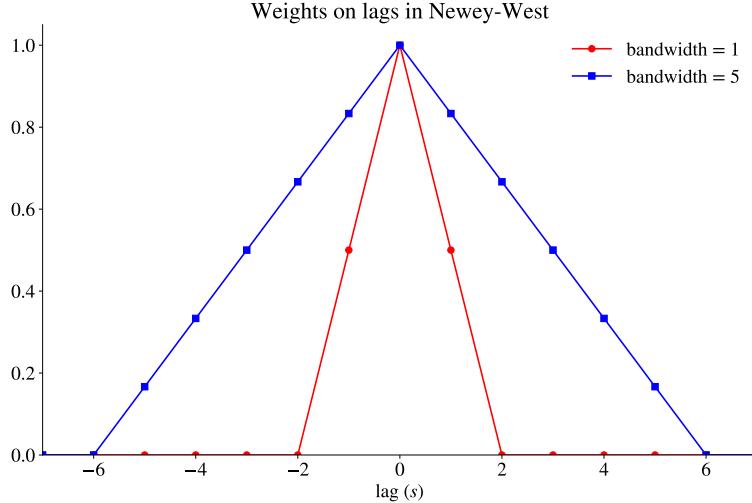


Figure 1.5: Weights in Newey-West

1.4 OLS

Consider the linear regression model

$$y_t = x'_t \beta + u_t, \quad (1.12)$$

where x_t and β are $k \times 1$ vectors. Least squares minimizes the sum of the squared fitted residuals

$$\sum_{t=1}^T (y_t - x'_t b)^2, \quad (1.13)$$

by choosing the vector b . The first order conditions (zero derivatives with respect to b) hold at the values $\hat{\beta}$

$$\hat{\beta} = \Sigma_{xx}^{-1} \Sigma_{xy}, \text{ where } \Sigma_{xx} = \sum_{t=1}^T x_t x'_t / T \text{ and } \Sigma_{xy} = \sum_{t=1}^T x_t y_t / T. \quad (1.14)$$

Under the *Gauss-Markov assumptions* (the residuals have zero means, constant variances and are not correlated across observations) the asymptotic distribution of the estimates is

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}, \text{plim } \Sigma_{xx}^{-1} \sigma^2), \quad (1.15)$$

where σ^2 is the variance of the residuals (u_t). Due to a central limit theorem, this result typically holds even if the residuals have a non-normal distribution. However, if the residuals have time-varying variances (heteroskedasticity) or are autocorrelated, then the

covariance matrix in (1.15) may need to be adjusted (by applying White's or Newey-West's covariance estimator).

1.5 MLE

1.5.1 Basic MLE

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

$$L = \text{pdf}(y_1, y_2, \dots, y_T; \theta) \quad (1.16)$$

$$= L_1 \times L_2 \times \dots \times L_T, \quad (1.17)$$

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, assuming we can split up the overall density function—which is possible if x_t and x_{t-1} are independent. For notational convenience, the dependence of L_t on 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

$$\hat{\theta} = \arg \max \ln L. \quad (1.18)$$

MLE is typically asymptotically normally distributed

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V), \text{ where } V = I(\theta)^{-1} \text{ with} \quad (1.19)$$

$$I(\theta) = -E \frac{\partial^2 \ln L}{\partial \theta \partial \theta'} / T \text{ or} \quad (1.20)$$

$$= -E \frac{\partial^2 \ln L_t}{\partial \theta \partial \theta'}, \quad (1.21)$$

where $I(\theta)$ is the “information matrix.” Notice that in the second line, the derivative is of the whole log likelihood function (1.16) so we divide by T , 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

$$J(\theta) = E \left[\frac{\partial \ln L_t}{\partial \theta} \frac{\partial \ln L_t}{\partial \theta'} \right]. \quad (1.22)$$

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. On the other hand, a major drawback of MLE is that it requires that we know what the right density (likelihood) function is.

1.5.2 QMLE

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

$$\frac{\partial \ln L}{\partial \theta} = \mathbf{0}. \quad (1.23)$$

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 (1.23) as a set of GMM moment conditions—and the usual GMM results apply. In fact, this quasi-MLE (or pseudo-MLE) has the same sort of distribution as in (1.19), but with the variance-covariance matrix

$$V = I(\theta)^{-1} J(\theta) I(\theta)^{-1}. \quad (1.24)$$

Example 1.12 (LS and QMLE) *In a linear regression, $y_t = x_t' \beta + \varepsilon_t$, the first order condition for MLE based on the assumption that $\varepsilon_t \sim N(0, \sigma^2)$ is $\sum_{t=1}^T (y_t - x_t' \hat{\beta}) x_t = \mathbf{0}$. This has an expected value of zero (at the true parameters), even if the shocks have a, say, t_{22} distribution.*

1.5.3 MLE Example: Estimate the Variance

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

$$\text{pdf}(x_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{x_t^2}{\sigma^2}\right). \quad (1.25)$$

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

$$\begin{aligned} L &= \text{pdf}(x_1) \times \text{pdf}(x_2) \times \dots \times \text{pdf}(x_T) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^T \frac{x_t^2}{\sigma^2}\right), \text{ so} \end{aligned} \quad (1.26)$$

$$\ln L = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^T x_t^2. \quad (1.27)$$

The first order condition for optimum is

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma^2} &= -\frac{T}{2} \frac{1}{2\pi\sigma^2} 2\pi + \frac{1}{2(\sigma^2)^2} \sum_{t=1}^T x_t^2 = 0 \text{ so} \\ \hat{\sigma}^2 &= \sum_{t=1}^T x_t^2 / T.\end{aligned}\quad (1.28)$$

Differentiate the log likelihood once again to get

$$\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} = \frac{T}{2} \frac{1}{\sigma^4} - \frac{1}{(\sigma^2)^3} \sum_{t=1}^T x_t^2, \text{ so} \quad (1.29)$$

$$E \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} = \frac{T}{2} \frac{1}{\sigma^4} - \frac{T}{(\sigma^2)^3} \sigma^2 = -\frac{T}{2\sigma^4} \quad (1.30)$$

The information matrix is therefore

$$I(\theta) = -E \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} / T = \frac{1}{2\sigma^4}, \quad (1.31)$$

so we have

$$\sqrt{T}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4). \quad (1.32)$$

1.6 GMM

1.6.1 The Basic GMM

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

$$\bar{g}(\beta) = \frac{1}{T} \sum_{t=1}^T g_t(\beta) = \mathbf{0}_{q \times 1}, \quad (1.33)$$

where $\bar{g}(\beta)$ is short hand notation for the sample average. The notation $g_t(\beta)$ is meant to show that moments conditions depend on the parameter vector (β) and on data for period t . We let β_0 denote the true value of the $k \times 1$ parameter vector.

The GMM estimator is

$$\hat{\beta}_{k \times 1} = \arg \min \bar{g}(\beta)' W \bar{g}(\beta), \quad (1.34)$$

where W is some symmetric positive definite $q \times q$ weighting matrix. When the model is exactly identified ($q = k$), then we do not have to perform an explicit minimization, since all sample moment conditions can be set equal to zero (there are as many parameters as there are moment conditions).

It can be shown that choosing $W = S_0^{-1}$, where S_0 is the covariance matrix of

$\sqrt{T}\bar{g}(\beta_0)$ evaluated at the true parameter values, gives the most efficient estimates (for a given set of moment conditions). To approximate this, an iterative procedure is often used: start with $W = I_q$ (or some other reasonable weighting matrix), estimate the parameters, estimate 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 you could also continue iterating until the point estimates converge.

Example 1.13 (*Moment condition for a mean*) To estimate the mean of x_t , use

$$g_t = x_t - \mu.$$

There is one parameter and one moment condition: exactly identified.

Example 1.14 (*Moments conditions for OLS*) Consider the linear model $y_t = x'_t \beta_0 + u_t$, where x_t and β are $k \times 1$ vectors. The k moments are

$$g_t = x_t(y_t - x'_t \beta).$$

There are as many parameters as moment conditions: exactly identified.

Example 1.15 (*Moment conditions for estimating a normal distribution*) Suppose you specify four moments for estimating the mean and variance of a normal distribution

$$g_t = \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix}$$

This case is overidentified ($q = 4$ and $k = 2$), so a weighting matrix is needed.

Example 1.16 (*Moment conditions for variances and a covariance*) For expositional simplicity, assume that both variables have zero means. The variances and the covariance can then be estimated by the moment conditions

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

1.6.2 Distribution of the Basic GMM

GMM estimators are typically asymptotically normally distributed, with a covariance matrix that depends on the covariance matrix of the moment conditions (S_0) and the mapping from the parameters to the moment conditions (D_0). The details of these matrices are discussed below. For now, notice that the distribution of the GMM estimates is

$$\begin{aligned}\sqrt{T}(\hat{\beta} - \beta_0) &\xrightarrow{d} N(0, V) \text{ if } W = S_0^{-1}, \text{ where} \\ V &= (D_0' S_0^{-1} D_0)^{-1}. \end{aligned}\quad (1.35)$$

This result assumes that we have used S_0^{-1} as the weighting matrix ($W = S_0^{-1}$) in (1.34). The choice of the weighting matrix is irrelevant if the model is exactly identified, so (1.35) can be applied to this case (even if we did not specify any weighting matrix at all). It can also be noticed that when the model is exactly identified, then we can typically rewrite the covariance matrix as $V = D_0^{-1} S_0 (D_0^{-1})'$, which might be easier to calculate.

Let S_0 be the $(q \times q)$ covariance matrix of $\sqrt{T}\bar{g}(\beta_0)$, evaluated at the true parameter values

$$S_0 = \text{Cov}[\sqrt{T}\bar{g}(\beta_0)], \quad (1.36)$$

where $\text{Cov}()$ is a matrix of covariances. When there is no autocorrelation of the moments, then (1.36) becomes

$$S_0 = \text{Cov}[g_t(\beta_0)], \text{ if } g_t \text{ is not autocorrelated.} \quad (1.37)$$

When there is autocorrelation, then we may use the Newey-West approach to estimate S_0 .

In practice, S_0 is estimated by using the estimated coefficients in the moments to get the data series $g_t(\hat{\beta})$, a $T \times q$ matrix, from which we estimate the covariances needed for (1.36) or (1.37).

Example 1.17 (*Estimating a mean, variance*) The moment in Example 1.13 (assuming iid data, so we can use (1.37)) gives

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

In practice, we replace the variance by a sample estimate. If we suspect that x_t is autocorrelated, then we may use the NW estimator of $\text{Var}(\sqrt{T}\bar{g})$.

Example 1.18 (*OLS, covariance*) For the moments in Example 1.14, using $u_t = y_t - x_t'\beta$,

we have

$$S_0 = \text{Cov} \left[\frac{\sqrt{T}}{T} \sum_{t=1}^T x_t u_t \right]$$

In practice, replace u_t by the fitted residuals and calculate a sample covariance. It can be shown that under the Gauss-Markov assumptions $S_0 = \sigma^2 \Sigma_{xx}$. If we suspect that the variance of u_t is related to x_t , then we should calculate the covariance matrix of g_t , which gives White's covariance estimator. In addition we suspect that g_t is autocorrelated, then we may use the NW estimator of $\text{Var}(\sqrt{T}\bar{g})$.

Example 1.19 (Estimating/testing a normal distribution, covariance) For the moments in Example 1.15 (assuming iid normally distributed data, so we can use (1.37)), it can be shown that we have

$$S_0 = \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}.$$

In practice, we would use the point estimates in the moments and calculate the sample covariance matrix. If we suspect that g_t is autocorrelated, then we may use the NW estimator of $\text{Var}(\sqrt{T}\bar{g})$.

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)

$$D_0 = \text{plim} \frac{\partial \bar{g}(\beta_0)}{\partial \beta'}. \quad (1.38)$$

In practice, the gradient D_0 is approximated by using the point estimates and the available sample of data.

Remark 1.20 (Jacobian) The Jacobian is of the following format

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

Example 1.21 (Estimating a mean, Jacobian) For the moment in Example 1.13

$$D_0 = \frac{\partial}{\partial \mu} \frac{1}{T} \sum_{t=1}^T (x_t - \mu) = -1,$$

which does not involve any parameters or any data.

Example 1.22 (OLS, Jacobian) For the moments in Example 1.14

$$D_0 = \text{plim} \left(-\frac{1}{T} \sum_{t=1}^T x_t x_t' \right) = -\Sigma_{xx}.$$

This does not contain any parameters either, but includes data. In practice, we replace Σ_{xx} by a sample estimate.

Example 1.23 (Estimating/testing a normal distribution, covariance) For the moments in Example 1.15 (assuming iid normally distributed data) we have (the rows are for the four different moment conditions, the columns for the parameters: μ and σ^2)

$$\begin{aligned} D_0 &= \text{plim} \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} -1 & 0 \\ -2(x_t - \mu) & -1 \\ -3(x_t - \mu)^2 & 0 \\ -4(x_t - \mu)^3 & -6\sigma^2 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}. \end{aligned}$$

Element (4,1) of the second equality holds only if the data has a symmetric distribution (for instance, a normal distribution). In practice, we would use the point estimates in the matrix on the first line and calculate the sample average.

Example 1.24 (Estimating a mean, distribution) For the moment condition in Example 1.13 we have (assuming iid data)

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

Example 1.25 (OLS, distribution) For the moment conditions in Example 1.14

$$V = (\Sigma_{xx} S_0^{-1} \Sigma_{xx})^{-1}.$$

Under the Gauss-Markov assumptions $S_0 = \sigma^2 \Sigma_{xx}$, so

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

Example 1.26 (Estimating/testing a normal distribution, distribution) For the moment conditions in Example 1.15 (assuming iid normally distributed data) we have that the asymptotic covariance matrix of the estimated mean and variance is then $((D_0' S_0^{-1} D_0)^{-1})$

$$\left(\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}' \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix} \right)^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

In an overidentified model ($k < q$), we can test if the k parameters make all q moment conditions hold. Notice that under the null hypothesis (that the model is correctly specified)

$$\sqrt{T} \bar{g}(\beta_0) \xrightarrow{d} N(\mathbf{0}_{q \times 1}, S_0), \quad (1.39)$$

where q is the number of moment conditions. Since $\hat{\beta}$ is chosen in such a way that k linear combinations of the moment conditions are zero, there are effectively only $q - k$ non-degenerate random variables. We can therefore test the hypothesis that $\bar{g}(\beta_0) = 0$ by the “J test”

$$T \bar{g}(\hat{\beta})' S_0^{-1} \bar{g}(\hat{\beta}) \xrightarrow{d} \chi_{q-k}^2, \text{ if } W = S_0^{-1}. \quad (1.40)$$

The left hand side equals T times of value of the loss function in (1.34) evaluated at the point estimates. With no overidentifying restrictions ($q = k$) there are no restrictions to test. Indeed, the loss function value is then always zero at the point estimates.

Example 1.27 (Estimating/testing a normal distribution, testing) After having estimated the mean and the variance, we can test if all four moment conditions in Example 1.15 hold. If data is drawn from a normal distribution, they should hold (give and take some randomness).

1.6.3 GMM with a Suboptimal Weighting Matrix

The distribution of the GMM estimates when we use a sub-optimal weighting matrix is similar to (1.35), but the variance-covariance matrix is different (basically, reflecting the fact that the approach does not produce the lowest possible variances anymore).

Example 1.28 (Estimating/testing a normal distribution) Example 1.15 is overidentified since there are four moment conditions but only two parameters. Instead of using the

optimal weighting matrix (the inverse of S_0 from Example 1.19, assuming the data is iid normally distributed), we could use any other (positive definite) 4×4 matrix. For instance, $W = I_4$ or a matrix that puts almost all weight on the first two moment conditions.

It can be shown that if we use another weighting matrix than $W = S_0^{-1}$, then the variance-covariance matrix in (1.35) should be changed to

$$\begin{aligned} V_2 &= V_{A2} D_0' W S_0 W' D_0 V_{A2}', \text{ where} \\ V_{A2} &= (D_0' W D_0)^{-1}. \end{aligned} \quad (1.41)$$

Similarly, the test of overidentifying restrictions becomes

$$T \bar{g}(\hat{\beta})' \Psi_2^+ \bar{g}(\hat{\beta}) \xrightarrow{d} \chi_{q-k}^2, \quad (1.42)$$

where Ψ_2^+ is a generalized inverse of

$$\begin{aligned} \Psi_2 &= \Psi_{A2} S_0 \Psi_{A2}', \text{ where} \\ \Psi_{A2} &= I_q - D_0 (D_0' W D_0)^{-1} D_0' W. \end{aligned} \quad (1.43)$$

The covariance matrix Ψ_2 has a reduced rank, so we must use a generalized inverse in the test.

Remark 1.29 (*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 $Y = X' \Sigma^+ X \sim \chi_r^2$ where Σ^+ is the pseudo inverse of Σ .

Example 1.30 (*Pseudo inverse of a square matrix*) For the matrix

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

1.6.4 GMM without a Loss Function

Suppose we sidestep the whole optimization issue and instead specify k linear combinations of the q moment conditions directly

$$\mathbf{0}_{k \times 1} = \underbrace{A}_{k \times q} \underbrace{\bar{g}(\hat{\beta})}_{q \times 1}, \quad (1.44)$$

where the matrix A is chosen by the researcher.

Example 1.31 (*Overidentified example: estimating/testing a normal distribution*) Example 1.15 is overidentified since there are four moment conditions but only two parameters. One possible A matrix would put all weight on the first two moment conditions

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

It is straightforward to show that the variance-covariance matrix in (1.35) should be changed to

$$\begin{aligned} V_3 &= V_{A3} A_0 S_0 A_0' V_{A3}', \text{ where} \\ V_{A3} &= (A_0 D_0)^{-1}, \end{aligned} \quad (1.45)$$

where A_0 is the probability limit of A (if it is random).

Similarly, in the test of overidentifying restrictions (1.42), we should replace Ψ_2 by

$$\begin{aligned} \Psi_3 &= \Psi_{A3} S_0 \Psi_{A3}', \text{ where} \\ \Psi_{A3} &= I_q - D_0 (A_0 D_0)^{-1} A_0. \end{aligned} \quad (1.46)$$

The covariance matrix Ψ_3 has a reduced rank, so we must again use a generalized inverse in the test.

Example 1.32 (*Estimating/testing a normal distribution*) Continuing Example 1.31, we have that $A_0 D_0$ in (1.45) is

$$V_{A3} = \left(\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0} \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \right)^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Example 1.33 (*Estimating/testing a normal distribution*) Continuing the previous exam-

ple, Ψ_{A3} in (1.46) is

$$\begin{aligned}\Psi_{A3} &= \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{I_4} - \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \left(\underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}}_{A_0 D_0} \right)^{-1} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0} \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}.\end{aligned}$$

Ψ_3 in (1.46) is therefore

$$\begin{aligned}\Psi_3 &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}' \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 6\sigma^6 & 0 \\ 0 & 0 & 0 & 24\sigma^8 \end{bmatrix}\end{aligned}$$

Example 1.34 (Estimating/testing a normal distribution) Continuing the previous example, we have that the test of the overidentifying restrictions (1.42) (assuming iid normally distributed data to calculate S_0) is (notice the generalized inverse of Ψ_3)

$$\begin{aligned}&= T \begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^T (x_t - \mu)^3 / T \\ \Sigma_{t=1}^T [(x_t - \mu)^4 - 3\sigma^4] / T \end{bmatrix}' \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/(6\sigma^6) & 0 \\ 0 & 0 & 0 & 1/(24\sigma^8) \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^T (x_t - \mu)^3 / T \\ \Sigma_{t=1}^T [(x_t - \mu)^4 - 3\sigma^4] / T \end{bmatrix} \\ &= \frac{T}{6} \frac{[\Sigma_{t=1}^T (x_t - \mu)^3 / T]^2}{\sigma^6} + \frac{T}{24} \frac{\{\Sigma_{t=1}^T [(x_t - \mu)^4 - 3\sigma^4] / T\}^2}{\sigma^8}.\end{aligned}$$

When we replace μ and σ by their estimates, then this is the same as the Jarque-Bera test of normality.

1.6.5 GMM Example: 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

$$\mathbb{E} R_t - \mu = \mathbf{0}_{N \times 1}, \quad (1.47)$$

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

$$\mathbb{E} \text{vech}(R_t R'_t) - \text{vech}(\Gamma) = \mathbf{0}_{N(N+1)/2 \times 1}. \quad (1.48)$$

Remark 1.35 (*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,*

$$\text{vech} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}.$$

Stack (1.47) and (1.48) and substitute the sample mean for the population expectation to get the GMM estimator

$$\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} R_t \\ \text{vech}(R_t R'_t) \end{bmatrix} - \begin{bmatrix} \hat{\mu} \\ \text{vech}(\hat{\Gamma}) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N \times 1} \\ \mathbf{0}_{N(N+1)/2 \times 1} \end{bmatrix}. \quad (1.49)$$

In this case, $D_0 = -I$, so the covariance matrix of the parameter vector $(\hat{\mu}, \text{vech}(\hat{\Gamma}))$ is just S_0 (defined in (1.36)), which is straightforward to estimate.

1.7 Appendix: A Primer on Using Numerical Optimization Routines*

Reference: Brandimarte (2006)

1.7.1 Unconstrained Minimization

Consider the loss function

$$f(\theta) = (x - 2)^2 + (4y + 3)^2, \quad (1.50)$$

where $\theta = (x, y)$ contains the two choice variables.

A numerical minimization routine searches different values of θ , typically starting from a guess supplied by the user, to find the values that makes $f(\theta)$ as small as possible. (The correct solution is $(x, y) = (2, -3/4)$.) Convergence criteria (often set by the user) determine when the search will stop (for instance, when the improvement in $f(\theta)$ is smaller than a certain threshold or when the θ values do not change much anymore). The starting guess is often important, so be sure to use reasonable values.

There are two main types of algorithms: those that use derivatives of the loss function (which needs to be coded by the user) as extra information and those that do not. The latter type is often slower, but sometimes more robust.

Most optimization algorithms are for minimizing a function value. In case you want to maximize, then just change the sign of the function and then minimize it. For instance, if you want to maximize $g(\theta)$, then you can do that by minimizing $-g(\theta)$.

1.7.2 Equality Constraints

If you want to add an *equality constraint* to the minimization problem, say

$$h_1(\theta) = x + 2y - 3 = 0, \quad (1.51)$$

then there are several possible ways to proceed. The best is perhaps to use the constraint to rewrite the loss function (in this case, we would use $x = 3 - 2y$ to replace x in (1.50)). If this is tricky, then we try to find a routine that can handle equality constraints. Finally (if there are no good routines available), we could construct such a routine ourselves.

The idea is to apply a penalty for deviations from the constraint, so the overall loss function becomes

$$f(\theta) + \lambda \sum_{i=1}^p h_i(\theta)^2, \quad (1.52)$$

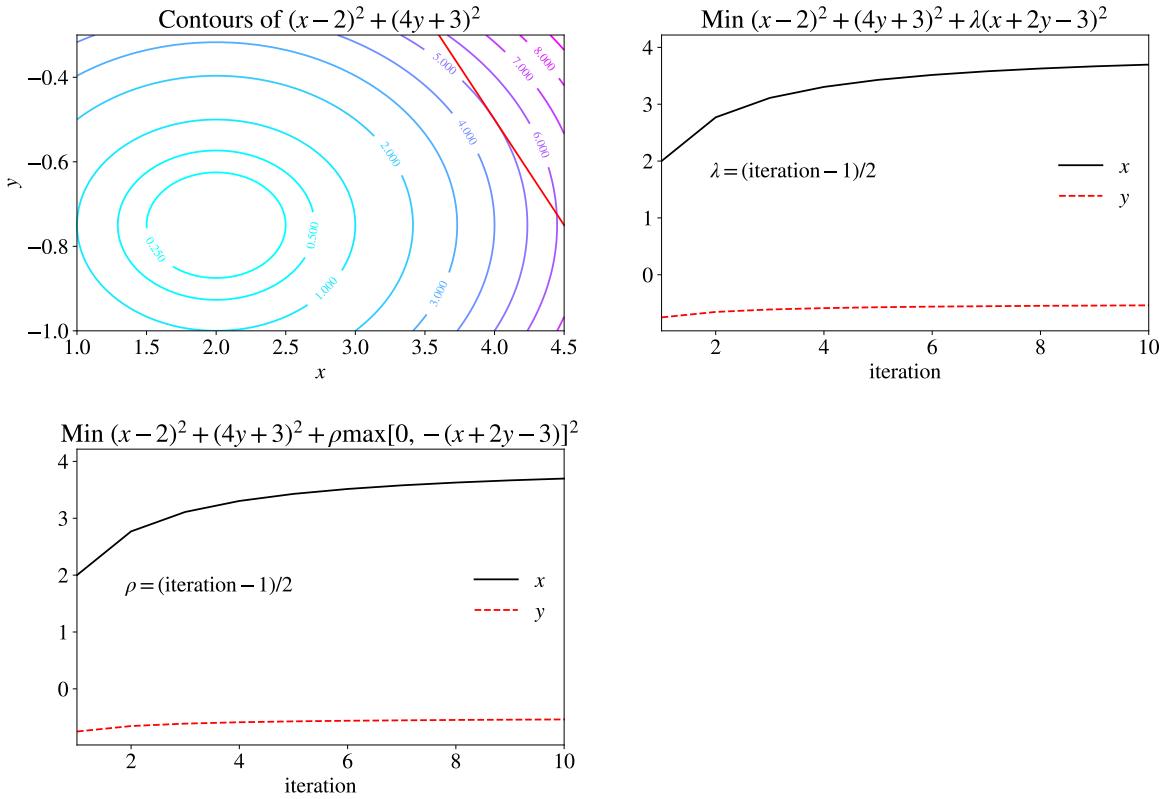


Figure 1.6: Numerical optimization with restrictions

where $h_i(\theta)^2$ is the square of the i th equality constraint. This expression allows for p different constraints (there is only one in (1.51)).

Start by setting $\lambda = 0$ and find the optimal value of θ , and call it θ_1 . Then, set $\lambda = 5$ and redo the optimization (using θ_1 as the starting guess) to get the optimal value θ_2 . Now, set $\lambda = 10$ and redo the optimization (using θ_2 as the starting guess). Keep doing this (at higher and higher values of λ) until the solutions do not change much anymore. It is often worthwhile to experiment a bit with the sequence of λ values. (In our case the solution should be very close to $(x, y) = (4, -1/2)$.)

See Figure 1.6 for an example.

1.7.3 Inequality Constraints

Instead, we now want to minimize (1.50) under the *inequality constraint*

$$g_1(\theta) = -(x + 2y - 3) \leq 0. \quad (1.53)$$

You can always rewrite a \geq inequality on the \leq form by multiplying both sides by -1 . This restriction says that $x + 2y \geq 3$, so (in this case) it should give the same solution to (1.50) as the equality restriction (1.51)

To do this we either find a routine that does the job, or we create one ourselves. It is the same ideas as for the equality constraints, except that we now use the overall loss function

$$f(\theta) + \rho \sum_{j=1}^q \max[0, g_j(\theta)]^2, \quad (1.54)$$

where ρ plays the same role as λ : start by solving for $\rho = 0$, then use that solution as a starting guess for the problem with $\rho = 5$, etc. See Figure 1.6 for an example.

Finally, we can combine equality and inequality constraints as

$$f(\theta) + \lambda \sum_{i=1}^p h_i(\theta)^2 + \rho \sum_{j=1}^q \max[0, g_j(\theta)]^2. \quad (1.55)$$

1.8 Appendix: Statistical Tables

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

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

1.9 Appendix: Data Sources

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

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

Table 1.2: Critical values of chi-square distribution (different degrees of freedom, *n*).

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

Chapter 2

Basic Asset Pricing Theory

References: Back (2010), Cochrane (2005) and Pennacchi (2008)

These notes summarise a large number of results in theoretical asset pricing. Proofs are given if they are short and easy. Otherwise, see the references.

2.1 Three Pricing Principles

The following definitions will be useful later.

Remark 2.1 (*Complete markets*) *Markets are complete if there are sufficiently many assets so that you could hedge against any possible outcome. For instance, in a binomial model (where the stock price can take a jump up or down), then two assets are required: a stock and a bond (or a stock and an option, or...)*

Remark 2.2 (*Law of one price*) *The price of a portfolio is the portfolio of the prices. This rules out trivial arbitrage.*

Remark 2.3 (*No arbitrage*) *Every asset whose payoff is always nonnegative and sometimes positive, has a positive price. This rules out a free lunch: you cannot get something for nothing.*

Remark 2.4 (*Law of iterated expectations*) *The law of iterated expectations says that $E_t E_{t+1} y_{t+2} = E_t y_{t+2}$. This means that the expected change (in $t + 1$) of the forecast of y_{t+2} is zero.*

2.1.1 Three Asset Pricing Principles

There are three main ways (in theoretical finance) of pricing financial assets. First, *by replication*. If the payoff of asset i (x_i) is a portfolio (linear combination) of the payoffs of assets a and b

$$x_i = \alpha x_a + \beta x_b, \quad (2.1)$$

then the law of one price requires that

$$P_i = \alpha P_a + \beta P_b. \quad (2.2)$$

where x_i is the (future) payoff and P_i is today's price. All time subscripts are dropped in order to save ink.

Example 2.5 (*Binomial model*) Assume a stock will be worth either 9.5 (low state) or 11 (high state), and that there is a bond that always pays off 1. A call option (with strike price 10) on the stock will be worth either 0 or 1. You can replicate this option by holding 2/3 of the stock and $-19/3$ of the bonds, since it gives 0 in the low state and 1 in the high state.

The second way is by a *stochastic discount factor* (SDF, also called “pricing kernel”), which is a variable (m) such that

$$P_i = E mx_i, \text{ for all assets } (i), \quad (2.3)$$

The expectation should be interpreted as being conditional on the information available today. Notice that it is the same m that prices all assets.

The third way is by using the *discounted risk neutral expected payoff*

$$P_i = \frac{1}{R_f} E^* x_i, \quad (2.4)$$

where R_f is the riskfree gross rate and $E^* x_i$ is the expected value of x_i according to the riskneutral distribution. This distribution is best thought of as a theoretical construction, which cannot be directly observed from data (more details later).

Remark 2.6 (*Black-Scholes*) Consider a European call option with a strike price K . The payoff at expiration is $\max(0, P_{i,1} - K)$, where $P_{i,1}$ is the price of the underlying asset at expiration. If the return of the underlying asset follows a continuous time random walk with normally distributed shocks, then a dynamically rebalanced portfolio of the

underlying and a safe asset replicates the call option—so the call option price must equal the price of the portfolio. Alternatively, the call option price equals $E[m \max(0, P_{i,1} - K)]$ and $E^* \max(0, P_{i,1} - K) / R_f$. If $\ln m$ and $\ln P_{i,1}$ have a joint normal distribution, then all three approaches give the Black-Scholes formula.

Remark 2.7 (Bond pricing) The price of an n -period zero-coupon bond equals the cross-moment between the pricing kernel (m) and the value of the same bond next period (then an $n - 1$ -period bond)

$$P_{n,0} = E m P_{n-1,1} = \frac{1}{R_f} E^* P_{n-1,1}.$$

2.2 Stochastic Discount Factors

Dividing both sides of (2.3) by the known price gives

$$E m R_i = 1, \text{ where } R_i = x_i / P_i, \quad (2.5)$$

so R_i is the gross return. (Another way to think about this: R_i is the payoff of an asset whose price today is 1.)

For a riskfree asset (which is not correlated with the SDF) we get

$$E m E R_f = 1, \text{ so } R_f = 1 / E m. \quad (2.6)$$

Remark 2.8 (Risk and asset prices) Using (2.6) in (2.3) gives

$$P_i = \frac{E x_i}{R_f} + \text{Cov}(m, x_i).$$

This says that the price equals the expected present value of payoff + risk adjustment. Idiosyncratic risk is not compensated (priced). For instance, $E x_i = 25$, $R_f = 1.1$ and $\text{Cov}(m, x_i) = -2$

$$P_i = \frac{25}{1.1} - 2 \approx 22.7 - 2 = 20.7.$$

The investor is only willing to pay 20.7 for an asset with expected present value of 22.7: this is considered to be a risky asset.

Combining (2.5) and (2.6) gives that an excess return should satisfy

$$E m R_i^e = 0. \quad (2.7)$$

Rewrite (2.7) get the risk premium

$$\mathbb{E} R_i^e = -R_f \operatorname{Cov}(m, R_i^e) \quad (2.8)$$

Risk premium is driven by the *systematic risk* (covariance with SDF). Idiosyncratic volatility does not matter for pricing (it can be diversified away).

Example 2.9 Using $R_f = 1.1$ and $\operatorname{Cov}(m, R_i^e) = -0.1$ in (2.8) gives the risk premium

$$\mathbb{E} R_i^e = 0.11, \text{ that is } 11\%.$$

Equation (2.8) together with $\mathbb{E} m = 1/R_f$ gives

$$\operatorname{Corr}(m, R_i^e) \sigma(m) = -\mathbb{E} m \frac{\mathbb{E} R_i^e}{\sigma(R_i^e)}. \quad (2.9)$$

Since $-1 \leq \operatorname{Corr}(m, R_i^e) \leq 1$, this means (take the absolute value of both sides, notice that $\operatorname{Corr}(m, R_i^e) \leq 1$, and rearrange)

$$\frac{\sigma(m)}{\mathbb{E}(m)} \geq \frac{|\mathbb{E} R_i^e|}{\sigma(R_i^e)}. \quad (2.10)$$

The following two remarks give examples of how this inequality can be used for testing a model of m .

Remark 2.10 (*A simple test of an SDF model*) Find the highest $|\mathbb{E} R_i^e| / \sigma(R_i^e)$ from a set of assets, and check if $\sigma(m) / \mathbb{E}(m)$ (from your model) is higher. See Figure 2.1.

Remark 2.11 (*How the SDF defines a MV frontier*) Reshuffle (2.10) as

$$\mathbb{E} R_i \text{ is } \begin{cases} \leq R_f + \frac{\sigma(m)}{\mathbb{E}(m)} \sigma(R_i) & \text{if } \mathbb{E} R_i \geq R_f \\ \geq R_f - \frac{\sigma(m)}{\mathbb{E}(m)} \sigma(R_i) & \text{if } \mathbb{E} R_i \leq R_f \end{cases}$$

which gives the two lines in Figure 2.2 (draw $\mathbb{E} R_i$ as a function of $\sigma(R_i)$). Use $1 / \mathbb{E}(m) = R_f$.

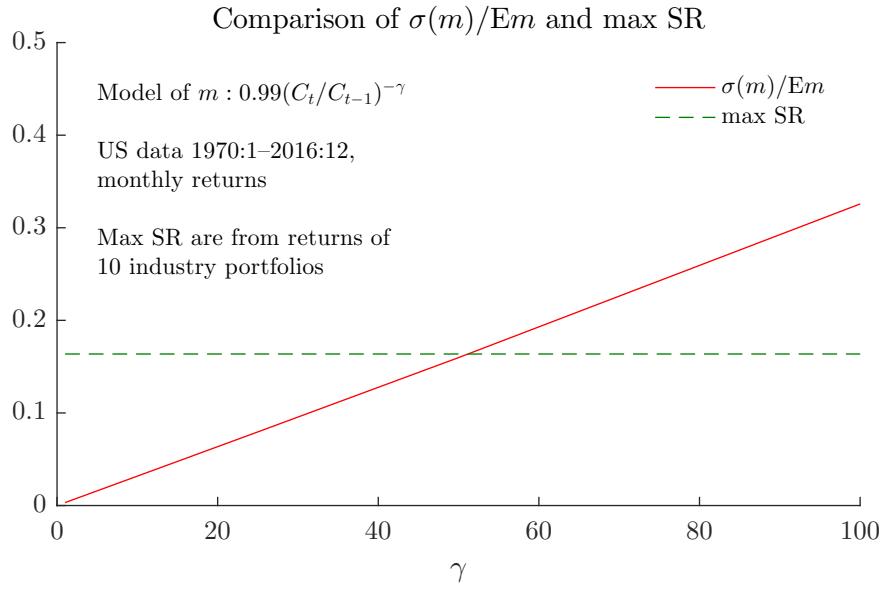


Figure 2.1: Comparison of SDF model with max SR from data on returns

2.2.1 Where Do SDFs Come from? Version 1: Optimization

Consider the two-period optimization problem

$$\max_{\{C_0, \theta_i\}} u(C_0) + \delta E u(C_1) \text{ subject to} \quad (2.11)$$

$$C_0 + \sum_{i=1}^n \theta_i P_i = W_0 \text{ and} \quad (2.12)$$

$$C_1 = y_1 + \sum_{i=1}^n \theta_i x_i. \quad (2.13)$$

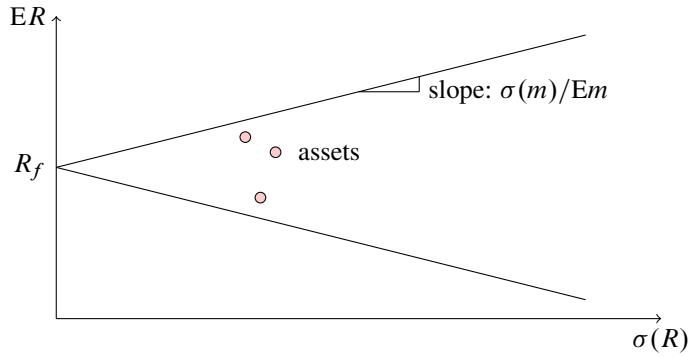


Figure 2.2: MV frontier (from SDF)

The first line defines expected utility from consuming now (period 0) and later (period 1). Notice: the subscripts on (C_0, C_1, y_1, W_0) are here used to indicate time. However, following the previous analysis we take it for granted that the price of asset i (P_i) refer to period 0 and the payoff (x_i) to period 1. The second line is the budget restriction today: consumption plus asset purchases (θ_i assets of type i , each at the price P_i) equals today's wealth. The third line defines what consumption in period 1 will be: any income y_1 plus the payoffs from the assets (x_i from asset i , of which we own θ_i).

To solve the optimization problem, substitute for C_1 and define the Lagrangian

$$\max u(C_0) + \delta E u(\underbrace{y_1 + \sum_{i=1}^n \theta_i x_i}_{C_1}) - \lambda [C_0 + \sum_{i=1}^n \theta_i P_i - W_0]. \quad (2.14)$$

The first order conditions (with respect to today's decision variables) are

$$\text{wrt } C_0: u'(C_0) = \lambda \quad (2.15)$$

$$\text{wrt } \theta_i: \delta E u'(C_1) x_i = \lambda P_i \text{ for } i = 1 \dots n. \quad (2.16)$$

Combine to get

$$\delta E u'(C_1) x_i = \lambda P_i \text{ for } i = 1 \dots n \quad (2.17)$$

$$E mx_i = P_i, \text{ where } m = \delta \frac{u'(C_1)}{u'(C_0)}. \quad (2.18)$$

Apply to a riskfree asset, $x_{i,1} = 1$ and $P_i = 1/R_f$, where R_f is the gross interest rate

$$E \delta \frac{u'(C_1)}{u'(C_0)} 1 = \frac{1}{R_f} \text{ so } R_f = \frac{1}{\delta} \frac{u'(C_0)}{E u'(C_1)}. \quad (2.19)$$

With a CRRA utility function, $u'(C) = C^{-\gamma}$, so the SDF in (2.18) is

$$m = \delta \left(\frac{C_1}{C_0} \right)^{-\gamma}. \quad (2.20)$$

Example 2.12 (Riskfree rate) With CRRA as in (2.20) with $\gamma = 3$ and $\delta = 0.95$, and $(C_0, C_1) = (1.95, 2)$ we get (assuming no uncertainty)

$$R_f = \frac{1}{0.95} \left(\frac{2}{1.95} \right)^3 \approx 1.14,$$

so the net interest rate is approximately 14%. Instead, with $C_1 = 2.1$ we get

$$R_f = \frac{1}{0.95} \left(\frac{2.1}{1.95} \right)^3 \approx 1.31,$$

so the net rate is approximately 31%. The reason is that if the future is bright, then investors want a large compensation for saving (rather, they would like to borrow).

Example 2.13 (*The SDF as a function of consumption*) With CRRA as in (2.20) with $\gamma = 3$ and $\delta = 0.95$, and $(C_0, C_1) = (1.95, 2)$ we get

$$m = 0.95 \left(\frac{2}{1.95} \right)^{-3} \approx 0.88$$

Instead, with $C_1 = 2.1$ we get

$$m = 0.95 \left(\frac{2.1}{1.95} \right)^{-3} \approx 0.76,$$

so the ratio is random (since C_1 is)—and moves inversely with C_1 .

Notice that $u'(C_1)$, and therefore m in (2.18), moves inversely with C_1 , see Figure 2.3. Approximate $m \approx a - bC_1/C_0$, where $b > 0$. Use in (2.8)

$$\begin{aligned} \mathbb{E} R_i^e &= -R_f \operatorname{Cov}(a - bC_1/C_0, R_i^e) \\ &= bR_f \operatorname{Cov}(C_1/C_0, R_i^e). \end{aligned} \quad (2.21)$$

Procyclical assets have high expected returns (as they pay off when marginal utility is low). The reason why risky assets have high risk premia is, of course, that otherwise no one would like to buy those assets. The equity premium puzzle (Mehra and Prescott (1985)) is that the covariance is too small to explain the historical average risk premium on US equity.

Example 2.14 (*From a consumption-based model to CAPM*) Suppose marginal utility is an affine function of the market excess return

$$m = a - bR_m^e, \text{ with } b > 0.$$

This would, for instance, be the case in a Lucas model where consumption equals the market return and the utility function is quadratic.

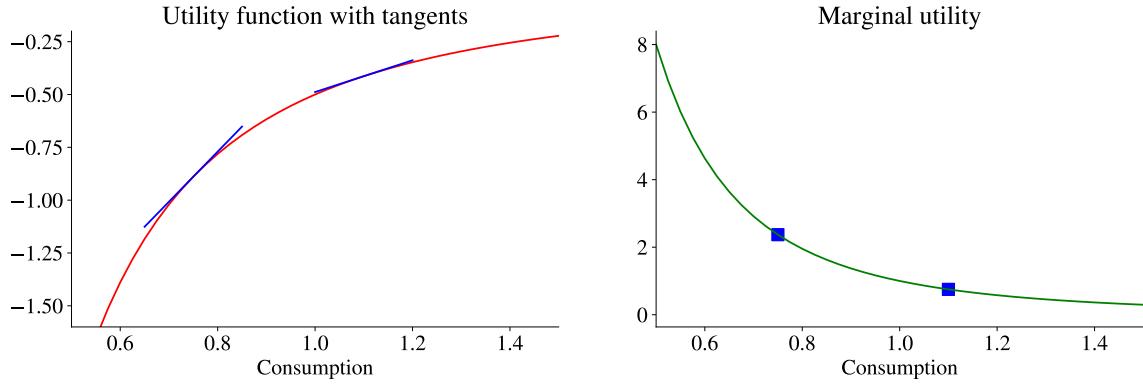


Figure 2.3: Utility function

Remark 2.15 (*Deriving CAPM from marginal utility*) If utility depends on the portfolio return R_p and is of CARA form, $u(R_p) = -\exp(-R_p k)$ where k is a measure of risk aversion, then the optimal portfolio will be on the mean-variance frontier. In equilibrium, CAPM holds.

Remark 2.16 (*The equity premium puzzle, log-normal version*) With CRRA (2.20), equation (2.5) says that the average gross return of an asset should be determined by

$$E \delta \left(\frac{C_1}{C_0} \right)^{-\gamma} R_i = 1.$$

Let $\Delta c = \ln(C_1/C_0)$ and $r_i = \ln R_i$ and write the expression as

$$E \exp(\delta - \gamma \Delta c + r_i) = 1.$$

Recall that if $x \sim N(\mu, \sigma^2)$, then $E \exp(x) = \exp(\mu + \sigma^2/2)$. Assume $-\gamma \Delta c + r_i$ is indeed normally distributed, then the previous equation can be written

$$\exp(\delta - \gamma E \Delta c + E r_i + \gamma^2 \sigma_c^2/2 + \sigma_r^2/2 - \gamma \sigma_{cr}) = 1.$$

Take logs to get

$$\delta - \gamma E \Delta c + E r_i + \gamma^2 \sigma_c^2/2 + \sigma_r^2/2 - \gamma \sigma_{cr} = 0, \text{ or}$$

$$E r_i = -\delta + \gamma E \Delta c - (\gamma^2 \sigma_c^2/2 + \sigma_r^2/2 - \gamma \sigma_{cr}).$$

Apply to a riskfree return to get

$$r_f = -\delta + \gamma E \Delta c - \gamma^2 \sigma_c^2 / 2.$$

Combine to have the risk premium

$$E r_i - r_f = \gamma \sigma_{cr} - \sigma_r^2 / 2.$$

2.2.2 Where Do SDFs Come from? Version 2: Law of One Price or No Arbitrage

We can (although it is a bit involved) prove the results in this section. In comparison with the optimization approach, they are very general: they rely on weak assumptions and tell us about whether there is an SDF. On the other hand, the results say fairly little about the economic mechanism behind the SDF.

- If markets are *complete*, then there is only one SDF. You may derive it from whatever approach, but you get the same SDF.

Instead, if markets are *incomplete* (the more realistic case), then we can still prove something about the existence of an SDF:

- The “law of one price” and incomplete markets together imply that there exists an SDF—and it can be written as linear function of available assets. However, there may be several alternative SDFs, for instance, written in terms of macro variables. They need not be the same, but they should clearly deliver the same asset prices ($E mx_i$ is the same for all possible m variables).
- No arbitrage and incomplete markets together imply that there exists at least one positive SDF. There may be other SDFs, and some of them can take on negative values (in some states of the world).

(SDFs that can have negative realizations are problematic, since they can assign negative prices to new derivatives that you consider. Notice that the optimization based approach creates SDFs that are always positive—provided marginal utilities are.)

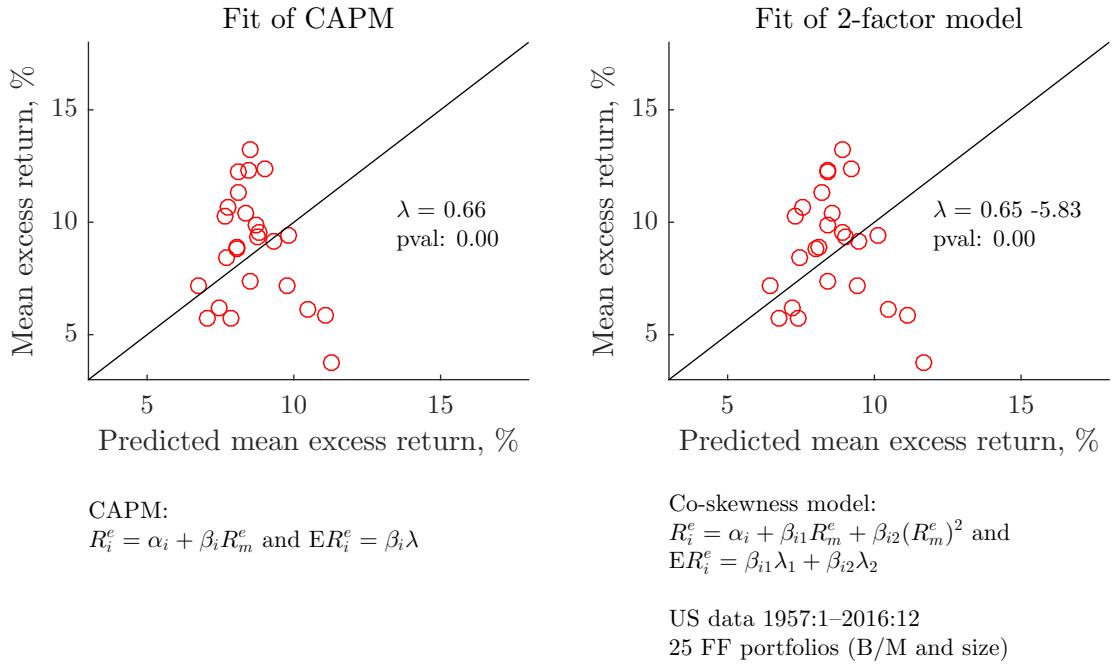


Figure 2.4: CAPM and quadratic model (co-skewness model)

2.3 Beta Pricing Models

2.3.1 Definition of a Beta Pricing Model

A beta pricing model says that the average excess return of any asset is

$$E R_i^e = \beta'_i \lambda, \quad (2.22)$$

where λ are factor risk premia and β_i are the regression coefficients from

$$R_{i,t}^e = \alpha_i + \beta'_i f_t + u_{i,t}. \quad (2.23)$$

See Figure 2.4 for an empirical example.

When there is a single factor which is an excess return, then we can apply (2.22) to f_t and notice that $\beta_i = 1$ (regressing f_t on itself gives a slope coefficient equal to one). Therefore,

$$\lambda = E f. \quad (2.24)$$

This is a way to identify the factor risk premia λ and it also holds when there are several excess return factors. However, this approach only works when the factors are excess

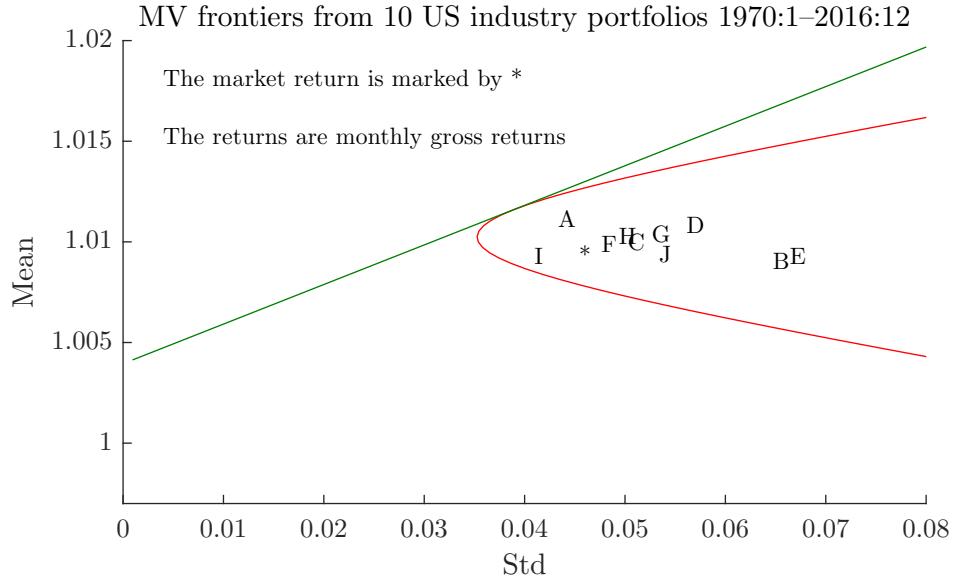


Figure 2.5: MV frontier from US industry indices

returns. If not, the factor risk premia λ must be estimated from data.

Remark 2.17 (*Some well-known factor models*) CAPM is a one-factor model, while the Fama-French model is a three-factor model. In CAPM the factor is an excess return since it is the difference between the returns of the market portfolio and another portfolio (a risk free asset). In the *Fama and French (1993)* model, the factors are also excess returns since they are also based on long-short portfolios (market minus risk free, value stock minus growth stocks, large stocks minus small stocks). *Carhart (1997)* adds a momentum factor and *Pastor and Stambaugh (2003)* a liquidity factor.

If the factors are excess returns, then (it can be shown that)

- the beta pricing model (2.22) hold if and only if there is a portfolio of the factors that is on the MV frontier (and is not equal to the minimum variance return or the riskfree return).

The *empirical implication* is that

- testing a beta pricing model is the same as testing if there is a portfolio of the excess return factors (for instance, the market return) that is MV efficient. See Figure 2.5 for an empirical example.

This can be done in several ways, but the easiest approach is run the regression (2.23) and to test if $\alpha_i = 0$. In contrast, if $\alpha \neq 0$, then (it can be shown that) the return factor is not on the MV frontier. Notice that this approach works also with several factors, provided all of them are excess returns.

Remark 2.18 (*Typical results from testing CAPM*) *It is often found that $\alpha > 0$ for small stocks and value stocks (and the opposite for growth stocks). Dynamic portfolios that bet on short-run reversal, medium-run momentum and long-run reversion also tend to have $\alpha > 0$. Finally, firms with unexpectedly high earnings growth tend to have high returns also after the surprise, but that firms with high uncertainty or high asset growth tend to underperform. The average mutual fund has a negative alpha—and it seems as if alphas of mutual funds are not particularly correlated over time.*

2.3.2 Factor Mimicking Portfolios

Pricing factors need not be returns (e.g. inflation or market volatility). In principle, that does not change much: the beta pricing model (2.22)–(2.23) would still be true, although we need another method to identify the factor risk premia than (2.24).

In practice, it is often convenient to work with “factor mimicking portfolios” instead. To construct such a portfolio, regress the factor on a constant and a vector of asset excess returns. Then use, the fitted values (minus the intercept) as the excess return of a factor mimicking portfolio.

Instead of using a regression, factor mimicking portfolios are often approximated by long-short based on asset sorts (for instance, small minus big firms).

Example 2.19 (*Fama-French*) *The Fama and French (1993) (see also Fama and French (1996)) SMB and HML portfolios can be thought of as factor mimicking portfolios—perhaps mimicking the credit cycle and the degree of optimism/pessimism on the market. See Figure 2.6. This figure shows (log) portfolio values relative to the value of a portfolio entirely invested into the riskfree asset. Each factor portfolio earns the excess return of the factor plus the riskfree rate (this makes it into a proper return which can be accumulated).*

Example 2.20 (*Carry trade factor*) *Lustig, Roussanov, and Verdelhan (2011) use a carry trade risk factor mimicking portfolio (labelled HML_{FX}) which is long currencies with high interest rates and short currencies with low interest rates. This turns out to be an important factor for explaining the cross-section of exchange rate returns. It is argued that HML_{FX} is strongly related to macro economic risk.*

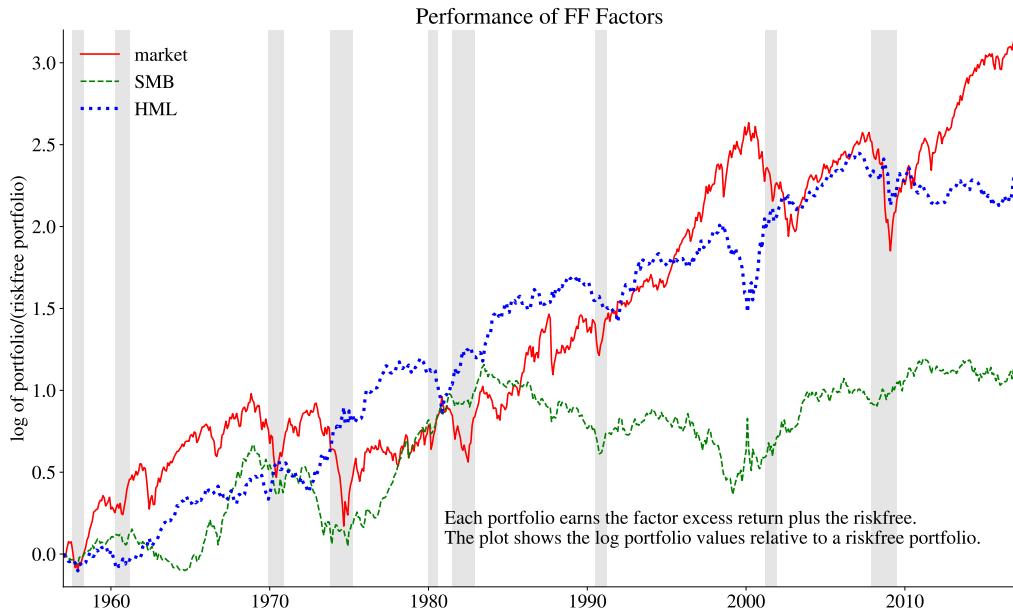


Figure 2.6: Performance of FF factors

2.3.3 SDFs versus Beta Pricing Models

We can always think of the SDF as *the* factor in a beta pricing model. To see this, rewrite (2.8) as

$$E R_i^e = \underbrace{\frac{\text{Cov}(m, R_i^e)}{\text{Var}(m)}}_{\beta_i} \underbrace{(-R_f) \text{Var}(m)}_{\lambda}. \quad (2.25)$$

However, we often want to go beyond this by providing more information about what is behind the SDF.

Example 2.21 ((2.25) with a utility based SDF) If the SDF is the ratio of marginal CRRA utilities as in (2.20), then (2.25) explains the expected average return as a function of the returns regression coefficient from $R_i^e = \gamma + \beta_i (C_1/C_0)^{-\gamma} + u$.

Remark 2.22 (Empirical performance of consumption based models) Standard consumption based asset pricing models (like the CRRA discussed above) have a problem with explaining the historical average returns on equity. It is perhaps somewhat better at explaining the cross-section of different equity returns. More sophisticated consumption based models that incorporate habits (Campbell and Cochrane (1999)) or that focus on longer horizons (Parker and Julliard (2005)) or allow for (a small) long-run movements in growth (Bansal and Yaron (2004)) do a bit better.

Given a beta pricing model against an SDF like (2.25), it is straightforward to show that

- there is always (another) beta representation—against a return on the MV frontier:

$$\mathbb{E} R_i^e = \beta_{i,mv} \lambda_{mv}, \quad (2.26)$$

where $\beta_{i,mv}$ is the regression coefficient of asset i against that return on the MV frontier.

This result says that MV frontiers are crucial for all asset pricing. However, it remains to be tested whether your favourite factor is really on the frontier. For instance, if the market return happens to be on the MV frontier, then CAPM holds—otherwise not.

Proof. (Proof of (2.25) \Rightarrow (2.26)*) Recall from (2.9)–(2.10) that an asset that is perfectly correlated (positively or negatively) with the SDF has the highest possible |Sharpe ratio|, which means that it is on the MV frontier. The perfect correlation also means that the return of this asset (R_{mv}) must be such that $m = \gamma + \delta R_{mv}$, where γ and δ are two constants. Using this in (2.25) gives the result.

$$\begin{aligned} \mathbb{E} R_i^e &= \frac{\text{Cov}(\gamma + \delta R_{mv}, R_i^e)}{\text{Var}(\gamma + \delta R_{mv})} (-R_f) \text{Var}(\gamma + \delta R_{mv}) \\ &= \underbrace{\frac{\text{Cov}(R_{mv}, R_i^e)}{\text{Var}(R_{mv})}}_{\beta_{i,mv}} \underbrace{(-R_f) \delta}_{\lambda_{mv}} \text{Var}(R_{mv}). \end{aligned}$$

■

It can also be shown that an SDF that is a linear function of some factors (f) is the same thing as having linear factor model. To be precise,

- there is a beta pricing model with the factors f if and only if there is an SDF that is linear (affine) in those factors, $m = a + b'f$.

This implies that if we know (a, b) in the model

$$m = a + b'(f - \mathbb{E} f), \quad 0 = \mathbb{E} m R_i^e, \quad (2.27)$$

then we can find an λ such that the beta pricing model (2.22) holds. Conversely, given λ in (2.22) we can find b such that (2.27) holds. (Notice: β_i can be estimated and a is only important if we use some returns, not just excess returns. In that case, $a = 1/R_f$.) See Figure 2.7 for a numerical example and Figure 2.8 for an empirical illustration.

Example 2.23 (*From consumption to other factors*) If the ratio of marginal utility (consumption) in Example 2.21 depends on a vector of factors in a linear way, then we have

an SDF like in (2.27). For instance, the macro-economic equilibrium might imply that marginal utility is a linear function of some key macroeconomic variables like output, interest rates and inflation

$$m = \gamma \text{Output} + \delta \text{Interest rate} + \kappa \text{Inflation},$$

where (γ, δ, κ) are constants. The factor model for the asset prices then include the same macro factors.

The empirical implication of (2.27) is that

- estimating/testing a linear SDF model is the same as estimating/testing an old-fashioned linear factor model.

The choice between a linear SDF model or a beta pricing model is therefore based on what is more convenient (or already established in the literature). For most asset classes this means a beta pricing model, although studies of bonds and derivatives often work with SDFs. What is more important is the assumption of linearity—and the choice of the factors.

Proof. (Proof of (2.27)*) Combine (2.27) and (2.22) to get

$$\lambda = -\frac{1}{a} \text{Var}(f)b \text{ or } b = -a \text{Var}(f)^{-1}\lambda,$$

where $\text{Var}(f)$ is the variance-covariance matrix of f . By definition, the betas are multiple regression coefficients, so they are

$$\beta_i = \text{Var}(f)^{-1} \text{Cov}(f, R_i^e).$$

■

Example 2.24 (CAPM⇒SDF) Suppose we know that (2.22) is $E R_i^e = \beta_i \lambda$ with $\lambda = 0.08$ and where β_i is the regression coefficient in $R_i^e = \alpha_i + \beta_i R_m^e + u_i$, where R_m^e is the market excess return ($f = R_m^e$). Suppose $\text{Var}(R_m^e) = 0.16^2$. From the proof of (2.27) and using $E m = a = 1$ (which is not important since we are dealing with excess returns), we have

$$\begin{aligned} \lambda &= -b \text{Var}(R_m^e), \text{ or} \\ 0.08 &= -b 0.16^2 \Rightarrow b = -3.125. \end{aligned}$$

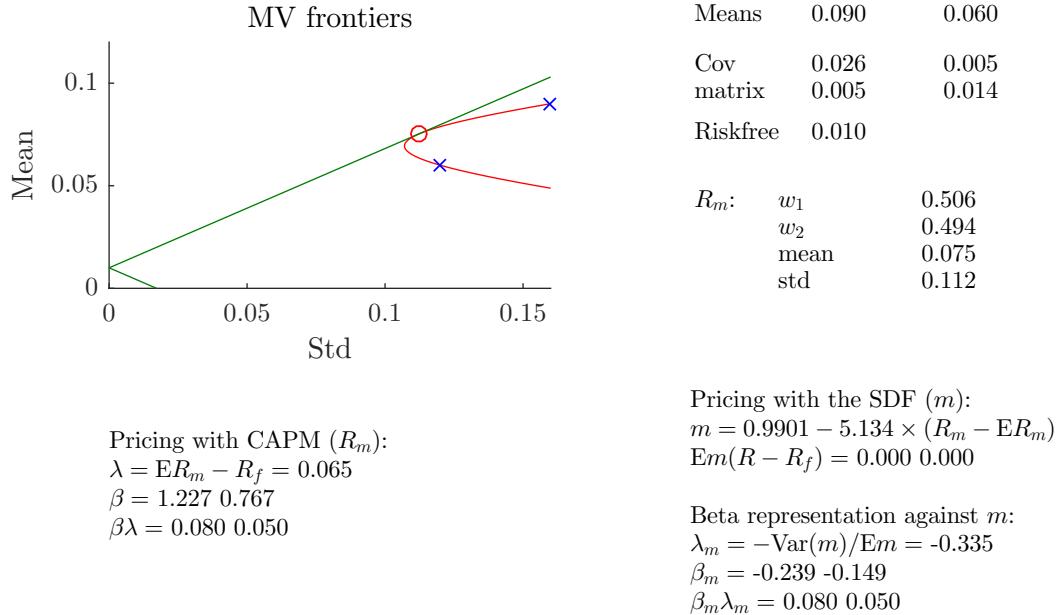


Figure 2.7: Calculations of SDF, beta representation against both SDF and R_m

Use in (2.27)

$$m = 1 - 3.125 \times (R_m^e - E R_m^e)$$

Notice the sign: m proxies marginal utility. In terms of risk premia, recall (2.8) and combine with the equation for m

$$E R_i^e = R_f 3.125 \times \text{Cov}(R_m^e, R_i^e),$$

which shows that a procyclical asset has a positive risk premium.

Example 2.25 ($SDF \Rightarrow CAPM$) Suppose we know that

$$m = 1 - 3.125 \times (R_m^e - E R_m^e)$$

and that $\text{Var}(f) = \text{Var}(R_m^e) = 0.16^2$. The proof of 2.27 gives

$$\lambda = -b' \text{Var}(f) = -0.16^2 \times (-3.125) = 0.08,$$

so the beta (CAPM) representation (2.22) is $E R^e = \beta_i \times 0.08$.

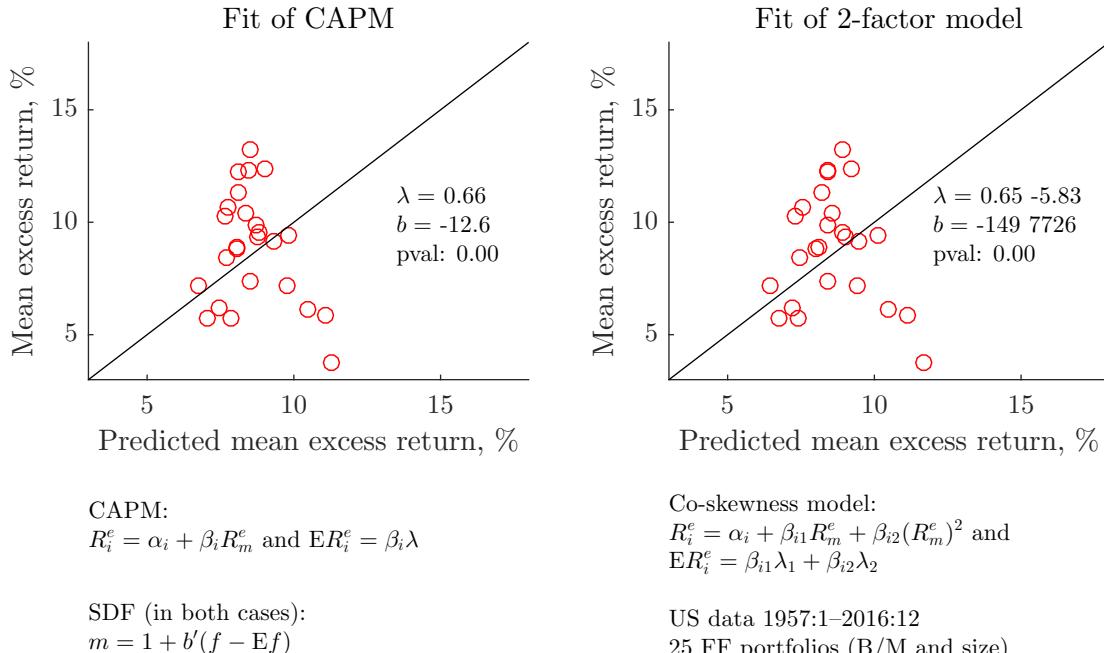


Figure 2.8: CAPM and quadratic model (co-skewness model)

2.3.4 Conditional Asset Pricing

(In this section I use time subscripts, since it is needed to clarify the concepts)

Most asset pricing theories are conditional

$$1 = E_t m_{t+1} R_{i,t+1}, \quad (2.28)$$

where E_t are the expectations at time of the portfolio formation (here called period t) and $R_{i,t+1}$ is the gross return of asset i in period $t + 1$. However, we typically want an expression in terms of an unconditional expectation—since that can be approximated by a sample average of available data.

Use *iterated expectations* to get

$$1 = E m_{t+1} R_{i,t+1}, \quad (2.29)$$

where the E denotes an unconditional expectation. This is correct, but explores only a very limited set of the model properties. For instance, the distribution of R_i might be time-varying (...predictability). The conditional asset pricing equation (2.28) says that our SDF m must be able to explain this time-variation (for instance, in terms of the riskiness).

However, (2.29) disregards this time-variation and only cares about the average across time. Conditional asset pricing models typically try to work with a simplified version of (2.28) where some information variable captures the changes in the distribution from period to period.

See Figure 2.9 for an example where the return distribution is different in different time periods (driven by a state variable which can take only two values, so we have states A and B). To keep the figures simple, there is only one risky asset so the (efficient part of the) mean-variance is just a straight line. Also see Figure 2.10 for an empirical illustration, where the factor loadings (betas) are modelled in terms of a state variable.

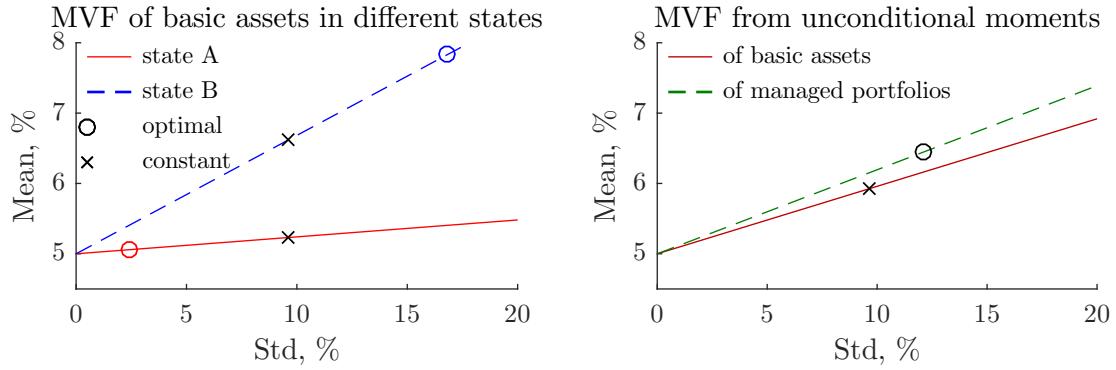


Figure 2.9: MV frontiers in two different states (left figure) and on average (right figure)

2.4 Risk Neutral Distributions

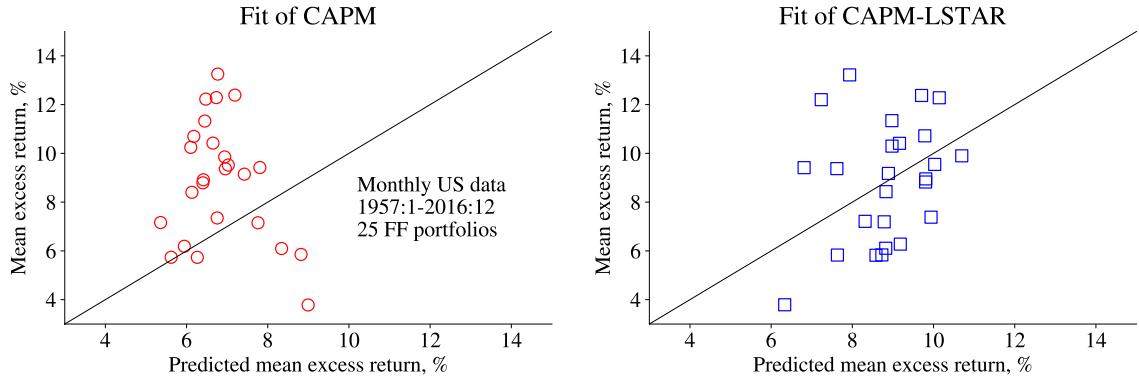
To simplify the analysis, assume that there are k different “states of the world” and that state j has the probability π_j . The asset pricing model (2.3) can then be written (for asset i)

$$P_i = \text{E} mx_i = \sum_{j=1}^k \pi_j m_j x_{ij}, \quad (2.30)$$

where x_{ij} is the payoff of asset i in state j (next period).

Define the risk neutral probabilities as

$$\pi_j^* = R_f m_j \pi_j, \quad (2.31)$$



$$R_i^e = \alpha + \beta R_m^e + \epsilon$$

$$R_i^e = \alpha + \beta_L [1 - G(z)] R_m^e + \beta_H G(z) R_m^e + \epsilon$$

z: lagged momentum return

Figure 2.10: Test of 1 and 2-factor models, 25 FF portfolios

where R_f is the gross riskfree rate. Then rewrite (2.30) as

$$\begin{aligned} P_i &= \sum_{j=1}^k \underbrace{\pi_j m_j}_{\pi^*(j)/R_f} x_{ij} \\ &= \sum_{j=1}^k \pi_j^* x_{ij} / R_f \\ &= E^* x_i / R_f, \end{aligned} \quad (2.32)$$

which says that the asset price equals the discounted “risk neutral expected value.”

Clearly, (2.32) can be rearranged as

$$\frac{E^* x_i}{P_i} = R_f, \quad (2.33)$$

which says that the risk neutral expected (gross) return on any asset equals the (gross) riskfree rate.

Risk neutral distributions are easy to work with, especially since the asset price can be calculated as the expected value of just one variable (the payoff x_i), instead of an expectation of a product ($m x_i$). For instance, once we have the risk neutral distribution of asset i for a given investment horizon, we can price any European derivative (whose expiration is at the investment horizon) on that asset. For instance, if we have the risk neutral distribution for the bond price 3 months from now, we can price also a European put option on that bond.

Recall that a forward contract has a zero price and a payoff $x_{i,1} = P_{i,1} - F_i$, where $P_{i,1}$ is the price of the underlying asset i next period and F_i the contracted forward price. Apply (2.32) to get

$$0 = \mathbb{E}^* (P_{i,1} - F_i) / R_f, \text{ so } F_i = \mathbb{E}^* P_{i,1}. \quad (2.34)$$

This shows that the mean of the risk neutral distribution (of asset i) equals the forward price.

Remark 2.26 (*Binomial model*) Consider the process for the underlying asset i

$$P_{i,1} = \begin{cases} P_{i,0}u & \text{with probability } \pi \\ P_{i,0}d & \text{with probability } 1 - \pi, \end{cases}$$

where $P_{i,0}$ is today's price and $P_{i,1}$ is the price next period. For simplicity, we assume no dividends. We know from basic option pricing that a derivative that is worth C_u in the up state and C_d in the down state must have the current price C_0

$$C_0 = \frac{1}{R_f} [\pi^* C_u + (1 - \pi^*) C_d] \text{ with } \pi^* = \frac{R_f - d}{u - d},$$

where π^* is the risk neutral probability. From (2.31) we know that $\pi_j^*/R_f = m_j \pi_j$, so

$$\begin{aligned} \pi^*/R_f &= m_u \pi \\ (1 - \pi^*)/R_f &= m_d (1 - \pi). \end{aligned}$$

We can therefore write the price of the derivative also as

$$C_0 = m_u \pi C_u + m_d (1 - \pi) C_d,$$

which is just $C_0 = \mathbb{E}(m \times \text{payoff of derivative})$.

Example 2.27 (*Binomial model*) If $u = 1.1$, $d = 0.95$, $\pi = 2/3$, $R_f = 1$, $C_u = 1$ and $C_d = 0$, then we get

$$C_0 = \pi^* 1 \text{ with } \pi^* = \frac{1 - 0.95}{1.1 - 0.95} = \frac{1}{3}.$$

It follows that $m_u = 1/2$ and $m_d = 2$. Notice that $\mathbb{E} m = 2/3 \times 1/2 + 1/3 \times 2 = 1$ so the gross riskfree rate is indeed 1. Also, notice that using m to price the derivative gives $C_0 = 2/3 \times 1/2 \times 1 = 1/3$.

2.4.1 Special Case: Lognormal Distribution (the Log Asset Price)

Suppose log asset price is a univariate normal (we call this is “physical distribution”)

$$\ln P_{i,1} = p_{i,1} \sim N(\mu_p, \sigma_{pp}), \quad (2.35)$$

and also that the distribution of the log SDF is also normal

$$\ln m_1 \sim N(\mu_m, \sigma_{mm}). \quad (2.36)$$

Direct calculations then give that today’s gross interest rate is (recall if $x \sim N(\mu, \sigma^2)$, then $E \exp(x) = \exp(\mu + \sigma^2/2)$)

$$\frac{1}{R_f} = E m_1 = \exp(\mu_m + \sigma_{mm}/2). \quad (2.37)$$

Similarly, today’s price of the underlying asset is (assuming no dividends, so the future price is the payoff)

$$\begin{aligned} P_{i,0} &= E m_1 P_{i,1} \\ &= E \exp(\ln m_1 + p_{i,1}) \\ &= \exp(\mu_m + \mu_p + \sigma_{mm}/2 + \sigma_{pp}/2 + \sigma_{mp}) \\ &= \underbrace{\exp(\mu_m + \sigma_{mm}/2)}_{1/R_f} \exp(\mu_p + \sigma_{pp}/2 + \sigma_{mp}) \\ &= \frac{1}{R_f} \exp(\mu_p + \sigma_{pp}/2 + \sigma_{mp}). \end{aligned} \quad (2.38)$$

Suppose the risk neutral distribution of the future log asset price is also normal

$$p_{i,1} \sim^* N(\mu_p^*, \sigma_{pp}), \text{ with } \mu_p^* = \mu_p + \sigma_{mp}. \quad (2.39)$$

This distribution has the same variance as the physical distribution of $p_{i,1}$ in (2.35), but a different mean. This simple result is due to the assumption of lognormally distributed variables. See Figure 2.11 for an example.

To illustrate that this works, notice that

$$\begin{aligned} P_{i,0} &= \frac{1}{R_f} E^* P_{i,1} \\ &= \frac{1}{R_f} E^* \exp(p_{i,1}) \\ &= \frac{1}{R_f} \exp(\mu_p + \sigma_{mp} + \sigma_{pp}/2), \end{aligned} \tag{2.40}$$

which is the same as the SDF approach (2.38) gives.

To apply the risk neutral distribution, we could, for instance, price a European put option with strike price K as

$$\begin{aligned} \text{Put}_0 &= \frac{1}{R_f} E^* \max(0, K - P_{i,1}) \\ &= \frac{1}{R_f} \int_{-\infty}^K [K - \exp(p)] \phi^*(p; \mu_p^*, \sigma_{pp}) dp, \end{aligned} \tag{2.41}$$

where $\phi^*(p; \mu_p^*, \sigma_{pp})$ is notation for the pdf for a $N(\mu_p^*, \sigma_{pp})$ distribution evaluated at p . The solution to this integral is the Black-Scholes formula for pricing a put option.

Example 2.28 (*Asset price under lognormality.*) If $\mu_m = -0.04$, $\sigma_{mm} = 0.04$, $\mu_p = 0.03$, $\sigma_{pp} = 0.01$ and $\sigma_{mp} = -0.015$, then (2.37)–(2.38) give

$$\begin{aligned} \frac{1}{R_f} &= \exp(-0.04 + 0.04/2) \approx 0.98 \text{ so } \ln R_f = 0.02 \\ P_{i,0} &= \exp(-0.04 + 0.04/2) \exp(0.03 - 0.015 + 0.01/2) = 1. \end{aligned}$$

Since $P_{i,0} = 1$, the payoff is actually a gross return.

Example 2.29 (*Risk neutral lognormal distribution*) In Example 2.28, the physical distribution of the log payoff ($p_{i,1}$) is $N(0.03, 0.01)$, while the risk neutral distribution (2.39) is $N(0.03 - 0.015, 0.01)$, that is, $N(0.015, 0.01)$. See Figure 2.11.

To interpret the risk neutral pdf in (2.39), notice that an asset with a negative covariance with the pricing kernel tends to pay off in the “wrong” states (for instance, in booms), so it is considered a risky asset and will have a low price. For a risk neutral investor to make an equally low valuation, he must be more pessimistic about the future payoff: the distribution is shifted down (to the left).

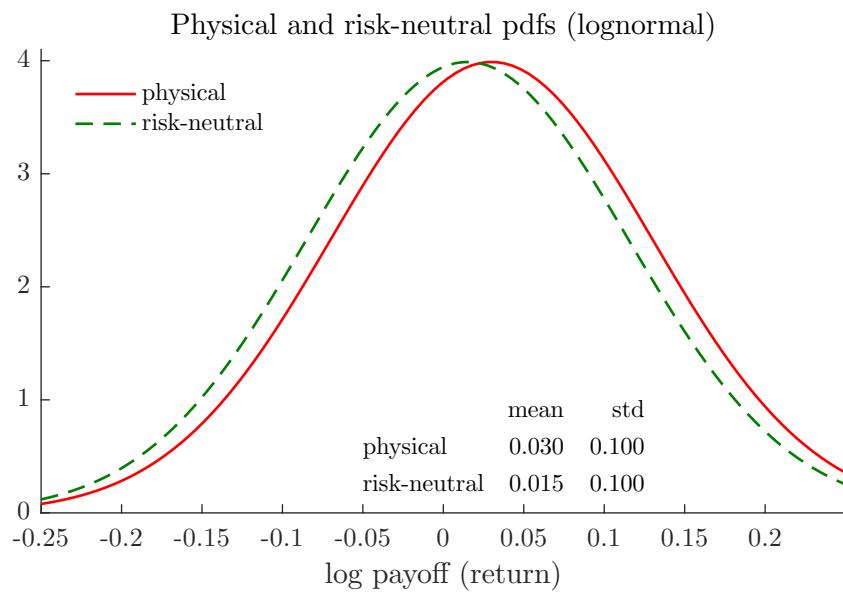


Figure 2.11: Physical and risk neutral distributions

Figures 2.12–2.13 illustrate a case where the distribution is more complicated—and the transformation from the physical to the risk neutral distribution involves much more than just a horizontal shift.

Joint distribution: $\ln m \sim N, R^e \sim \text{mix}N$

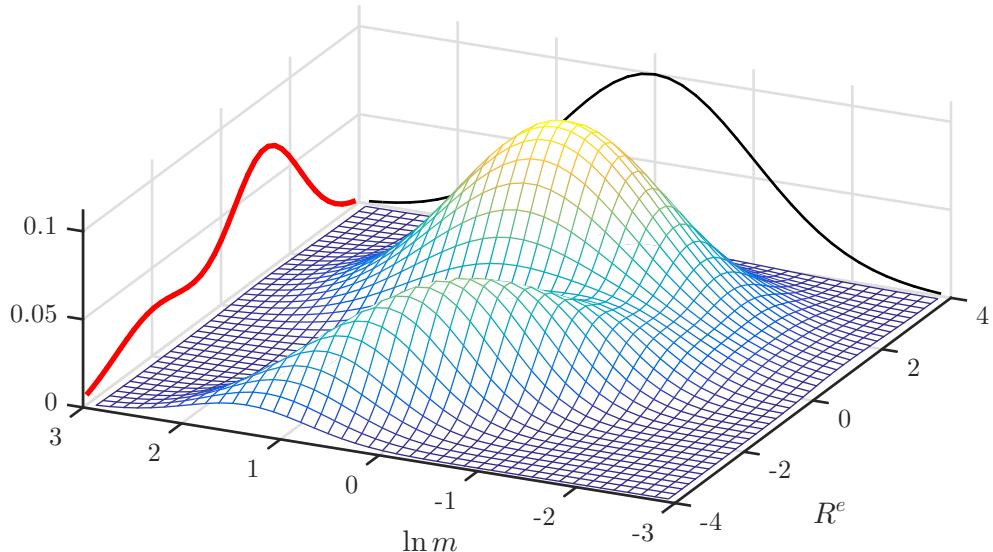


Figure 2.12: Physical and risk neutral distributions

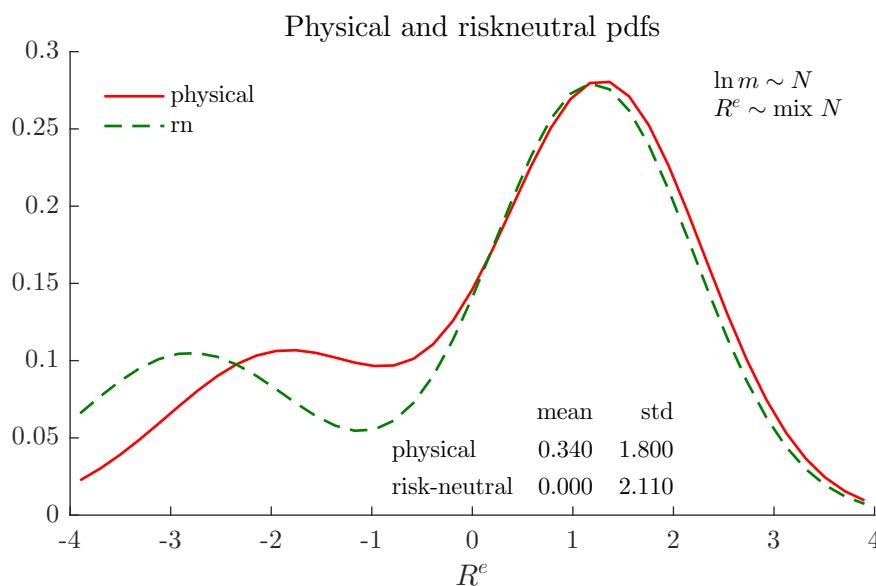


Figure 2.13: Physical and risk neutral distributions

Chapter 3

Simulating the Finite Sample Properties

Reference: Greene (2000) 5.3 and Horowitz (2001)

Additional references: Cochrane (2001) 15.2; Davidson and MacKinnon (1993) 21; Davidson and Hinkley (1997); Efron and Tibshirani (1993) (bootstrapping, chap 9 in particular); and Berkowitz and Kilian (2000) (bootstrapping in time series models)

3.1 Introduction

We know the small sample properties of regression coefficients in linear models with fixed regressors and iid normal error terms. When these conditions are not satisfied, then we must rely on asymptotic results or apply Monte Carlo/bootstrap simulations to approximate the small sample properties. For instance, if the regression residuals have autocorrelation and/or heteroskedasticity, then we may either use a consistent estimator of the covariance matrix (Newey-West, White, etc) and apply the usual test by comparing with the asymptotically correct $N(0, 1)$ or χ_q^2 distributions. Alternatively, we can compare the test statistic (based on either the classical covariance matrix or a consistent one) with a simulated distribution. The advantage of the simulations is that they might provide better approximations of the small sample properties than the asymptotic distribution does.

The results from the simulations can be used to study, for instance, (a) the distribution of a point estimate (to create confidence bands or a standard deviation) or (b) the distribution of a test statistic (to generate appropriate critical values).

How these simulations should be implemented depends crucially on the properties of the model and data: if the residuals are autocorrelated, heteroskedastic, or perhaps correlated across regressions equations. These notes summarize a few typical cases.

It should be noticed that the need for using Monte Carlos or bootstraps varies across applications and data sets. For a case where it is not really needed, see Table 3.1, and

for a case where it matters, compare the traditional and bootstrapped t-stats in Tables 11.1–11.2.

	α	t (LS)	t (NW)	t (boot)
A (NoDur)	3.56	2.81	2.62	2.28
B (Durlb)	-1.31	-0.67	-0.68	-0.66
C (Manuf)	0.52	0.57	0.55	0.51
D (Enrgy)	3.05	1.41	1.40	1.44
E (HiTec)	-1.73	-1.02	-1.03	-0.94
F (Telcm)	1.93	1.23	1.19	1.07
G (Shops)	1.24	0.91	0.87	0.86
H (Hlth)	2.28	1.39	1.42	1.36
I (Utils)	3.06	1.78	1.74	1.82
J (Other)	-0.51	-0.51	-0.50	-0.43

Table 3.1: Estimates of CAPM on US industry portfolios 1970:1-2016:12. NW uses 1 lag. The bootstrap samples (y_t, x_t) pairs, in blocks of 10 observations and has 3000 simulations.

	2y	3y	4y	5y
factor	1.00 (6.70)	1.87 (6.82)	2.68 (6.99)	3.46 (7.15)
constant	-0.00 (-0.00)	-0.00 (-0.34)	-0.00 (-0.68)	-0.00 (-1.02)
R2	0.14	0.14	0.15	0.16
obs	624.00	624.00	624.00	624.00

Table 3.2: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. Numbers in parentheses are t-stats. U.S. data for 1964:1-2016:12.

3.2 Monte Carlo Simulations

3.2.1 Monte Carlo Simulations in the Simplest Case

Monte Carlo simulations is essentially a way to generate many artificial (small) samples from a parameterised model and then estimate the statistic (for instance, a slope coefficient) on each of those samples. The distribution (across the artificial samples) of the statistic is then used as an approximation of the small sample distribution of the estimator.

	2y	3y	4y	5y
factor	1.00 (4.05)	1.87 (4.19)	2.68 (4.33)	3.46 (4.48)
constant	-0.00 (-0.00)	-0.00 (-0.16)	-0.00 (-0.32)	-0.00 (-0.48)
R2	0.14	0.14	0.15	0.16
obs	624.00	624.00	624.00	624.00

Table 3.3: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. U.S. data for 1964:1-2016:12. Numbers in parentheses are t-stats. Bootstrapped standard errors, with blocks of 10 observations.

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

$$y_t = x_t' \beta + u_t, \quad (3.1)$$

where u_t is iid $N(0, \sigma^2)$ and x_t is stochastic but independent of $u_{t \pm s}$ for all s . (This means that x_t cannot include lags of y_t .)

Suppose we want to find the small sample distribution of a function of the estimate, $g(\hat{\beta})$. To do a Monte Carlo experiment, we need information on (i) the coefficients β ; (ii) the variance of u_t , σ^2 ; (iii) and a process for x_t .

The process for x_t is typically estimated from the data on x_t (for instance, a VAR system $x_t = A_1 x_{t-1} + A_2 x_{t-2} + e_t$). Alternatively, we could simply use the actual sample of x_t and repeat it.

The values of β and σ^2 are often a mix of estimation results and theory. In some case, we simply take the point estimates. In other cases, we adjust the point estimates so that $g(\beta) = 0$ holds, that is, so you *simulate the model under the null hypothesis* in order to study the size of 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 this discussion a bit more concrete, suppose you want to use these simulations to get a 5% critical value for testing the null hypothesis $g(\beta) = 0$. The Monte Carlo experiment follows these steps.

1. Construct an artificial sample of the regressors (see above), \tilde{x}_t for $t = 1, \dots, T$.

Draw random numbers \tilde{u}_t for $t = 1, \dots, T$ from a prespecified distribution (for instance, $N(0, \sigma^2)$) and use those together with the artificial sample of \tilde{x}_t to calculate an artificial sample \tilde{y}_t for $t = 1, \dots, T$ from

$$\tilde{y}_t = \tilde{x}'_t \beta + \tilde{u}_t, \quad (3.2)$$

by using the prespecified values of the coefficients β (perhaps your point estimates).

2. Calculate an estimate $\tilde{\beta}$ and record it along with the value of $g(\tilde{\beta})$ and perhaps also the test statistic of the hypothesis that $g(\beta) = 0$.
3. Repeat the previous steps N (3000, say) times. The more times you repeat, the better is the approximation of the small sample distribution.
4. Sort your simulated $\tilde{\beta}$, $g(\tilde{\beta})$, and the test statistic in ascending order. For a one-sided test (for instance, a chi-square test), take the $(0.95N)$ th observations in this sorted vector as your 5% critical value. 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 could also record how many times the 5% critical values from the asymptotic distribution would reject a true null hypothesis.
5. You may also want to plot a histogram of $\tilde{\beta}$, $g(\tilde{\beta})$, and the test statistic to see if there is a small sample bias, and how the distribution looks like. Is it close to normal? How wide is it? You could also estimate the variance-covariance matrix of $\tilde{\beta}$ by treating each estimate (from each simulation) as an observation—and then estimate the covariance matrix across these observations.

We use the same basic procedure when y_t is a vector, except that we have to consider correlations across the elements of the vector of residuals u_t . For instance, we could generate the vector \tilde{u}_t from a $N(\mathbf{0}, \Sigma)$ distribution—where Σ is the variance-covariance matrix of u_t .

Remark 3.1 (*Generating $N(\mu, \Sigma)$ random numbers*) Suppose you want to draw an $n \times 1$ vector ε_t of $N(\mu, \Sigma)$ variables. Use the Cholesky decomposition of Σ to calculate the lower triangular P such that $\Sigma = PP'$. Draw u_t from an $N(0, I_n)$ distribution, and define $\varepsilon_t = \mu + Pu_t$. Note that $\text{Cov}(\varepsilon_t) = E P u_t u'_t P' = P I P' = \Sigma$.

It is straightforward to sample the errors from other distributions than the normal, for instance, a student-t distribution. Equipped with uniformly distributed random numbers, you can always (numerically) invert the cumulative distribution function (cdf) of any

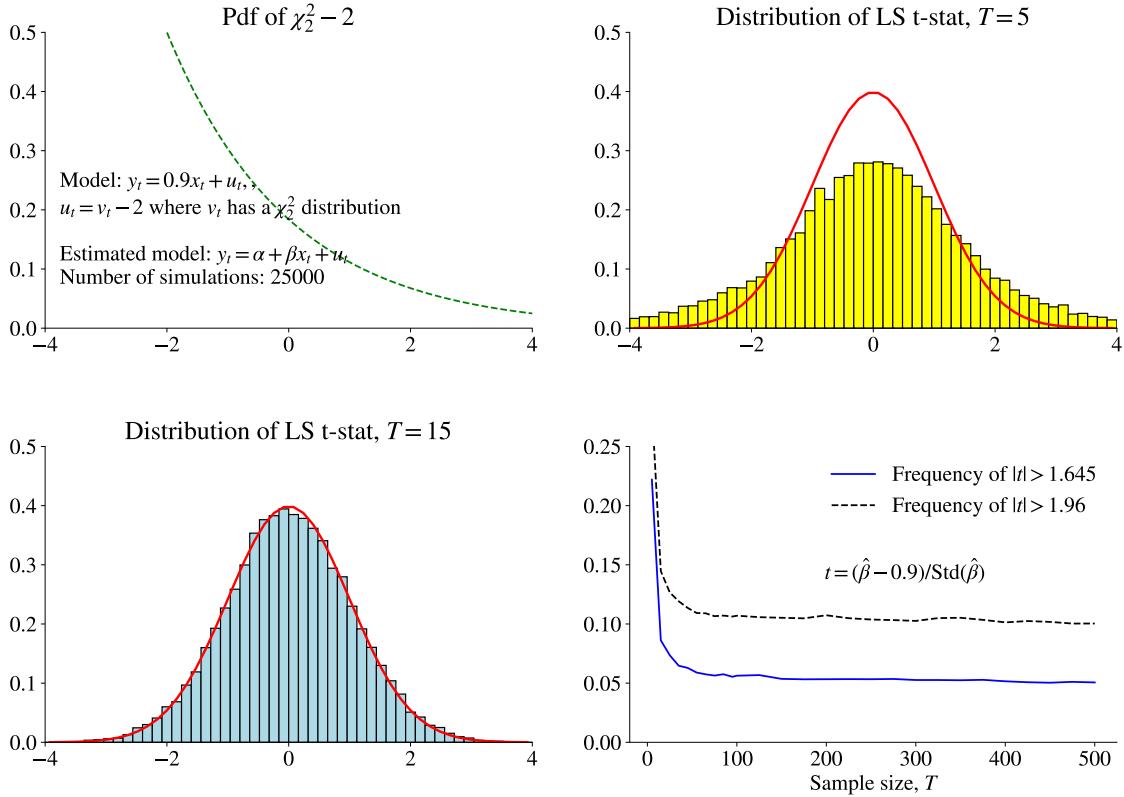


Figure 3.1: Results from a Monte Carlo experiment with fat-tailed errors

distribution to generate random variables from any distribution by using the probability transformation method. See *Figure 3.1* for an example.

Remark 3.2 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 .

3.2.2 Monte Carlo Simulations when x_t Includes Lags of y_t

When x_t contains lags of y_t , then we must set up the simulations so that temporal link is preserved in every artificial sample which we create. For instance, suppose x_t includes y_{t-1} and another vector z_t of variables which are independent of $u_{t \pm s}$ for all s

$$\begin{aligned} y_t &= x_t' \beta + u_t \\ &= \gamma y_{t-1} + \phi' z_t + u_t. \end{aligned} \tag{3.3}$$

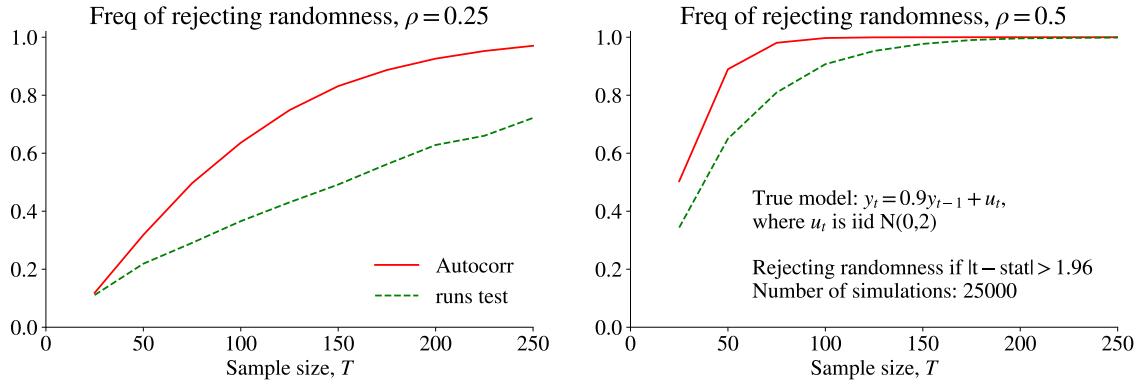


Figure 3.2: Results from a Monte Carlo experiment on two methods of testing for randomness.

We can then generate an artificial sample as follows. First, create a sample \tilde{z}_t for $t = 1, \dots, T$ by some time series model (for instance, a VAR) or by taking the observed sample itself. Second, observation t of $(\tilde{x}_t, \tilde{y}_t)$ is generated recursively as

$$\tilde{y}_t = \tilde{x}'_t \beta + \tilde{u}_t \text{ for } t = 1, \dots, T \text{ where} \quad (3.4)$$

$$\tilde{x}_t = \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{z}_t \end{bmatrix}. \quad (3.5)$$

Notice that this makes sure that \tilde{y}_{t-1} is the lagged value of \tilde{y}_t (from the same artificial sample). We clearly need the initial value \tilde{y}_0 (for instance, a randomly picked number from the sample of y_t) to start up the artificial sample—and then the rest of the sample ($t = 1, 2, \dots$) is calculated recursively. To reduce the importance of the initial value, you may choose to generate $100 + T$ values and then discard the first 100 observations. See Figures 3.2–3.3 for examples.

Remark 3.3 (*Monte Carlo for a VAR system*) For a VAR(2) model (where there is no z_t)

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

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 $t = 1, \dots, T$ and construct an artificial sample recursively (first $t = 1$, then $t = 2$ and so forth) as

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

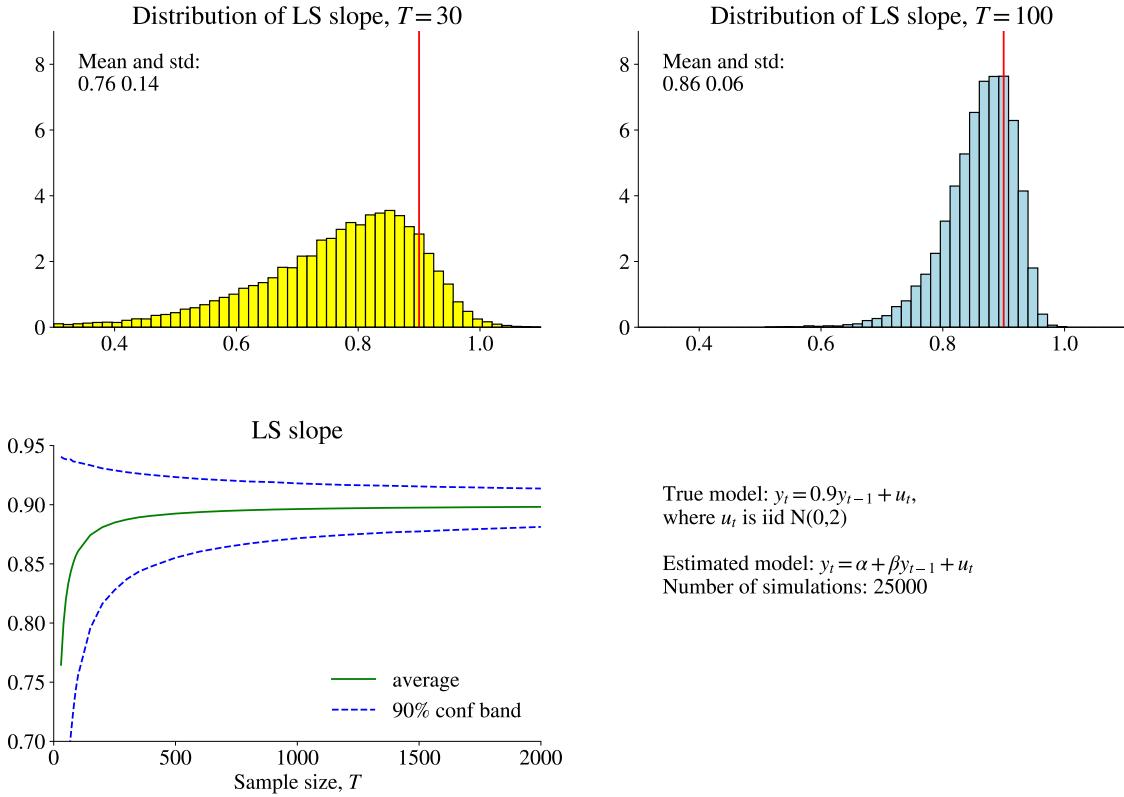


Figure 3.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

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

3.2.3 Monte Carlo Simulations with non-iid Errors

It is more difficult to handle non-iid errors, like those with autocorrelation and heteroskedasticity. We then need to model the error process and generate the errors from that model.

When the errors are *autocorrelated*, then we could estimate the error process from the fitted errors and then generate artificial samples of errors (here by an AR(2))

$$\tilde{u}_t = a_1 \tilde{u}_{t-1} + a_2 \tilde{u}_{t-2} + \tilde{\varepsilon}_t, \quad (3.6)$$

where $\tilde{\varepsilon}_t$ are iid.

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

model

$$u_t \sim N(0, \sigma_t^2), \text{ where } \sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2. \quad (3.7)$$

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

$$\varepsilon_t^2 = c' w_t + \eta_t, \quad (3.8)$$

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

3.3 Bootstrapping

3.3.1 Bootstrapping in the Simplest Case

Bootstrapping is another way to do simulations, where we construct artificial samples by sampling from the actual data (this is sometimes called a non-parametric bootstrap, whereas a parametric bootstrap is basically a Monte Carlo simulation). 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). This means 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, among other things, that x_t cannot include lags of y_t .) We here consider bootstrapping the linear model (3.1), for which we have point estimates (perhaps from LS) and fitted residuals. The procedure is then similar to the Monte Carlo approach, except that the artificial sample is generated somewhat differently. In particular, Step 1 in the Monte Carlo simulation is replaced by the following:

1. Construct an artificial sample \tilde{y}_t for $t = 1, \dots, T$ by

$$\tilde{y}_t = x_t' \hat{\beta} + \tilde{u}_t, \quad (3.9)$$

where \tilde{u}_t is drawn with replacement from the fitted residuals ($\tilde{u}_t = \hat{u}_s$ where s is the random draw) and where $\hat{\beta}$ is the point estimate from the original sample. Clearly, x_t is just the original data.

Example 3.4 With $T = 3$, the artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x'_1 \hat{\beta} + \hat{u}_2, x_1) \\ (x'_2 \hat{\beta} + \hat{u}_1, x_2) \\ (x'_3 \hat{\beta} + \hat{u}_2, x_3) \end{bmatrix}.$$

The approach in (3.9) works also when y_t is a vector of dependent variables. In this case we draw the whole vector \tilde{u}_t together to retain the cross-sectional correlation of the residuals.

The theoretical motivation for why bootstraps work is that the distribution of the fitted residuals converge to the true distribution as the sample size increases. In this sense, the bootstrap relies on asymptotic results, just like most traditional tests rely on a central limit theorem. The key point, however, is that the bootstrap often has smaller distortions (for instance, to the rejection frequency) than traditional tests have.

Remark 3.5 (*Bootstrapped confidence bands*) Using the simulated 0.025th and 0.975th quantiles of the bootstrapped $\tilde{\beta}$ values is a way of creating a 95% confidence band, sometimes called Efron's "bootstrap percentile method". The "bootstrap percentile t-method" (also suggested by Efron) is often considered to be an improvement. To implement it, first define $\tilde{t} = (\tilde{\beta} - \hat{\beta}) / \text{Std}(\tilde{\beta})$, where $\text{Std}(\tilde{\beta})$ is the standard deviation across the bootstrap estimates. (Sometimes the centering is done by subtracting the average of $\tilde{\beta}$ values instead of the point estimate $\hat{\beta}$). Let $Q(\tilde{t}; 0.025)$ be the 0.025th quantile of \tilde{t} (that is, the 2.5th percentile) and $Q(\tilde{t}; 0.975)$ be the 0.975th quantile. Then, we could define a 95% confidence band as $[\hat{\beta} + Q(\tilde{t}; 0.025) \text{Std}(\hat{\beta}), \hat{\beta} + Q(\tilde{t}; 0.975) \text{Std}(\hat{\beta})]$, where $\text{Std}(\hat{\beta})$ is a consistent estimate of the standard deviation of $\hat{\beta}$.

One issue with the bootstrap is that it does not directly create observations that obey the null hypothesis, or even a given alternative hypothesis. For instance, it is not straightforward to create samples where a particular coefficient is zero ($\beta_2 = 0$, say). Actually, (3.9) creates a *distribution around the point estimate*, that is, of $\tilde{\beta} - \hat{\beta}$, where $\hat{\beta}$ is the point estimate from the original sample. This is not important if we just want to understand the standard error of a coefficient (since the standard deviation across the bootstrap simulations is already defined in terms of squared deviations around the average value in the bootstraps.). However, it is crucial when we want to understand percentiles of coefficients, t -stats, or χ^2 -stats.

An additional complication is that the average (across bootstrap simulations) estimate $\tilde{\beta}$ may not always equal the point estimate $\hat{\beta}$. If we still want to use the bootstraps to find

critical values, then we have to center the test statistics on the the average estimate in the bootstraps, $\tilde{\beta}$. For instance, for a t -test we calculate

$$t = \frac{\tilde{\beta} - \text{average } \tilde{\beta}}{\text{Std}(\tilde{\beta})} \quad (3.10)$$

for each simulation and then take the $(0.025N)$ th and $(0.975N)$ th observations as the 5% critical values (instead of the ± 1.96 from a standard normal distribution). These critical values can then be used for t -tests based on the original sample, for instance, that $\beta_2 = k$. In most cases, there is little difference between centering on the average $\tilde{\beta}$ and the point estimate $\hat{\beta}$.

A similar reasoning applies to joint tests of coefficients. Consider a linear combination of the coefficients, $R\tilde{\beta}$. If V is the OLS variance-covariance matrix, then for each sample we would calculate the quadratic form

$$\mathcal{E} = [R(\tilde{\beta} - \text{average } \tilde{\beta})]'(RVR')^{-1}[R(\tilde{\beta} - \text{average } \tilde{\beta})]. \quad (3.11)$$

Once again, we could take the $(0.95N)$ th simulated τ as the 5% critical value (instead of the 95th percentiles for a χ_q^2 distribution, for instance, 5.99 for $q = 2$). This critical value can be used to hypotheses like $R\beta = k$ based on the original sample.

In general, the bootstraps of test statistics like the t and \mathcal{E} are more precise than the bootstraps of the regression coefficients themselves—provided that we use consistent estimates of the covariance matrix. (In the limit, these statistics do not depend on model parameters—they asymptotically “pivotal”—which often improves the convergence rate.) For instance, in the case of autocorrelated residuals, this suggests that it might be better to create a bootstrap simulation for t -stats calculated with a Newey-West covariance matrix than a “ t -stat” based on a standard OLS covariance matrix since the latter will have an asymptotic distribution which depends on the autocorrelation (that is, model parameters).

3.3.2 Bootstrapping when x_t Includes Lags of y_t

When x_t contains lagged values of y_t , then we have to modify the approach in (3.9) since \tilde{u}_t can become correlated with x_t . For instance, if x_t includes y_{t-1} and we happen to sample $\tilde{u}_t = \hat{u}_{t-1}$, then we get a non-zero correlation between regressor and residual. The easiest way to handle this is as in the Monte Carlo simulations in (3.4), but where \tilde{u}_t are drawn (with replacement) from the sample of fitted residuals. The same carries over to the VAR model in Remark 3.3.

3.3.3 Bootstrapping when Errors Are Heteroskedastic

Suppose now that the errors are heteroskedastic, but serially uncorrelated. If the heteroskedasticity is unrelated to the regressors, then we can still use (3.9).

However, if the heteroskedasticity is related to the regressors, then it would be wrong to pair x_t with just any $\tilde{u}_t = \hat{u}_s$ since that destroys the relation between x_t and the variance of the residual. (This is the case that White's test for heteroskedasticity tries to identify.)

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 for observation t be $(\tilde{y}_t, \tilde{x}_t) = (y_s, x_s)$ for some random draw of s . Since $(y_s, x_s) = (x'_s \hat{\beta} + \hat{u}_s, x_s)$ we are effectively pairing the fitted residual \hat{u}_s with the contemporaneous regressors x_s . This is called a *paired bootstrap*. Notice that we are 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 3.6 With $T = 3$, the artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (y_2, x_2) \\ (y_3, x_3) \\ (y_3, x_3) \end{bmatrix} = \begin{bmatrix} (x'_2 \hat{\beta} + \hat{u}_2, x_2) \\ (x'_3 \hat{\beta} + \hat{u}_3, x_3) \\ (x'_3 \hat{\beta} + \hat{u}_3, x_3) \end{bmatrix}$$

It could be argued (see, for instance, Davidson and MacKinnon (1993)) that bootstrapping the pairs (y_s, x_s) makes little sense when x_s contains lags of y_s , since the random sampling of the pair (y_s, x_s) destroys the autocorrelation pattern of the regressors.

See Table 7.4 for an application.

Remark 3.7 (*The wild Bootstrap*) The wild bootstrap is also aimed at solving the heteroskedasticity problem. In this case, the artificial sample is generated as in (3.9), but we use $\tilde{u}_t = \hat{u}_t \tilde{\varepsilon}_t$ where \hat{u}_t is the fitted (OLS) residual for observation t and $\tilde{\varepsilon}_t$ is drawn from an iid random variable with mean 0 and variance 1. For instance, $\tilde{\varepsilon}_t$ could have a two-point distribution where it is either -1 or 1 with equal probabilities.

3.3.4 Bootstrapping when Errors Are Autocorrelated

It is quite hard to handle the case when the errors are serially dependent, since we must sample in such a way that we do not destroy the autocorrelation structure of the data. A

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

Table 3.4: Standard error of OLS slope (Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1 + \gamma|z_t| + \alpha|x_t|)^2$, where z_t is iid $N(0, 1)$ and independent of x_t . Sample length: 200. Number of simulations: 25000. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

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 3-observations blocks randomly (with replacement) and stack them to form a longer series.

Example 3.8 With $T = 9$ and a block size of 3, the artificial sample could be

$$\underbrace{\hat{u}_2, \hat{u}_3, \hat{u}_4}_{block\ 2}, \underbrace{\hat{u}_7, \hat{u}_8, \hat{u}_9}_{block\ 7}, \underbrace{\hat{u}_4, \hat{u}_5, \hat{u}_6}_{block\ 4}.$$

To handle end point effects (so that all data points have the same probability to be drawn), we also create blocks by “wrapping” the data around a circle. In practice, this means that we add the following blocks: $(\hat{u}_{10}, \hat{u}_1, \hat{u}_2)$ and $(\hat{u}_9, \hat{u}_{10}, \hat{u}_1)$.

The length of the blocks should clearly depend on the degree of autocorrelation, but $T^{1/3}$ is sometimes recommended as a rough guide. An alternative approach is to have non-overlapping blocks. See [Berkowitz and Kilian \(2000\)](#) for some other approaches.

See Table 3.5 for an illustration.

3.3.5 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

rho:	0.0	0.75
Simulated	5.8	10.2
OLS formula	5.8	7.3
Newey-West	5.7	9.6
VARHAC	5.7	11.1
Bootstrapped	5.5	9.5

Table 3.5: Standard error of OLS intercept (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \xi_t + \rho\xi_{t-1}$, ξ_t is iid $N(0, 1)$. NW uses 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

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.

Chapter 4

Return Distributions

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

4.1 Estimating and Testing Distributions

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

4.1.1 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 ,

$$F_i(x) = \Pr(X_i \leq x). \quad (4.1)$$

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) $f_i(x)$ is the “height” of the distribution in the sense that the cdf $F(x)$ is the integral of the pdf from minus infinity to x

$$F_i(x) = \int_{s=-\infty}^x f_i(s)ds. \quad (4.2)$$

(Conversely, the pdf is the derivative of the cdf, $f_i(x) = \partial F_i(x)/\partial x$.) The Gaussian pdf (the normal distribution) is bell shaped.

Remark 4.1 (*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 4.1 for an illustration.

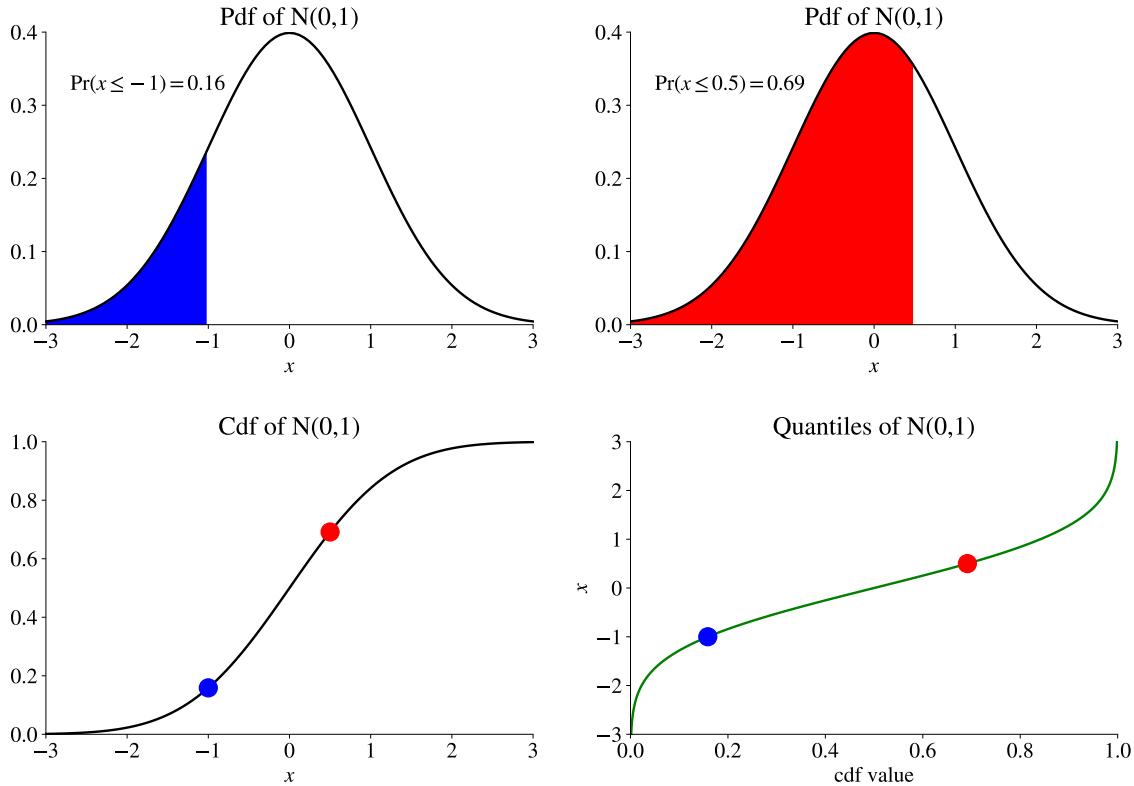


Figure 4.1: Finding quantiles of a $N(\mu, \sigma^2)$ distribution

4.1.2 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 4.2–4.5 for illustrations.

Remark 4.2 (*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 empirical as well as according to the theoretical distribution. For instance, if the 2th percentile (0.02 percentile) is at -10 in the empirical distribution, but at only -3 in the theoretical distribution, then this indicates that the two distributions have very different left tails.

There is one caveat to this way of studying data: it only provides evidence on the unconditional distribution. Suppose instead that we have estimated a model for time-variation in the mean and variance (denoted μ_t and σ_t^2 , respectively), then it makes more sense to study the distribution (QQ plot) of the standardised return

$$\tilde{R}_t = \frac{R_t - \mu_t}{\sigma_t}. \quad (4.3)$$

As a simple example, the mean could be estimated by an AR(1) model (so we would have $\mu_t = a + \rho R_{t-1}$) and the variance by a GARCH model (so we would have $\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2$ where u_{t-1} is the surprise to the return in $t-1$). See Figure 4.6 for an illustration.

4.1.3 Parametric Tests of a Normal Distribution

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

	Test statistic	Distribution
skewness	$= \frac{1}{T} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)^3$	$N(0, 6/T)$
kurtosis	$= \frac{1}{T} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)^4$	$N(3, 24/T)$
Bera-Jarque	$= \frac{T}{6} \text{skewness}^2 + \frac{T}{24} (\text{kurtosis} - 3)^2$	χ_2^2

(4.4)

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (4.4) are under the null hypothesis that x_t is iid $N(\mu, \sigma^2)$. The “excess kurtosis” is defined as the kurtosis minus 3 (as in the Bera-Jarque test).

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

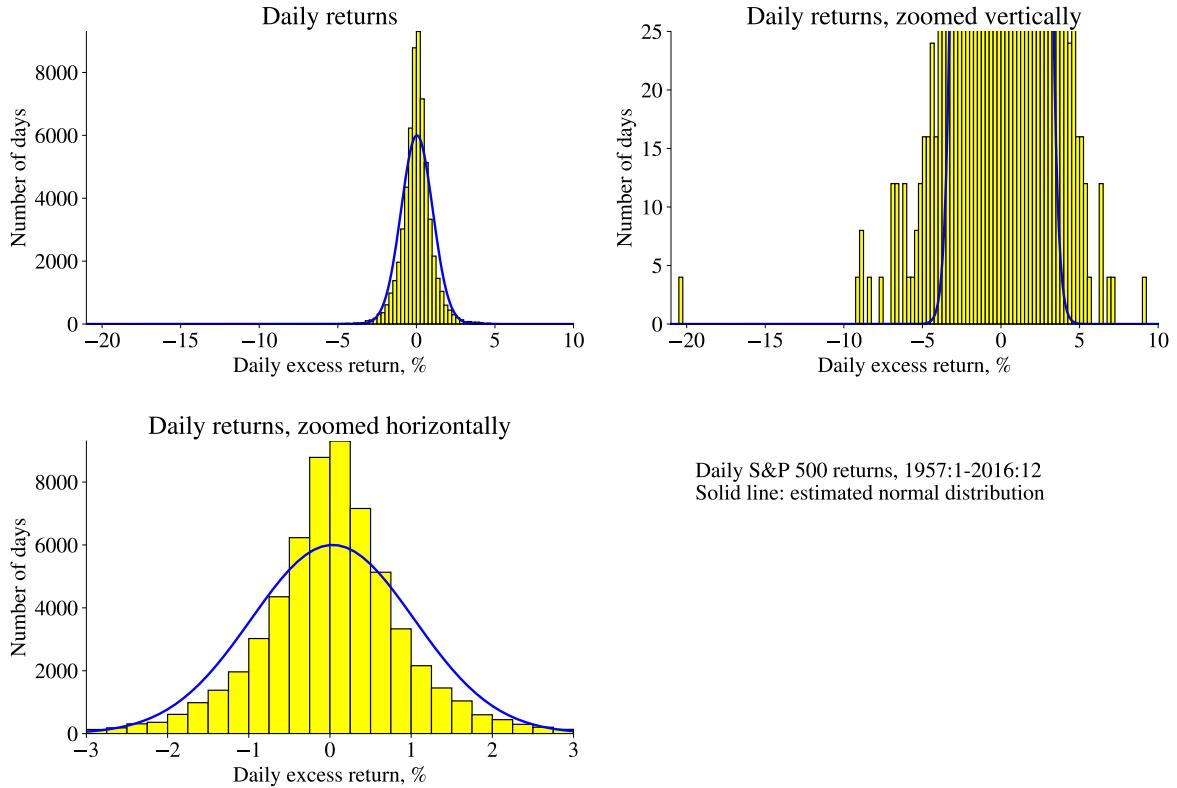


Figure 4.2: Distribution of daily S&P returns

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

$$g(\mu, \sigma^2) = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix}, \quad (4.5)$$

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 $N(\mu, \sigma^2)$.

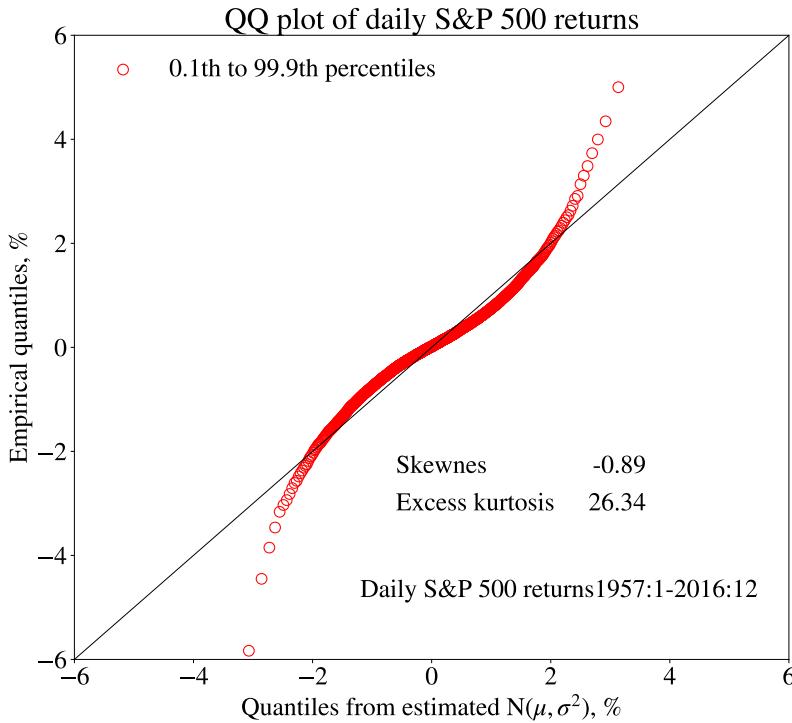


Figure 4.3: Quantiles of daily S&P returns

4.1.4 Nonparametric Tests of General Distributions

The *Kolmogorov-Smirnov* test is designed to test if an empirical distribution function, $\text{EDF}(x)$, conforms with a theoretical cdf, $F(x)$. The empirical distribution function is defined as the fraction of observations which are less or equal to x , that is,

$$\text{EDF}(x) = \frac{1}{T} \sum_{t=1}^T \delta(x_t \leq x), \text{ where} \quad (4.6)$$

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

The $\text{EDF}(x_t)$ and $F(x_t)$ are often plotted against the sorted (in ascending order) sample $\{x_t\}_{t=1}^T$. See [Figure 4.7](#) for an illustration.

Example 4.3 (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 4.7](#).

To perform a (Kolmogorov-Smirnov test of a distribution, first define the absolute

value of the maximum distance

$$D_T = \max_{x_t} |\text{EDF}(x_t) - F(x_t)|. \quad (4.7)$$

See Figure 4.8 for an illustration. Then, reject the null hypothesis that $\text{EDF}(x) = F(x)$ if $\sqrt{T}D_T > c$, where c is a critical value. For instance, the 10%, 5% and 1% critical values are (1.22, 1.36, 1.63). There is a corresponding test for comparing two empirical cdfs.

Remark 4.4 (*Critical values for the K-S test**) The cdf of $\sqrt{T}D_T$ is

$$\lim_{T \rightarrow \infty} \Pr(\sqrt{T}D_T \leq c) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2c^2}.$$

It can be approximated by replacing ∞ with a large number (for instance, 500). To find the 5% critical value, set the left hand side to 0.95 and find the c value (on the right hand side) that makes the equation hold.

Pearson's χ^2 test does the same thing as the K-S test but for a discrete distribution. Suppose you have K categories with N_i values in category i . The theoretical distribution predicts that the fraction p_i should be in category i , with $\sum_{i=1}^K p_i = 1$. Then

$$\sum_{i=1}^K \frac{(N_i - Tp_i)^2}{Tp_i} \sim \chi^2_{K-1}. \quad (4.8)$$

There is a corresponding test for comparing two empirical distributions.

4.1.5 Comparing Distributions

Just comparing histograms is sometimes useful. See Figure 4.9 for an illustration.

Instead, a box plot (which shows the 25th, 50th and 75th percentile, with a notch to indicate a 95% confidence band of the 50th percentile) are also useful. See Figure 4.10 for an illustration. Alternatively, a QQ-plot which plots the quantiles against each other may be a good way of illustrating the difference. See Figure 4.11, which illustrates that growth stocks have more extreme tails than value stocks.

4.1.6 Fitting a Mixture Normal Distribution to Data*

Reference: Hastie, Tibshirani, and Friedman (2001) 8.5

A normal distribution often fits return data 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)

$$f(x_t; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi) = (1 - \pi)\phi(x_t; \mu_1, \sigma_1^2) + \pi\phi(x_t; \mu_2, \sigma_2^2), \quad (4.9)$$

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

See Figures 4.12–4.14 for an illustration.

Remark 4.5 (*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 (4.9). 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

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

- (3) Calculate (in this order)

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

Iterate over (2) and (3) until the parameter values converge. (This is an example of the EM algorithm.) Notice that the calculation of σ_i^2 uses μ_i from the same (not the previous) iteration.

4.1.7 Kernel Density Estimation

Reference: Silverman (1986)

A histogram is just a count of the relative frequency of observations that fall in (pre-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 pdf. Formally, the scaled histogram at the point x (say, $x = 2.3$) is defined as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{h} K\left(\frac{x_t - x}{h}\right), \text{ where} \quad (4.10)$$

$$K\left(\frac{x_t - x}{h}\right) = \begin{cases} 1 & \text{if } \left|\frac{x_t - x}{h}\right| \leq 1/2 \\ 0 & \text{else.} \end{cases}$$

Notice that $K() = 1$ if x_t is in the interval $x \pm h/2$ and zero otherwise. The $K()$ is called the “kernel” and the formula shows that our estimate of the pdf (here: the scaled histogram) is an average of kernel values (scaled by $1/h$) of the data. See Figure 6.2.

We can gain efficiency and get a smoother (across x values) estimate by using another kernel. In particular, a kernel that tapers off slowly instead of suddenly dropping to zero, as the one in (4.10) does, improves the properties.

We clearly want the estimated pdf be non-negative and to have an integral equal to one. To guarantee these properties, we require that $K(u) \geq 0$ and $\int_{u=-\infty}^{\infty} K(u) du = 1$. In practice, we also require the kernel to be symmetric around zero, so $K(u) = K(-u)$. This means that the kernel function must be some type of density function. (Do not get lost here: we are trying to estimate an unknown density function from data and use a kernel as a simplifying device. The fact that the kernel looks like a pdf does not force our estimate to be similar to the kernel.)

Remark 4.6 (*Requirements on the kernel**) It is clear that using $K(u) \geq 0$ in (4.10) gives $\hat{f}(x) \geq 0$. To show that it integrates to one, notice that

$$\int_{x=-\infty}^{\infty} \hat{f}(x) dx = \frac{1}{T} \sum_{t=1}^T \frac{1}{h} \int_{x=-\infty}^{\infty} K\left(\frac{x - x_t}{h}\right) dx.$$

We use the symmetry of the kernel to write the argument as $(x - x_t)/h$, since it simplifies a bit. On the right hand side, change variable from x to u where $u = (x - x_t)/h$ (so $x = uh + x_t$) and notice that $dx/du = h$. We then get (since the integration limits are

still $-\infty$ and ∞)

$$\frac{1}{T} \sum_{t=1}^T \frac{1}{h} \int_{u=-\infty}^{\infty} K(u) h du = 1,$$

since the h terms cancel and the integral is 1.

A $N(0, 1)$ pdf is a common choice for the kernel. The kernel density estimator of the pdf at some point x is then as in (4.10), but with

$$\frac{1}{h} K\left(\frac{x_t - x}{h}\right) = \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - x}{h}\right)^2\right]. \quad (4.11)$$

Notice that (4.11) is really the pdf of a $N(x, h^2)$ distribution. See Figure 4.16 for an example of the values from (4.11).

It can be shown (see Silverman (1986) 3.3) that under the assumption that x_t is iid, the mean squared error, variance and bias of the estimator at the value x are approximately (for general kernel functions)

$$\begin{aligned} \text{MSE}(x) &= \text{Var}[\hat{f}(x)] + \text{Bias}[\hat{f}(x)]^2, \text{ with} \\ \text{Var}[\hat{f}(x)] &= \frac{1}{Th} f(x) \times \int_{-\infty}^{\infty} K(u)^2 du \\ \text{Bias}[\hat{f}(x)] &= h^2 \times \frac{1}{2} \frac{d^2 f(x)}{dx^2} \times \int_{-\infty}^{\infty} K(u) u^2 du. \end{aligned} \quad (4.12)$$

In these expressions, $f(x)$ is the true (unknown) density of x and $K(u)$ the kernel (pdf) used as a weighting function for $u = (x_t - x)/h$. With an $N(0, 1)$ kernel these expressions can be simplified to

$$\begin{aligned} \text{Var}[\hat{f}(x)] &= \frac{1}{Th} f(x) \times \frac{1}{2\sqrt{\pi}} \\ \text{Bias}[\hat{f}(x)] &= h^2 \times \frac{1}{2} \frac{d^2 f(x)}{dx^2}. \end{aligned} \quad (4.13)$$

Proof. (of (4.13)) 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 (4.12). ■

It can then be shown that (with iid data and a Gaussian kernel) the asymptotic distri-

bution is

$$\sqrt{Th}[\hat{f}(x) - f(x)] \xrightarrow{d} N\left[0, \frac{1}{2\sqrt{\pi}}f(x)\right], \quad (4.14)$$

provided h is decreased (as T increases) slightly faster than $T^{-1/5}$ (for instance, suppose $h = T^{-1.1/5}h_0$, where h_0 is a constant). Notice the \sqrt{Th} term on left hand side (the usual expression in parametric models include only \sqrt{T}).

Remark 4.7 (*Asymptotic bias*) *The condition that h decreases faster than $T^{-1/5}$ ensures that the bias of $\sqrt{Th}\hat{f}(x)$ vanishes as $T \rightarrow \infty$. This is seen by noticing that the bias in (4.13) is proportional to h^2 . Combining gives the bias of $\sqrt{Th}\hat{f}(x)$ as being proportional to $T^{1/2}h^{5/2}$. If indeed $h = T^{-1.1/5}h_0$, then we have*

$$T^{1/2}h^{5/2} = T^{-0.05}h_0^{5/2}$$

which is decreasing to zero as T increases.

The value $h = 1.06 \text{Std}(x_t)T^{-1/5}$ is sometimes recommended, since it can be shown to be the optimal choice (in MSE sense) if data is normally distributed and the Gaussian kernel is used.(See below for a proof.) The bandwidth h could also be chosen by a leave-one-out cross-validation technique. See Figure 4.17 for an example and Figure 4.18 for a cross-validation approach to determine the bandwidth.

Proof. (Best bandwidth if $f(x) \sim N$) Use (4.13) to write the mean integrated squared error (MISE) as

$$\text{MISE}(x) = \int \text{MSE}(x)dx = \frac{1}{Th} \int f(x)dx \times \frac{1}{2\sqrt{\pi}} + h^4 \times \frac{1}{4} \int \left[\frac{d^2 f(x)}{dx^2} \right]^2 dx.$$

Notice that $\int f(x)dx = 1$ and if $f(x)$ is the pdf of $N(0, \sigma^2)$, then the last integral is $3/(8\sigma^5\sqrt{\pi})$. Combining gives

$$\text{MISE}(x) = \frac{1}{Th} \frac{1}{2\sqrt{\pi}} + h^4 \times \frac{1}{4} \frac{3}{8\sigma^5\sqrt{\pi}}.$$

The first order condition with respect to h is

$$\begin{aligned} 0 &= \frac{-1}{Th^2} + h^3 \frac{3}{4\sigma^5}, \text{ or} \\ \frac{4\sigma^5}{3T} &= h^5, \text{ so} \\ h &= \left(\frac{4}{3}\right)^{1/5} \sigma T^{-1/5}. \end{aligned}$$

Notice that $(4/3)^{1/5} \approx 1.06$. ■

Remark 4.8 (*Cross-validation to choose h*) If x_t is iid, then the following cross-validation approach can be used to find the best bandwidth. Define the integrated squared error (ISE) as $\int [f(x) - \hat{f}(x)]^2 dx$. (ISE is specific for a sample, while the MISE discussed above is the expected value of ISE across all samples. The difference is small in large samples.) Expand as $ISE = \int [f(x)^2 + \hat{f}(x)^2 - 2f(x)\hat{f}(x)]dx$. For the first term in ISE, notice that $\int f(x)^2 dx$ does not depend on the bandwidth, so we can treat it as a constant, α . If we use an $N(0, 1)$ kernel, then it can be shown that the middle term in ISE is

$$\int \hat{f}(x)^2 dx = \frac{1}{T} \sum_{t=1}^T \hat{g}(x_t), \text{ where } \hat{g}(x_t) = \frac{1}{T} \sum_{s=1}^T \frac{1}{h} \eta\left(\frac{x_t - x_s}{h}\right).$$

where $\eta(x)$ is the pdf of a $N(0, 2)$. The last term in ISE is the same as $-2 \mathbb{E}_x \hat{f}(x)$, where \mathbb{E}_x denotes that the expectation is over which x value that is realized. If we had access to another sample (\tilde{x}_i) , then we could plug in those values in the (already estimated) $\hat{f}()$ function to estimate $\mathbb{E}_x \hat{f}(x)$ as the average $\hat{f}(\tilde{x}_i)$ value. Instead, we use a leave-one-out approach which preserves the property that $\hat{f}()$ does not depend on the x_i value: estimate $\hat{f}_{-t}(x)$ by excluding data point t from the sample, and then evaluate it at $x = x_t$, $\hat{f}_{-t}(x_t)$. (The alternative of using the same sample for both estimation and evaluation would lead to a classical overfitting, driving the h value towards zero.) Then, estimate $\mathbb{E}_x \hat{f}(x)$ by the average $\hat{f}_{-t}(x_t)$ value. To sum up, pick h to minimize $ISE(h)$, where

$$ISE(h) = \alpha + \frac{1}{T} \sum_{t=1}^T [\hat{g}(x_t) - 2\hat{f}_{-t}(x_t)].$$

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 $d \times 1$ vector and $\hat{\Omega}$ be the estimated covariance matrix of x_t . We can then estimate the pdf at a point x by using a multivariate Gaussian kernel as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \frac{1}{(2\pi)^{d/2} |H^2 \hat{\Omega}|^{1/2}} \exp\left[-\frac{1}{2} (x_t - x)' (H^2 \hat{\Omega})^{-1} (x_t - x)\right]. \quad (4.15)$$

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 4.9 ((4.15) with $d = 1$) With just one variable, (4.15) becomes

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

which is the same as (4.11) if $h = H \text{Std}(x_t)$.

4.1.8 “Foundations of Technical Analysis...” by Lo, Mamaysky and Wang (2000)

Reference: Lo, Mamaysky, and Wang (2000)

Topic: is the distribution of the return different after a “signal?” This paper uses kernel regressions to identify and implement some technical trading rules, and then tests if the 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 4.19–4.20 for an illustration.

4.2 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)

4.2.1 The Breeden-Litzenberger Approach

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

$$C = E m \max(0, S - X), \quad (4.16)$$

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

$$C = \exp(-ik) E^* \max(0, S - X), \quad (4.17)$$

where E^* is the risk-neutral expectation.

Example 4.10 (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 (4.17) 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.

(4.17) can also be written as

$$C = \exp(-ik) \int_X^\infty (S - X) h^*(S) dS, \quad (4.18)$$

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 (4.18) with respect to the strike price and rearranging gives the risk-neutral distribution function

$$\Pr^*(S \leq X) = 1 + \exp(i k) \frac{\partial C(X)}{\partial X}. \quad (4.19)$$

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 $S = X$

$$\text{pdf}^*(X) = \exp(i k) \frac{\partial^2 C(X)}{\partial X^2}. \quad (4.20)$$

Figure 4.21 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.)

A difference quotient approximation of the derivative in (4.19)

$$\frac{\partial C}{\partial X} \approx \frac{1}{2} \left[\frac{C(X_{i+1}) - C(X_i)}{X_{i+1} - X_i} + \frac{C(X_i) - C(X_{i-1})}{X_i - X_{i-1}} \right] \quad (4.21)$$

gives the approximate distribution function. The approximate probability density func-

tion, obtained by a second-order difference quotient

$$\frac{\partial^2 C}{\partial X^2} \approx \left[\frac{C(X_{i+1}) - C(X_i)}{X_{i+1} - X_i} - \frac{C(X_i) - C(X_{i-1})}{X_i - X_{i-1}} \right] / \left[\frac{1}{2} (X_{i+1} - X_{i-1}) \right] \quad (4.22)$$

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 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.

4.2.2 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

$$\text{pdf} \left(\begin{bmatrix} \ln m \\ \ln S \end{bmatrix} \right) = \sum_{j=1}^n \alpha^{(j)} \phi \left(\begin{bmatrix} \ln m \\ \ln S \end{bmatrix}; \begin{bmatrix} \mu_m^{(j)} \\ \mu_s^{(j)} \end{bmatrix}, \begin{bmatrix} \sigma_{mm}^{(j)} & \sigma_{ms}^{(j)} \\ \sigma_{ms}^{(j)} & \sigma_{ss}^{(j)} \end{bmatrix} \right), \quad (4.23)$$

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 $\Phi(\cdot)$ be the standardized (univariate) normal distribution function. If $\mu_m^{(j)} = \mu_m$ and $\sigma_{mm}^{(j)} = \sigma_{mm}$ in (4.23), then the marginal distribution of the log SDF is Gaussian (while that of the underlying asset price is not). In this case the European call option price (4.16) has a closed form solution in terms of the spot interest rate, strike price, and the

parameters of the bivariate distribution¹

$$C = \exp(-ik) \sum_{j=1}^n \alpha^{(j)} \left[\exp\left(\mu_s^{(j)} + \sigma_{ms}^{(j)} + \frac{1}{2}\sigma_{ss}^{(j)}\right) \Phi\left(\frac{\mu_s^{(j)} + \sigma_{ms}^{(j)} + \sigma_{ss}^{(j)} - \ln X}{\sqrt{\sigma_{ss}^{(j)}}}\right) - X\Phi\left(\frac{\mu_s^{(j)} + \sigma_{ms}^{(j)} - \ln X}{\sqrt{\sigma_{ss}^{(j)}}}\right) \right]. \quad (4.24)$$

(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

$$C = \sum_{j=1}^n \alpha^{(j)} C^{(j)}. \quad (4.25)$$

(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 (4.24) as

$$F = \sum_{j=1}^n \alpha^{(j)} \exp\left(\mu_s^{(j)} + \sigma_{ms}^{(j)} + \frac{\sigma_{ss}^{(j)}}{2}\right). \quad (4.26)$$

There are several reasons for assuming a mixture of normal distributions. First, nonparametric 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 (4.26), $F = \exp(\mu_s + \sigma_{ms} + \sigma_{ss}/2)$, in the option price (4.24) to get

$$C = \exp(-ik) F \Phi\left(\frac{\ln F/X + \sigma_{ss}/2}{\sqrt{\sigma_{ss}}}\right) - \exp(-ik) X \Phi\left(\frac{\ln F/X - \sigma_{ss}/2}{\sqrt{\sigma_{ss}}}\right), \quad (4.27)$$

¹Without these restrictions, $\alpha^{(j)}$ in (4.24) 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.

which is indeed Black's formula.

We want to estimate the marginal distribution of the future asset price, S . From (4.23), 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 (4.24)–(4.26). For that we need data on at least as many different strike prices as there are parameters to estimate.

Remark 4.11 Figures 4.21–4.22 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..

Remark 4.12 Figures 4.23–4.25 show results for the CHF/EUR exchange rate around the period of active (Swiss) central bank interventions on the currency market.

Remark 4.13 (Robust measures of the standard deviation and skewness) Let P_α be the α th quantile (for instance, quantile 0.1) of a distribution. A simple robust measure of the standard deviation is just the difference between two symmetric quantiles,

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

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.

One of the classical robust skewness measures was suggested by Hinkley

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

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.

4.3 Threshold Exceedance and Tail Distribution*

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

In risk control, the focus is on the distribution of losses beyond some threshold level. This has three direct implications. First, the object under study is the loss

$$X = -R, \tag{4.28}$$

that is, the negative of the return. Second, the attention is on how the distribution looks like beyond a threshold and also on the probability of exceeding this threshold. In contrast, the exact shape of the distribution below that point is typically disregarded. Third, modelling the tail of the distribution is best done by using a distribution that allows for a much heavier tail than suggested by a normal distribution. The generalized Pareto (GP) distribution is often used. See *Figure 4.26* for an illustration.

Remark 4.14 (*Cdf and pdf of the generalized Pareto distribution*) *The generalized Pareto distribution is described by a scale parameter ($\beta > 0$) and a shape parameter (ξ). The cdf ($\Pr(Z \leq z)$, where Z is the random variable and z is a value) is*

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

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

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

The mean is defined (finite) if $\xi < 1$ and is then $E(Z) = \beta/(1-\xi)$. Similarly, the variance is finite if $\xi < 1/2$ and is then $\text{Var}(Z) = \beta^2/[(1-\xi)^2(1-2\xi)]$. See *Figure 4.27* for an illustration.

Remark 4.15 (*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*

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

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

$$F(x) = P_u + G(x - u)(1 - P_u), \text{ for } x > u, \quad (4.29)$$

where $G(z)$ is the cdf of the generalized Pareto distribution. Noticed that, the cdf value is P_u at $x = u$ (or just slightly above u), and that it becomes one as x goes to infinity.

Clearly, the pdf is

$$f(x) = g(x - u)(1 - P_u), \text{ for } x > u, \quad (4.30)$$

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 4.29 for an illustration.

It is often to calculate the *tail probability* $\Pr(X > x)$, which in the case of the cdf in (4.29) is

$$1 - F(x) = (1 - P_u)[1 - G(x - u)], \quad (4.31)$$

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

The VaR_α (say, $\alpha = 95\%$) is the α -th quantile of the loss distribution

$$\text{VaR}_\alpha = \text{cdf}_X^{-1}(\alpha), \quad (4.32)$$

where $\text{cdf}_X^{-1}()$ is the inverse cumulative distribution function of the losses. That is, VaR_α is the α quantile of the loss distribution. For instance, $\text{VaR}_{95\%}$ is the 0.95 quantile of the loss distribution. This clearly means that the probability of the loss to be less than VaR_α equals α

$$\Pr(X \leq \text{VaR}_\alpha) = \alpha. \quad (4.33)$$

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

Assuming $\text{VaR}_\alpha \geq u$ (that is, $\alpha \geq P_u$), the cdf (4.29) together with the form of the generalized Pareto distribution give

$$\text{VaR}_\alpha = \begin{cases} u + \frac{\beta}{\xi} \left[\left(\frac{1-\alpha}{1-P_u} \right)^{-\frac{1}{\xi}} - 1 \right] & \text{if } \xi \neq 0 \\ u - \beta \ln \left(\frac{1-\alpha}{1-P_u} \right) & \xi = 0 \end{cases}, \text{ for } \alpha \geq P_u. \quad (4.34)$$

Proof. (of (4.34)) Set $F(x) = \alpha$ in (4.29) and use $z = x - u$ in the cdf from Remark 4.14 and solve for x . ■

If we assume $\xi < 1$ (to make sure that the mean is finite), then straightforward integration using (4.30) shows that the expected shortfall is

$$\begin{aligned} \text{ES}_\alpha &= \mathbb{E}(X | X \geq \text{VaR}_\alpha) \\ &= \frac{\text{VaR}_\alpha}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}, \text{ for } \alpha > P_u \text{ and } \xi < 1. \end{aligned} \quad (4.35)$$

The expected exceedance of a GP distribution (with $\xi < 1$) for any threshold $v > u$ is

$$\begin{aligned} e(v) &= \mathbb{E}(X - v | X > v) \\ &= \frac{\xi v}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}, \text{ for } v > u \text{ and } \xi < 1. \end{aligned} \quad (4.36)$$

Proof. (of (4.36)) Substitute v for VaR_α in the expected shortfall (4.35)

$$\mathbb{E}(X | X \geq v) = \frac{v}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}$$

and subtract v from both sides to get (4.36). ■

The expected exceedance of a generalized Pareto distribution (with $0 < \xi < 1$) is increasing with the threshold level v . This indicates that the tail of the distribution is very long. In contrast, a normal distribution would typically show a negative relation (see Figure 4.29 for an illustration). This provides a way of assessing which distribution that best fits the tail of the historical histogram.

Remark 4.16 (*Expected exceedance from a normal distribution*) If $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} \mathbb{E}(X - v | X > v) &= \mu + \sigma \frac{\phi(v_0)}{1 - \Phi(v_0)} - v, \\ \text{with } v_0 &= (v - \mu)/\sigma \end{aligned}$$

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

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

$$\begin{aligned} \hat{e}(v) &= \frac{\sum_{t=1}^T (X_t - v) \delta(X_t > v)}{\sum_{t=1}^T \delta(X_t > v)}, \text{ where} \\ \delta(q) &= \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{else.} \end{cases} \end{aligned} \quad (4.37)$$

If it is found that $\hat{e}(v)$ is increasing (more or less) linearly with the threshold level (v) as in (4.36), 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 (4.37) with the theoretical (4.36) can help. Suppose we calculate the empirical exceedance for

different values of the threshold level (denoted v_i —all large enough so the relation looks linear), then we can estimate (by LS)

$$\hat{e}(v_i) = a + bv_i + \varepsilon_i. \quad (4.38)$$

Then, the theoretical exceedance (4.36) for a given starting point of the GPD u is related to this regression according to

$$\begin{aligned} a &= \frac{\beta - \xi u}{1 - \xi} \text{ and } b = \frac{\xi}{1 - \xi}, \text{ or} \\ \xi &= \frac{b}{1 + b} \text{ and } \beta = a(1 - \xi) + \xi u. \end{aligned} \quad (4.39)$$

See Figure 4.30 for an illustration.

Remark 4.17 (*Log likelihood function of the loss distribution*) Since we have assumed that the threshold exceedance $(X_t - u)$ has a generalized Pareto distribution, Remark 4.14 shows that the log likelihood for the observation of the loss above the threshold $(X_t > u)$ is

$$\begin{aligned} L &= \sum_{t \text{ st. } X_t > u} L_t \\ \ln L_t &= \begin{cases} -\ln \beta - (1/\xi + 1) \ln [1 + \xi (X_t - u) / \beta] & \text{if } \xi \neq 0 \\ -\ln \beta - (X_t - u) / \beta & \xi = 0. \end{cases} \end{aligned}$$

This allows us to estimate ξ and β by maximum likelihood. Typically, u is not estimated, but imposed *a priori* (based on the expected exceedance).

Example 4.18 (*Estimation of the generalized Pareto distribution on S&P daily returns*). Figure 4.30 (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.

4.4 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

$$\text{Corr}(x_t, y_t | h_1 < x_t \leq h_2, k_1 < y_t \leq k_2). \quad (4.40)$$

For instance, by setting the lower boundaries (h_1 and k_1) to $-\infty$ and the upper boundaries (h_2 and k_2) to 0, we get the correlation in down markets.

A (bivariate) normal distribution has little probability mass at very low returns, which leads to the correlation being squeezed towards zero as we only consider data far out in the tail. 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 4.31.

In contrast, Figures 4.32–4.33 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.

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

Spearman's rank correlation (called Spearman's rho and often denoted ρ_S) measures to what degree two variables have a monotonic relation: it is the correlation of their respective ranks. It measures if one variable tends to be high when the other also is—without imposing the restriction that this relation must be linear.

It is computed in two steps. First, the data is *ranked* from the smallest (rank 1) to

the largest (ranked T , where T is the sample size). Ties (when two or more observations have the same values) are handled by averaging the ranks. The following illustrates this for two variables

x_t	rank(x_t)	y_t	rank(y_t)
2	2.5	7	2
10	4	8	3
-3	1	2	1
2	2.5	10	4

(4.41)

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

$$\text{Spearman's } \rho = \text{Corr}[\text{rank}(x_t), \text{rank}(y_t)]. \quad (4.42)$$

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, $d_t = \text{rank}(x_t) - \text{rank}(y_t)$, $\rho = 1 - 6\sum_{t=1}^T d_t^2 / (T^3 - T)$. It gives the same result if there are no tied ranks.) See Figure 4.34 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

$$\sqrt{T-1} \times \text{Spearman's } \rho \xrightarrow{d} N(0, 1). \quad (4.43)$$

(For samples of 20 to 40 observations, it is often recommended to use $\sqrt{(T-2)/(1-\hat{\rho}_S^2)}\hat{\rho}_S$ where $\hat{\rho}_S$ denotes Spearman's ρ . This has a t_{T-2} distribution.)

Remark 4.19 (*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, \dots, x_{t-1}) with the corresponding changes of y_t . For instance, with three data points $((x_1, y_1), (x_2, y_2), (x_3, y_3))$ we first calculate

Changes of x	Changes of y
$x_2 - x_1$	$y_2 - y_1$
$x_3 - x_1$	$y_3 - y_1$
$x_3 - x_2$	$y_3 - y_2$,

(4.44)

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

$$\begin{aligned} ij \text{ is concordant if } (x_j - x_i)(y_j - y_i) &> 0 \\ ij \text{ is discordant if } (x_j - x_i)(y_j - y_i) &< 0. \end{aligned} \quad (4.45)$$

Finally, we count the number of concordant (T_c) and discordant (T_d) pairs and calculate Kendall's tau as

$$\text{Kendall's } \tau = \frac{T_c - T_d}{T(T - 1)/2}. \quad (4.46)$$

It can be shown that

$$\text{Kendall's } \tau \xrightarrow{d} N\left(0, \frac{4T + 10}{9T(T - 1)}\right), \quad (4.47)$$

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

Example 4.20 (*Kendall's tau*) Suppose the data is

$$\begin{array}{cc} \underline{x} & \underline{y} \\ 2 & 7 \\ 10 & 9 \\ -3 & 10. \end{array}$$

We then get the following changes

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

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

$$\text{Spearman's rho} = \frac{6}{\pi} \arcsin(\rho/2) \approx \rho \quad (4.48)$$

$$\text{Kendall's tau} = \frac{2}{\pi} \arcsin(\rho). \quad (4.49)$$

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 4.35. 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

$$G_\alpha = \Pr(x \leq \xi_{x,\alpha}, y \leq \xi_{y,\alpha}), \quad (4.50)$$

$\xi_{x,\alpha}$ 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

$$\begin{aligned} \hat{G}_\alpha &= \frac{1}{T} \sum_{t=1}^T \delta_t, \text{ where} \\ \delta_t &= \begin{cases} 1 & \text{if } x_t \leq \hat{\xi}_{x,\alpha} \text{ and } y_t \leq \hat{\xi}_{y,\alpha} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (4.51)$$

See Figure 4.36 for an illustration based on a joint normal distribution.

4.6 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 joint (non-normal) distributions: using copulas is one. This approach has the advantage that it proceeds in two steps: first we estimate the marginal distribution of each return separately, then we model the comovements by a copula.

4.6.1 Multivariate Distributions and Copulas

Any pdf can also be written as

$$f_{1,2}(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2), \text{ with} \quad (4.52)$$

$$u_i = F_i(x_i),$$

where $c()$ is a *copula density* function and $u_i = F_i(x_i)$ is short-hand notation for the cdf value as in (4.1). The extension to three or more random variables is straightforward.

Equation (4.52) 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 marginal (univariate) 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 4.21 (*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 4.22 (*Joint cdf*) A joint cdf of two random variables (X_1 and X_2) is defined as

$$F_{1,2}(x_1, x_2) = \Pr(X_1 \leq x_1 \text{ and } X_2 \leq x_2).$$

This cdf is obtained by integrating the joint pdf $f_{1,2}(x_1, x_2)$ over both variables

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

(Conversely, the pdf is the mixed derivative of the cdf, $f_{1,2}(x_1, x_2) = \partial^2 F_{1,2}(x_1, x_2) / \partial x_1 \partial x_2$.) See Figure 4.37 for an illustration.

Remark 4.23 (*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 $f_1(x_1) = \int_{x_2=-\infty}^{\infty} f_{1,2}(x_1, x_2) dx_2$.

Remark 4.24 (*Joint pdf and copula density, n variables*) For n variables (4.52) generalizes to

$$f_{1,2,\dots,n}(x_1, x_2, \dots, x_n) = c(u_1, u_2, \dots, u_n) f_1(x_1) f_2(x_2) \dots f_n(x_n), \text{ with}$$

$$u_i = F_i(x_i),$$

Remark 4.25 (*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 (4.52) 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 $C()$.

4.6.2 The Gaussian and Other Copula Densities

The bivariate Gaussian copula density function is

$$c(u_1, u_2) = \frac{1}{\sqrt{1 - \rho^2}} \exp\left(-\frac{\rho^2 \xi_1^2 - 2\rho \xi_1 \xi_2 + \rho^2 \xi_2^2}{2(1 - \rho^2)}\right), \text{ with} \quad (4.53)$$

$$\xi_i = \Phi^{-1}(u_i),$$

where $\Phi^{-1}()$ is the inverse of an $N(0, 1)$ distribution.

Notice that when using this function in (4.52) 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 $\xi_i = \Phi^{-1}(u_i) = \Phi^{-1}[F_i(x_i)]$. This is used in (4.53) (and finally in (4.52)). See Figure 4.38 for an illustration.

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.

A zero correlation ($\rho = 0$) makes the copula density (4.53) 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

$$\rho = \text{Spearman's rho}, \quad (4.54)$$

as suggested by (4.48).

Alternatively, the ρ parameter can be calibrated to give a joint probability of both x_1 and x_2 being lower than some quantile—to match the properties of data: see (4.51). The value 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 4.36 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 4.26 (*The Gaussian copula function**) *The distribution function corresponding to the Gaussian copula density (4.53) 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 (4.53) 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 routines for calculating this cdf.*

Remark 4.27 (*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 i th 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

$$c(u_1, u_2) = (-1 + u_1^{-\alpha} + u_2^{-\alpha})^{-2-1/\alpha} (u_1 u_2)^{-\alpha-1} (1 + \alpha), \quad (4.55)$$

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 4.38 for an illustration.

For the Clayton copula we have

$$\text{Kendall's } \tau = \frac{\alpha}{\alpha + 2}, \text{ so} \quad (4.56)$$

$$\alpha = \frac{2\tau_K}{1 - \tau_K}, \quad (4.57)$$

where τ_K denotes Kendall's τ . The easiest way to calibrate a Clayton copula is therefore to set the parameter α according to (4.57).

Figure 4.39 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 4.36 for the Gaussian copula, see (4.48)–(4.49). 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 4.28 (*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 4.29 (*Clayton copula function**) *The copula function (the cdf) corresponding to (4.55) 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 (4.1).
2. Calculate the copula density as follows (for the Gaussian or Clayton copulas, respectively):

- (a) for the Gaussian copula (4.53)
 - 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 (4.55)
 - i. assume (or estimate/calibrate) an α to use in the Clayton copula (typically based on Kendall's τ as in (4.57))
 - ii. calculate the copula density value $c(u_1, u_2)$
3. Combine the marginal pdfs and the copula density as in (4.52), $f_{1,2}(x_1, x_2) = c(u_1, u_2)f_1(x_1)f_2(x_2)$, where $u_i = F_i(x_i)$ is the cdf value according to the marginal distribution of variable i .

See Figures 4.40–4.41 for illustrations.

Remark 4.30 (*Tail Dependence**) *The measure of lower tail dependence starts by finding the probability that X_1 is lower than its q th quantile ($X_1 \leq F_1^{-1}(q)$) given that X_2 is lower than its q th quantile ($X_2 \leq F_2^{-1}(q)$)*

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

and then takes the limit as the quantile goes to zero

$$\lambda_l = \lim_{q \rightarrow 0} \Pr[X_1 \leq F_1^{-1}(q) | X_2 \leq F_2^{-1}(q)].$$

It can be shown that a Gaussian copula gives zero or very weak tail dependence, unless the correlation is 1. It can also be shown that the lower tail dependence of the Clayton copula is

$$\lambda_l = 2^{-1/\alpha} \text{ if } \alpha > 0$$

and zero otherwise.

4.7 Joint Tail Distribution*

The methods for estimating the (marginal) distribution of the lower tail for each return 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 GPD. 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 4.42–4.44 provide an illustration.

Remark 4.31 *Figure 4.42 suggests that the joint occurrence (of these two assets) of really 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 (4.44) 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 ($u_i, i = 1, \dots, n$). Second, calculate the random number (“return”) by inverting the cdf $u_i = F_i(x_i)$ in (4.52) as

$$x_i = F_i^{-1}(u_i), \quad (4.58)$$

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

Remark 4.32 *(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 4.33 *(To draw n random numbers from a Clayton copula) First, draw x_i for $i = 1, \dots, n$ from a uniform distribution (between 0 and 1). Second, draw v from a $\text{gamma}(1/\alpha, 1)$ distribution. Third, calculate $u_i = [1 - \ln(x_i)/v]^{-1/\alpha}$ for $i = 1, \dots, n$. These u_i values are the marginal cdf values.*

Remark 4.34 *(Inverting a normal and a generalised Pareto cdf) Must numerical software packages contain a routine for inverting a normal cdf. Remark 4.14 shows how to generate random numbers for a generalised Pareto 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 4.45–4.46 give an illustration of how the movements in the lower get more synchronised as the α parameter in the Clayton copula increases.

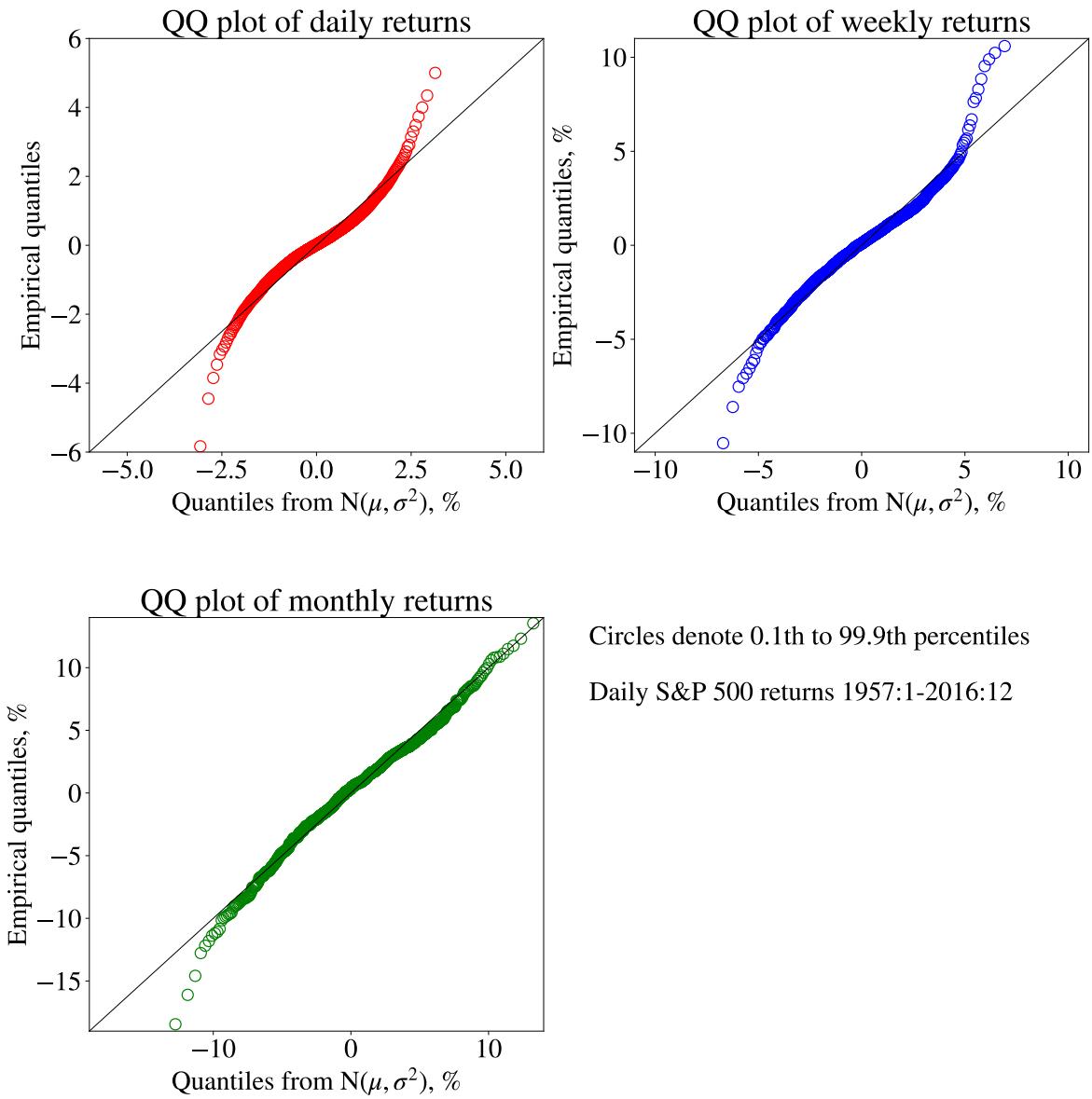


Figure 4.4: Distribution of S&P returns (different horizons)

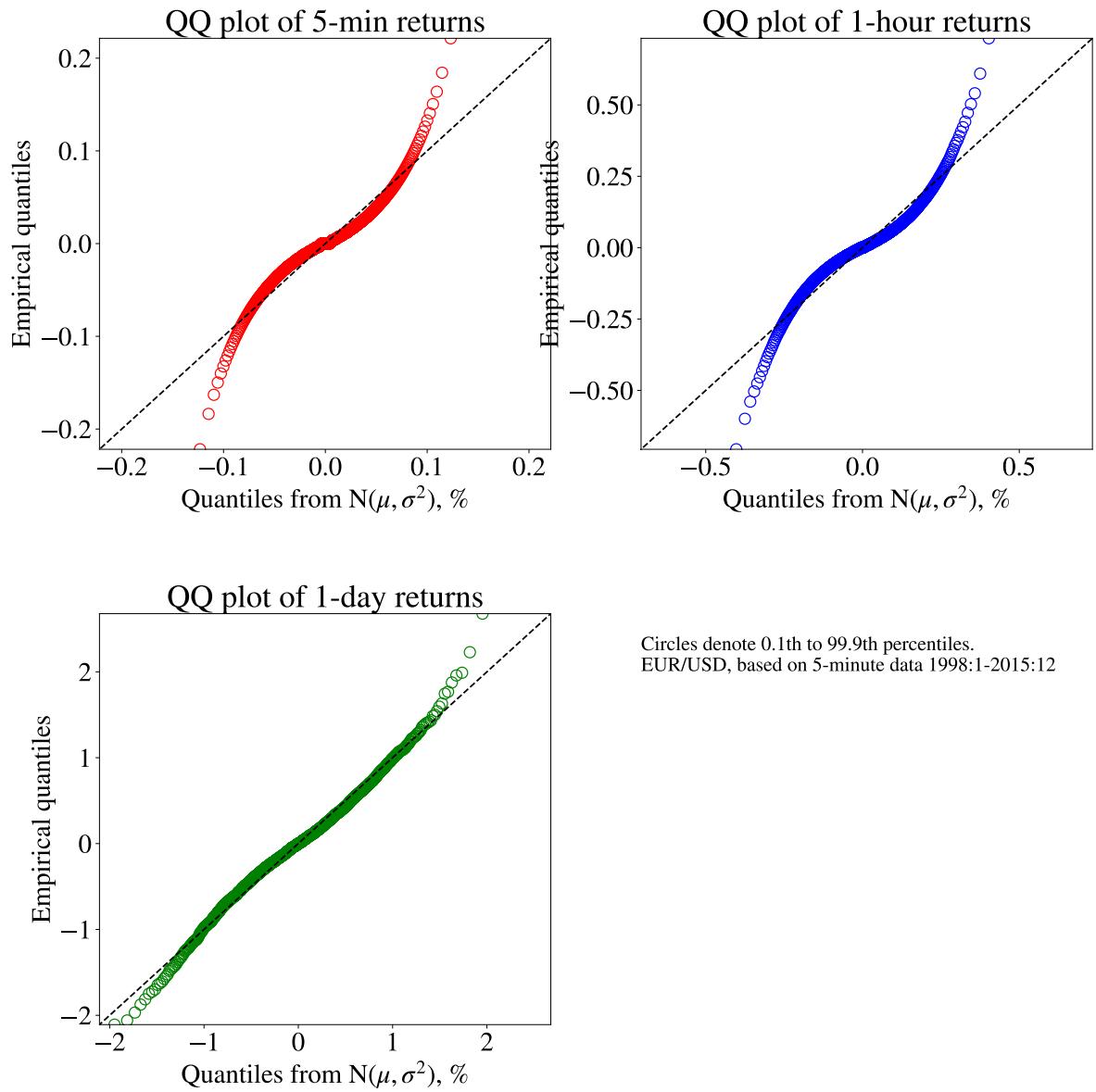


Figure 4.5: Distribution of exchange rate returns (different horizons)

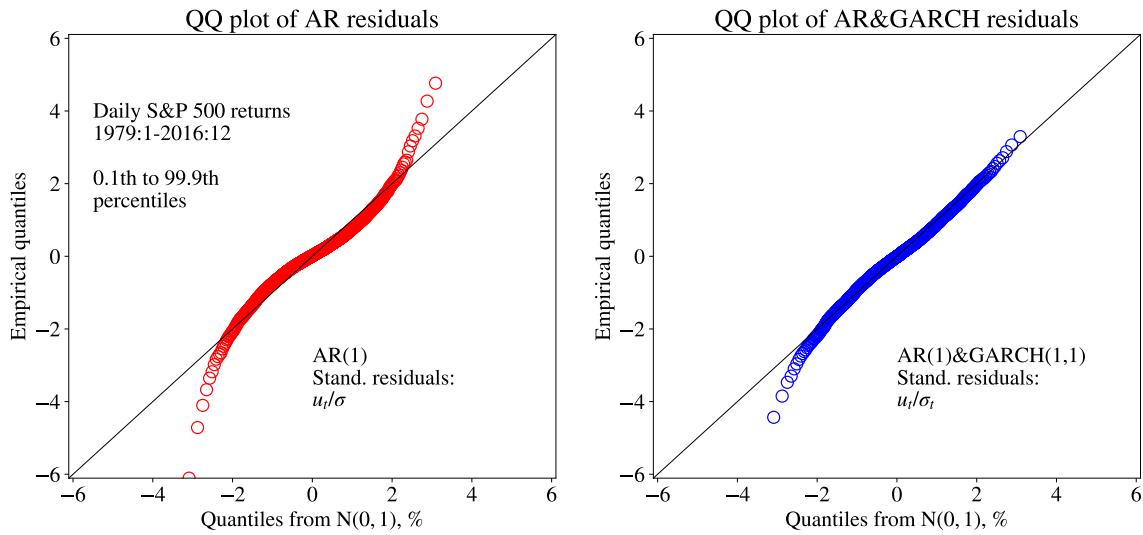


Figure 4.6: QQ-plot of residuals

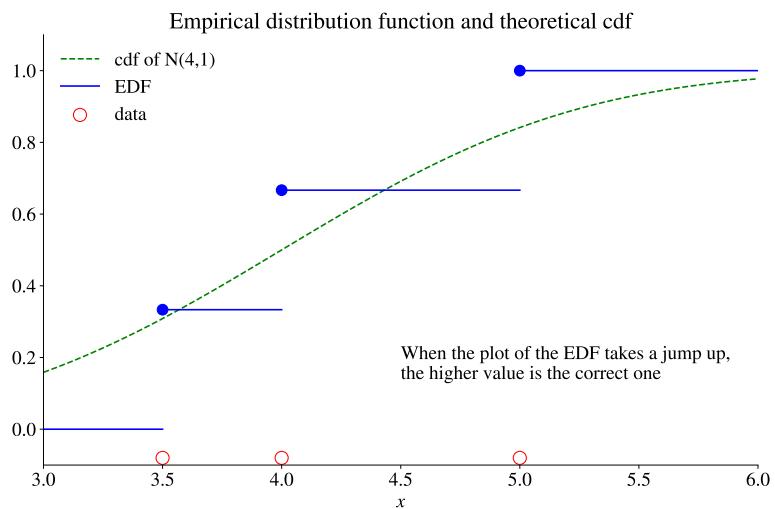


Figure 4.7: Example of empirical distribution function

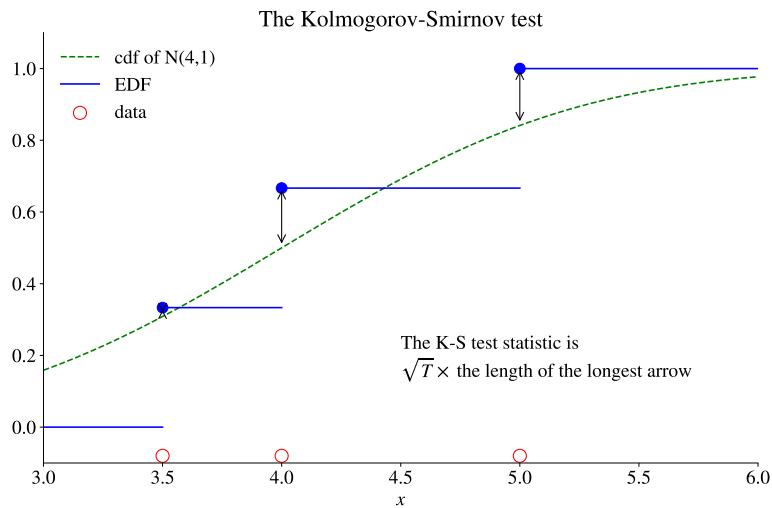


Figure 4.8: K-S test

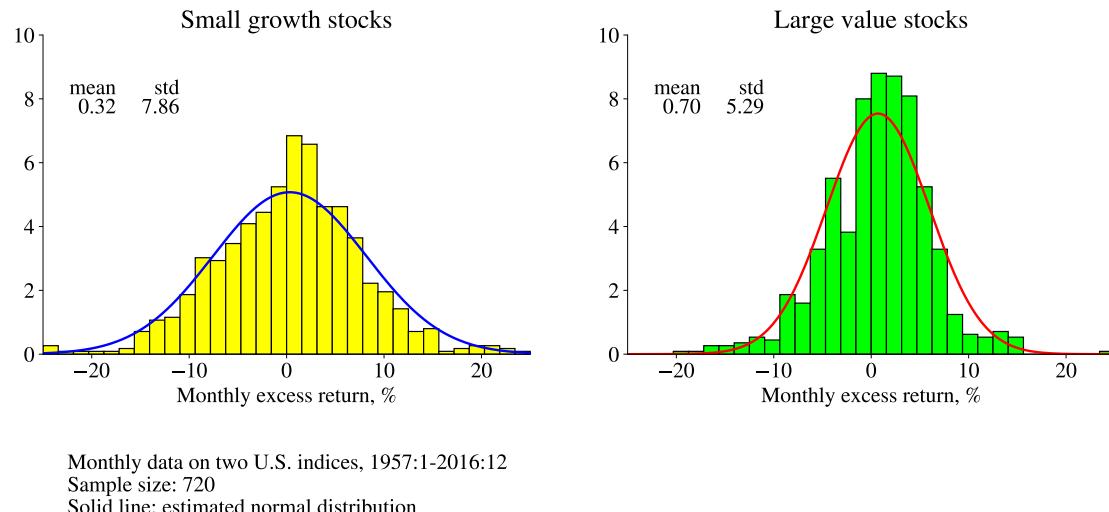


Figure 4.9: Histogram of returns, the curve is a normal distribution with the same mean and standard deviation as the return series

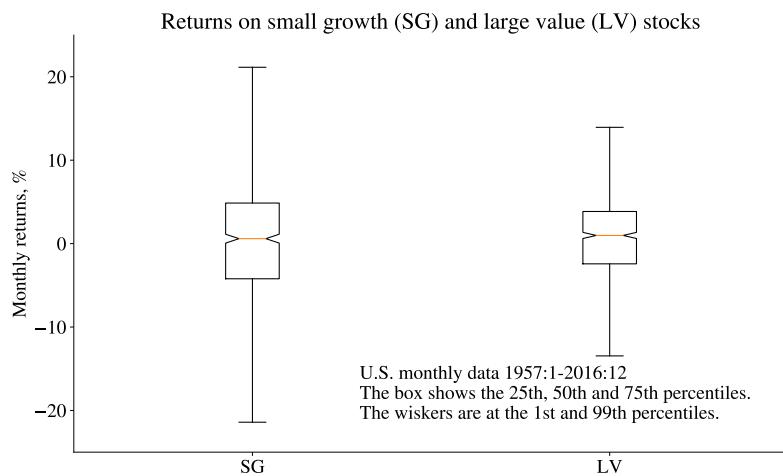


Figure 4.10: Box plot of returns

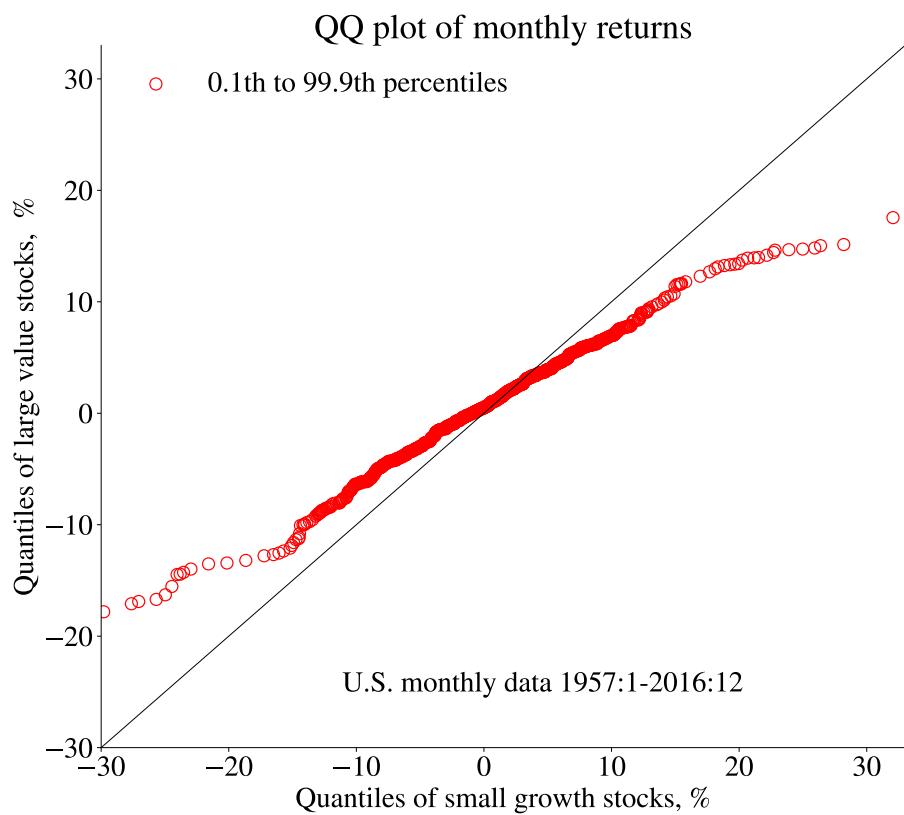


Figure 4.11: QQ-plot of returns, one-month U.S. equity returns

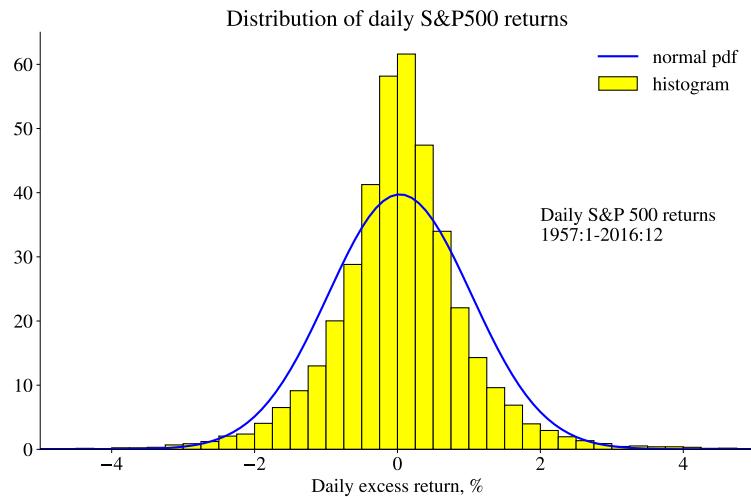


Figure 4.12: Histogram of returns and a fitted normal distribution

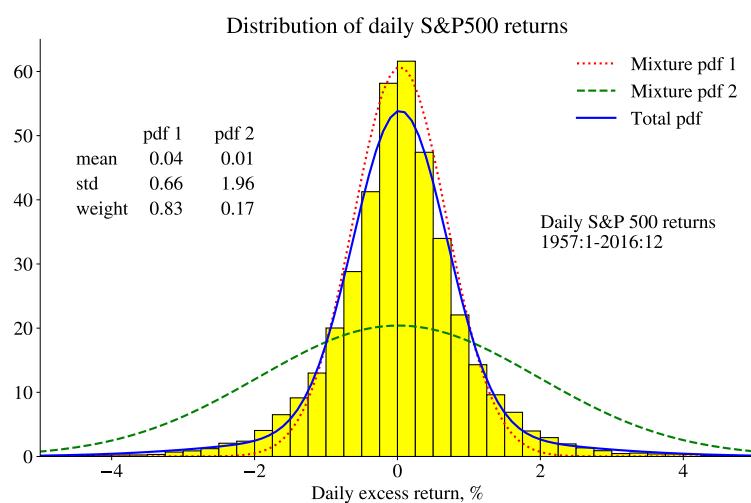


Figure 4.13: Histogram of returns and a fitted mixture normal distribution

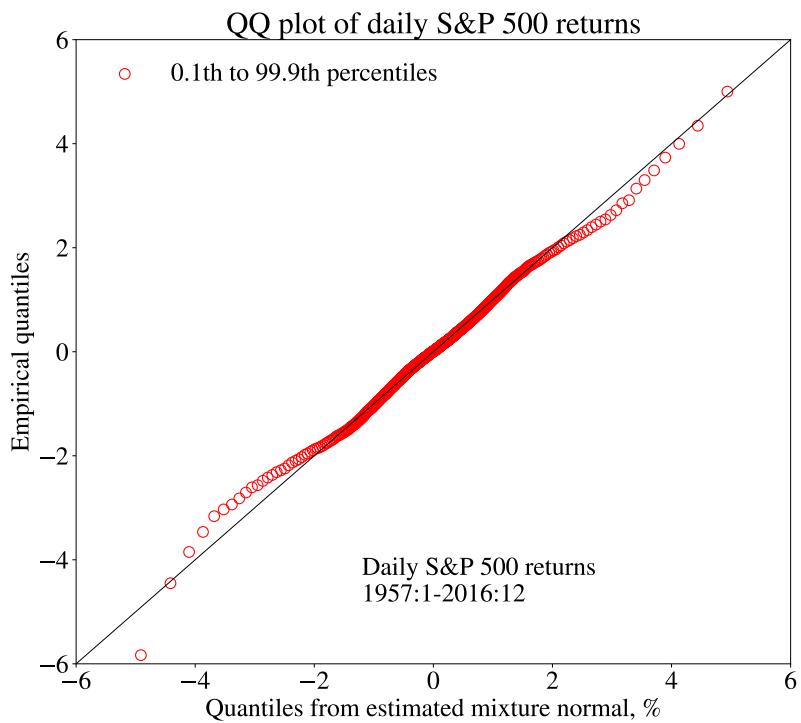


Figure 4.14: Quantiles of daily S&P returns

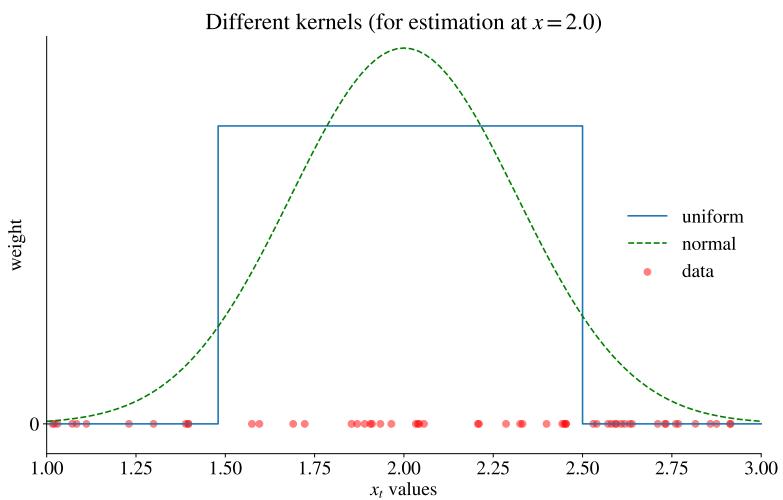


Figure 4.15: Different weighting functions for non-parametric regression

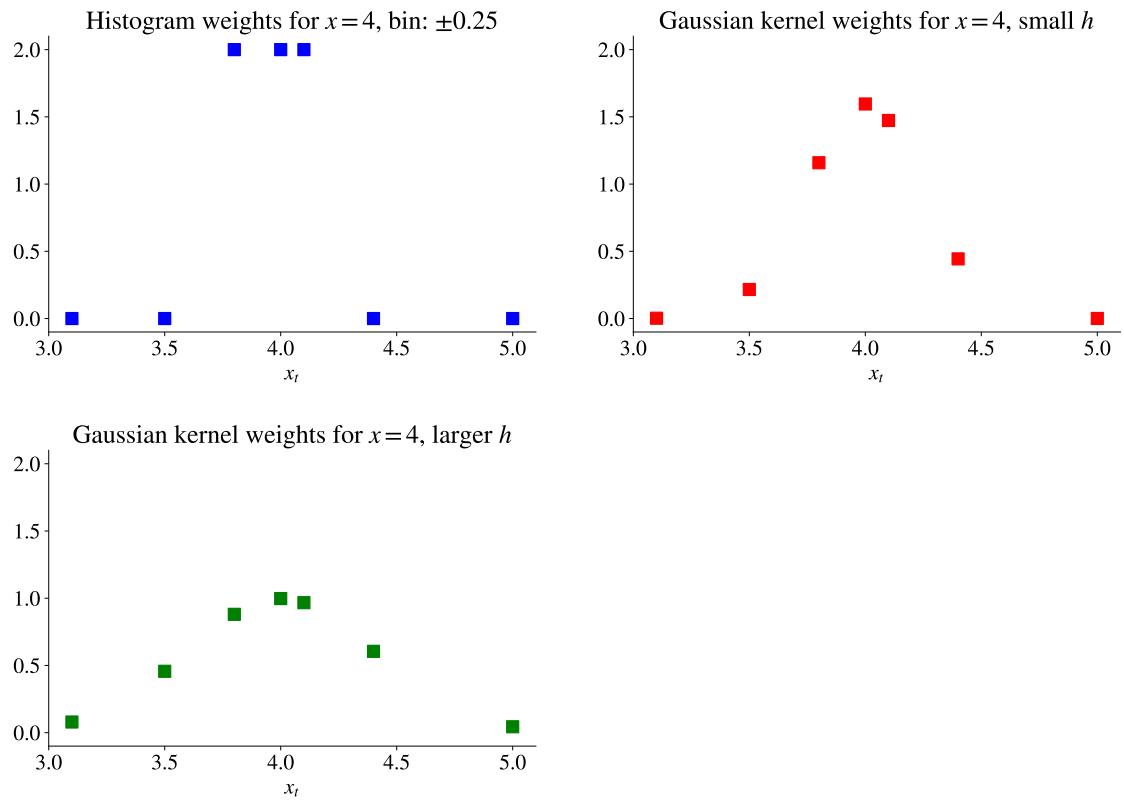


Figure 4.16: Calculation of the pdf at $x = 4$

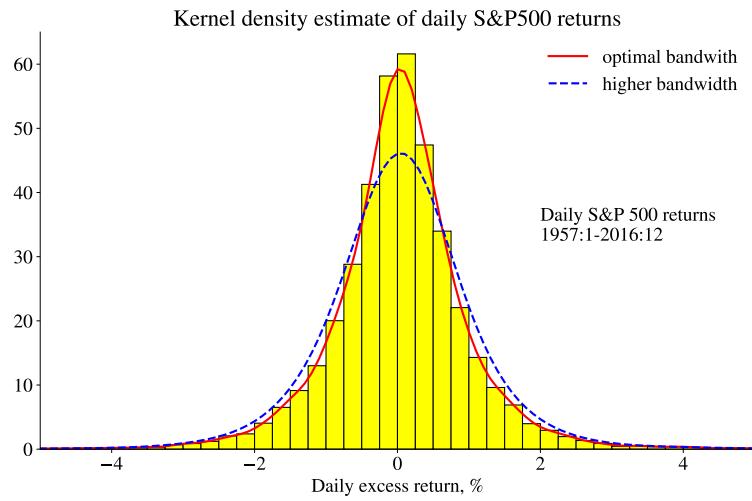


Figure 4.17: Stock returns

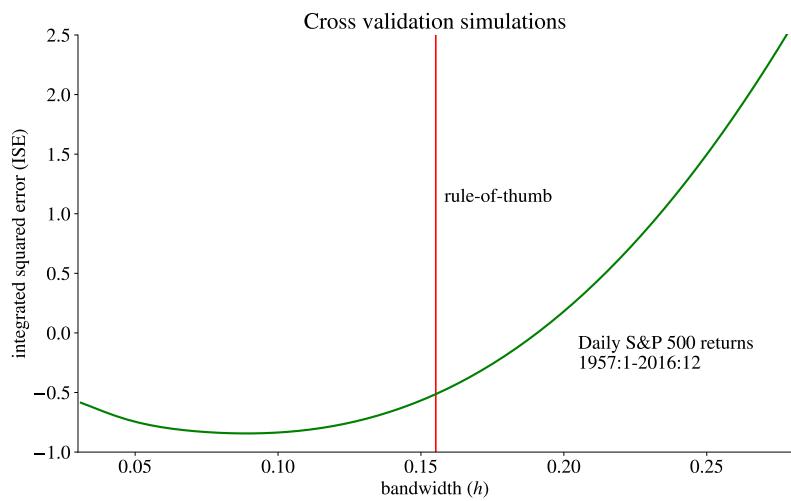


Figure 4.18: Cross-validation to determine the best band width

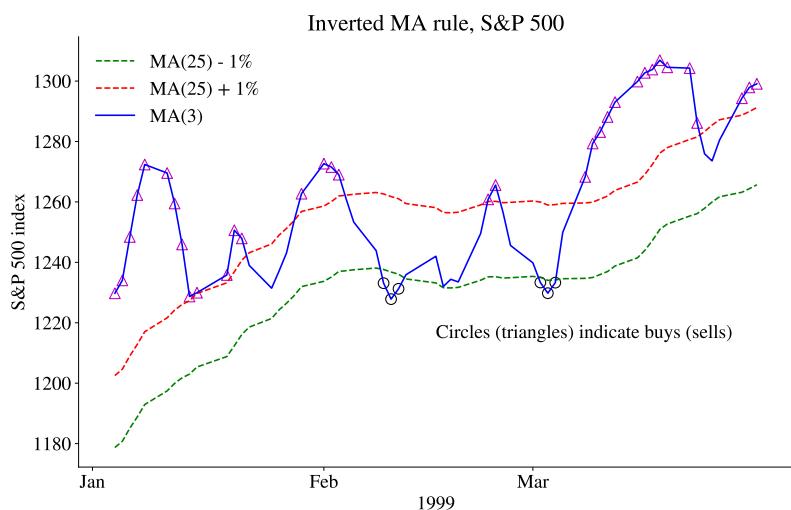


Figure 4.19: Examples of trading rules

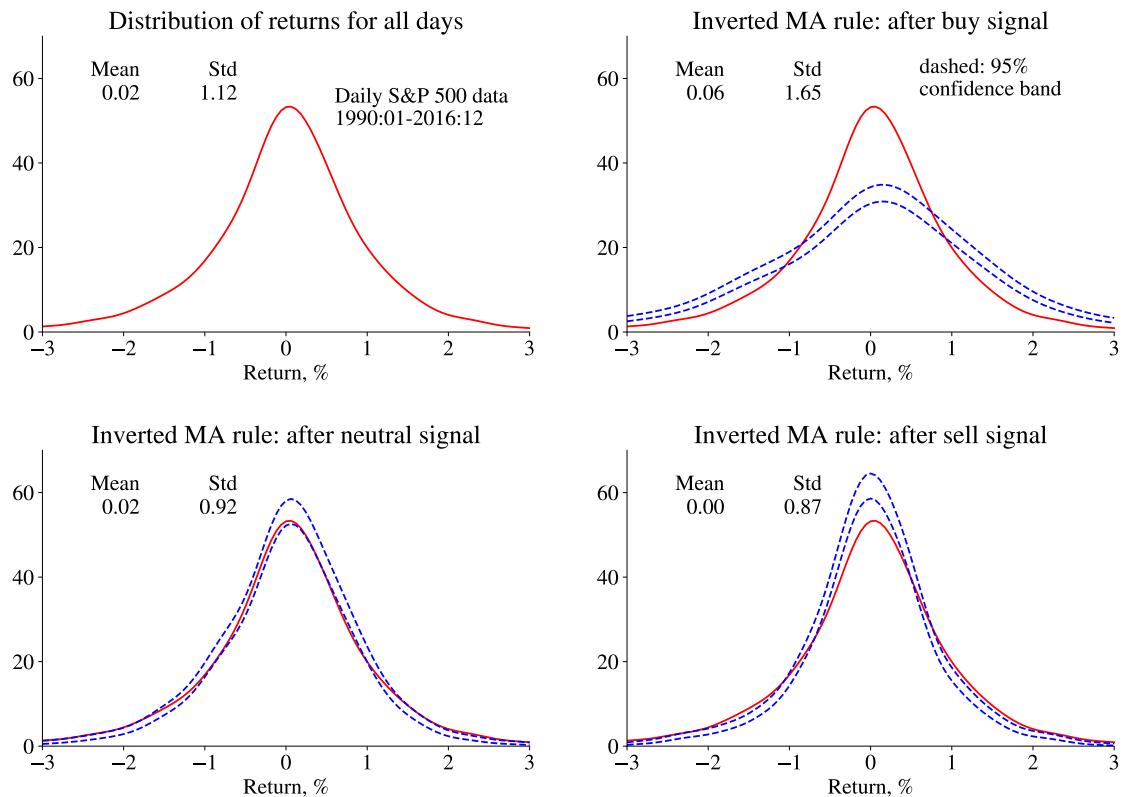


Figure 4.20: Distribution of returns after different trading signals

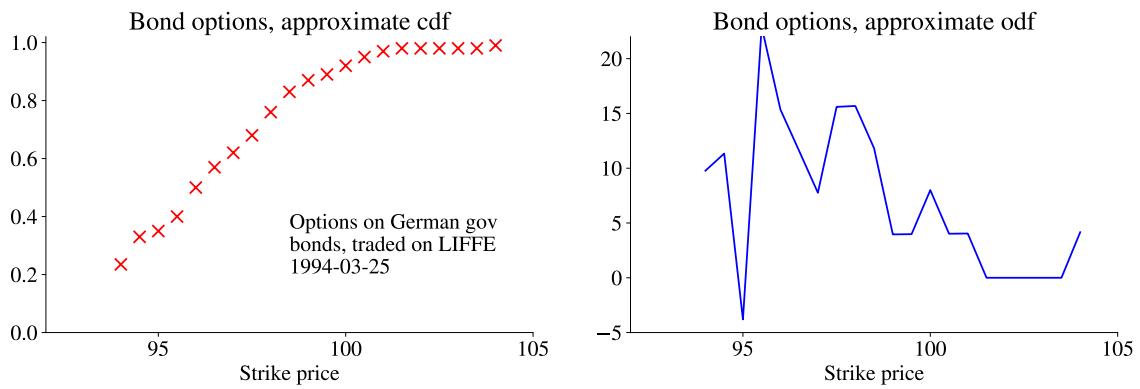


Figure 4.21: Bund options 6 April 1994. Options expiring in June 1994.

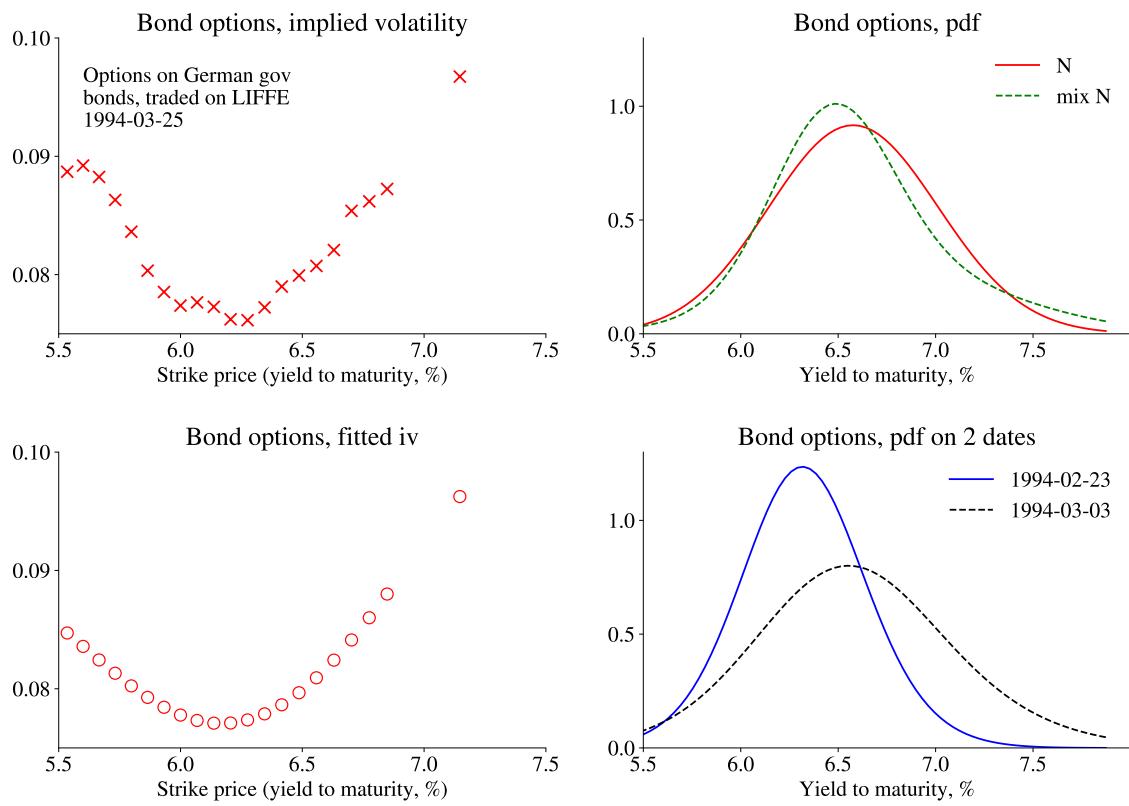


Figure 4.22: Bund options 23 February and 3 March 1994. Options expiring in June 1994.

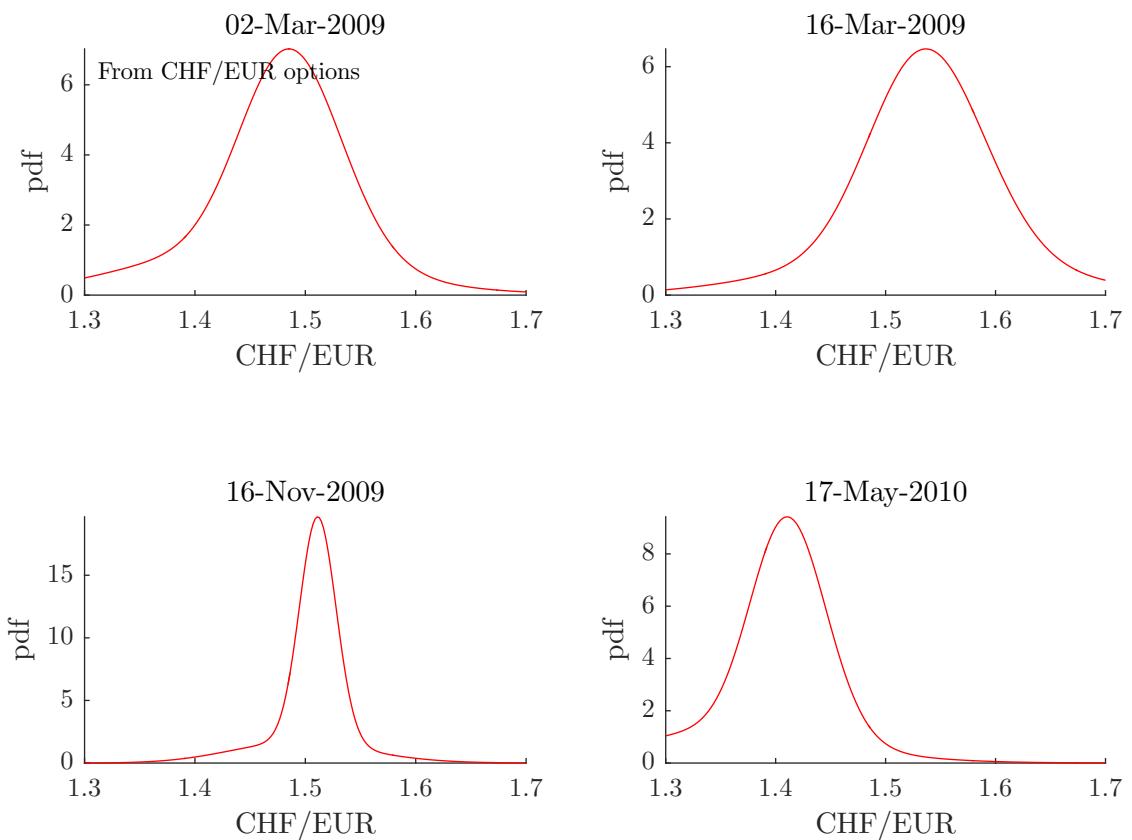


Figure 4.23: Riskneutral distribution of the CHF/EUR exchange rate

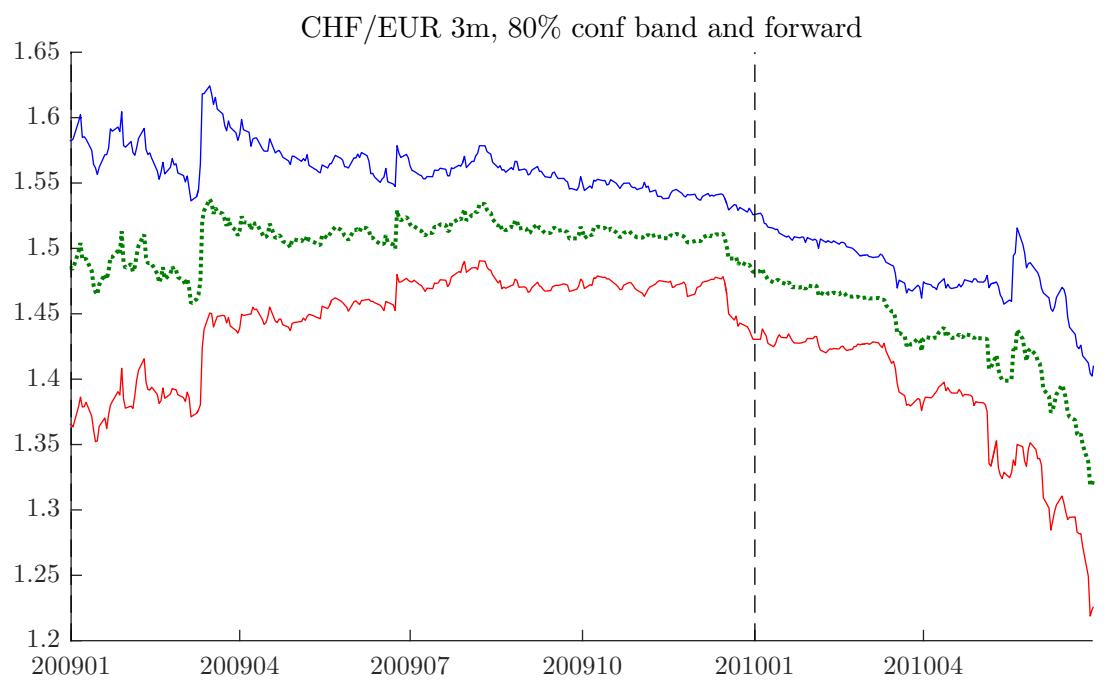


Figure 4.24: Riskneutral distribution of the CHF/EUR exchange rate

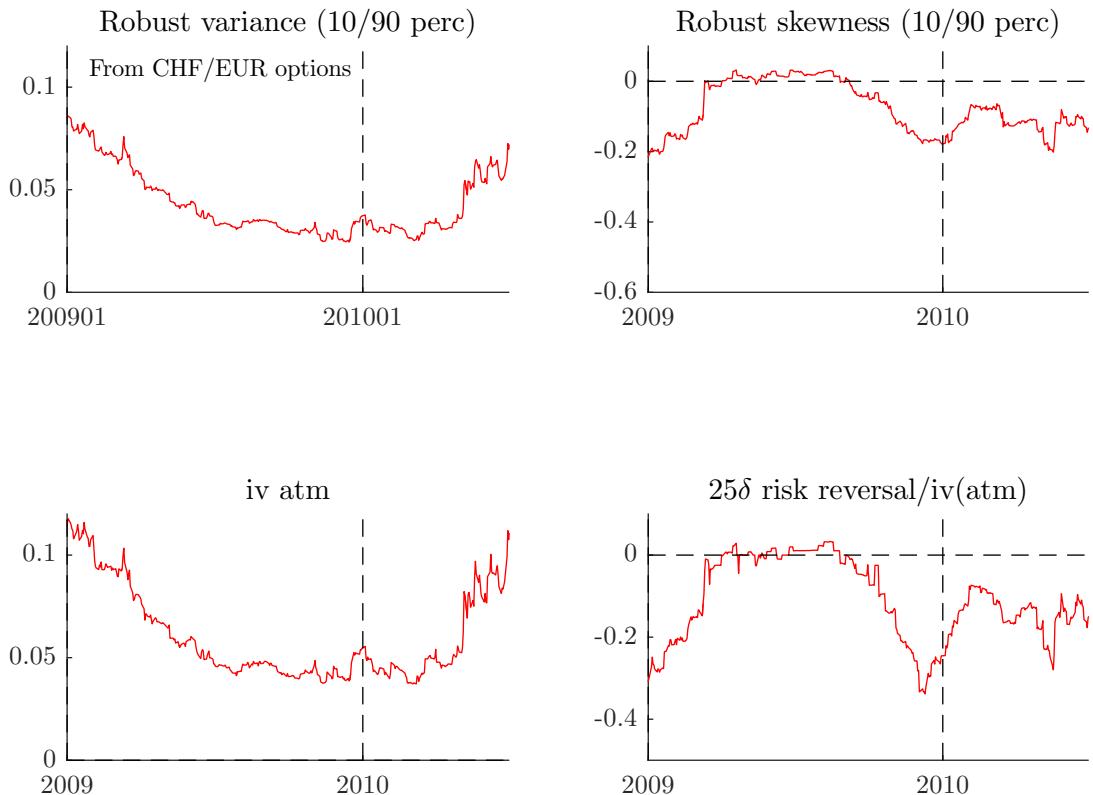


Figure 4.25: Riskneutral distribution of the CHF/EUR exchange rate

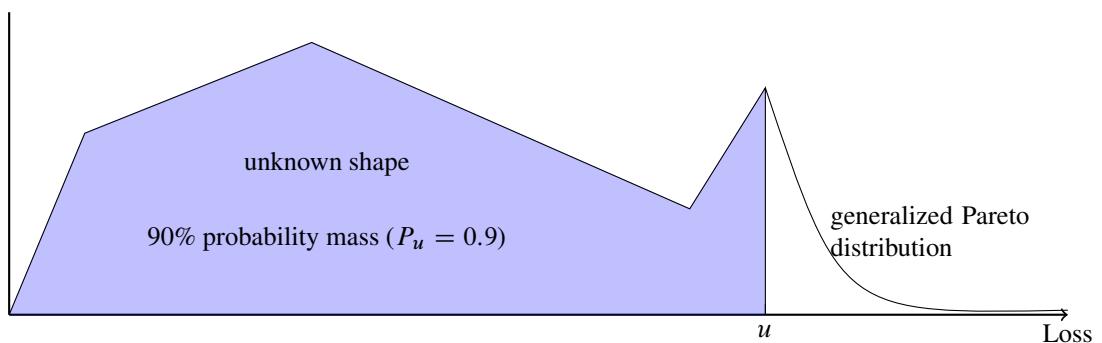


Figure 4.26: Loss distribution

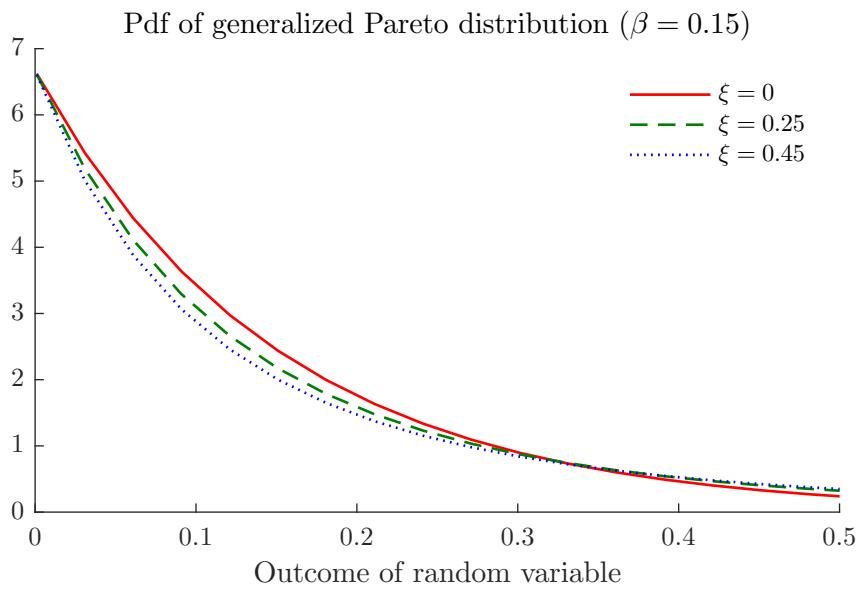


Figure 4.27: Generalized Pareto distributions

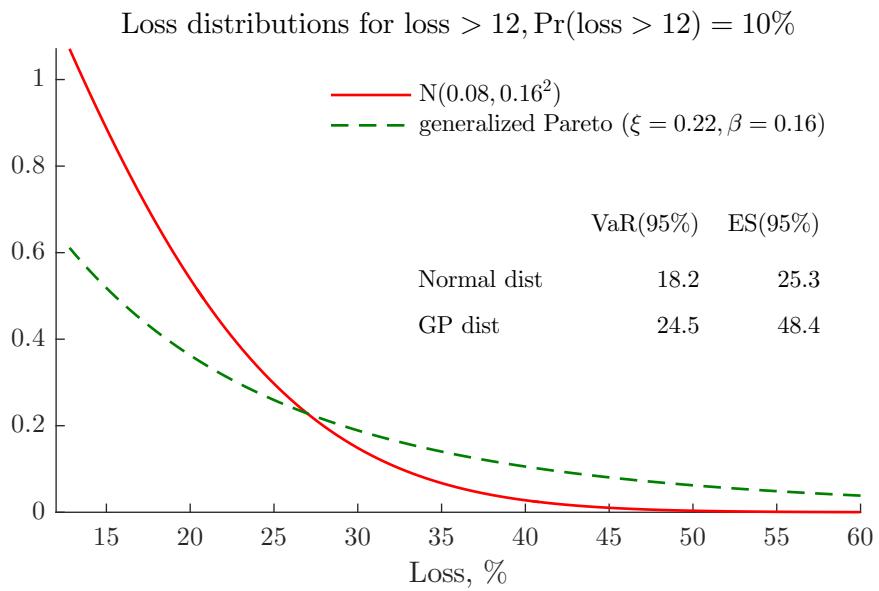


Figure 4.28: Comparison of a normal and a generalized Pareto distribution for the tail of losses

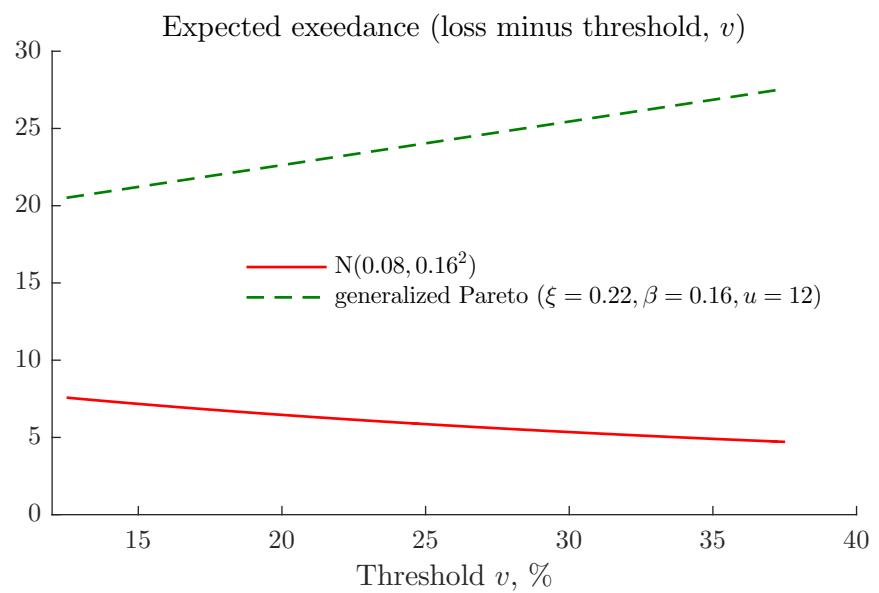


Figure 4.29: Expected exceedance, normal and generalized Pareto distribution

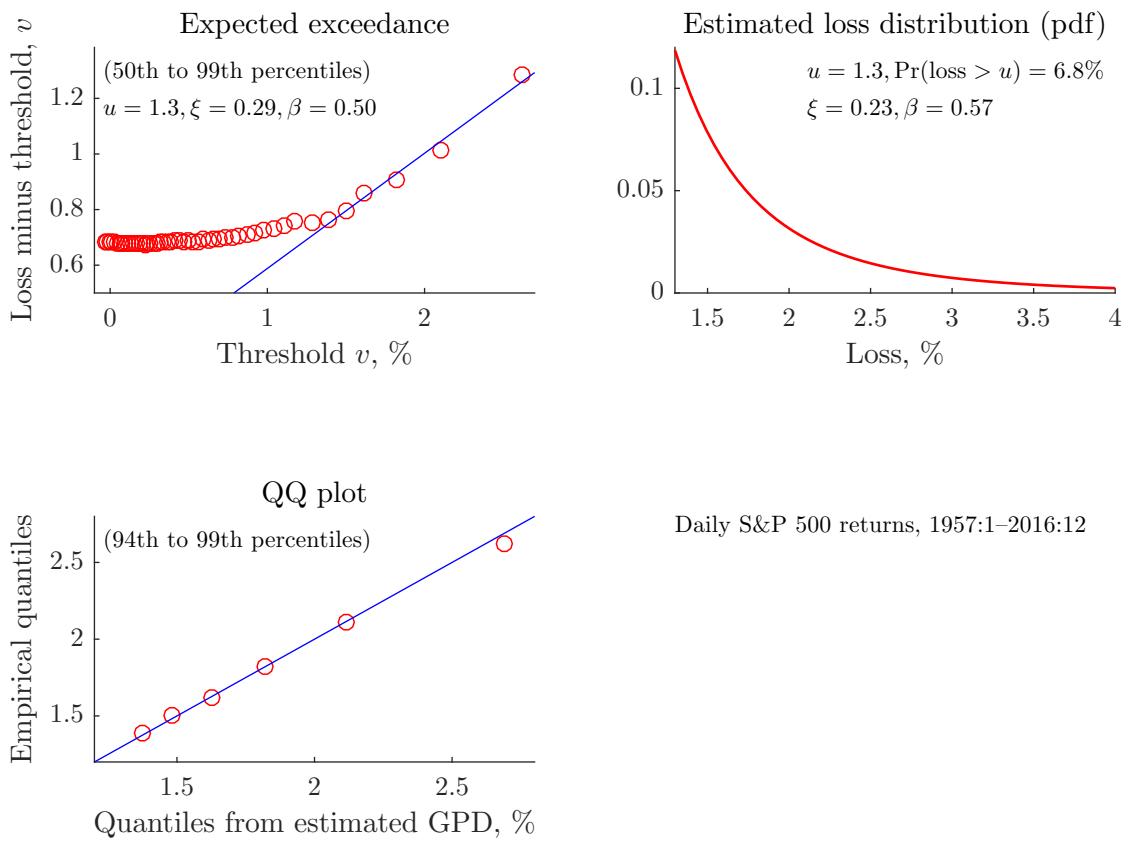


Figure 4.30: Results from S&P 500 data

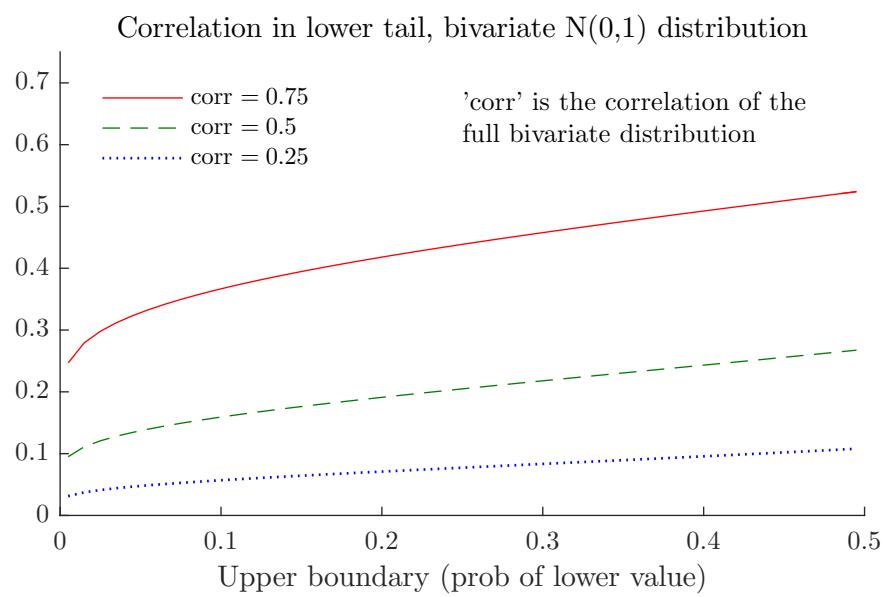


Figure 4.31: Correlation in lower tail when data is drawn from a normal distribution with correlation ρ

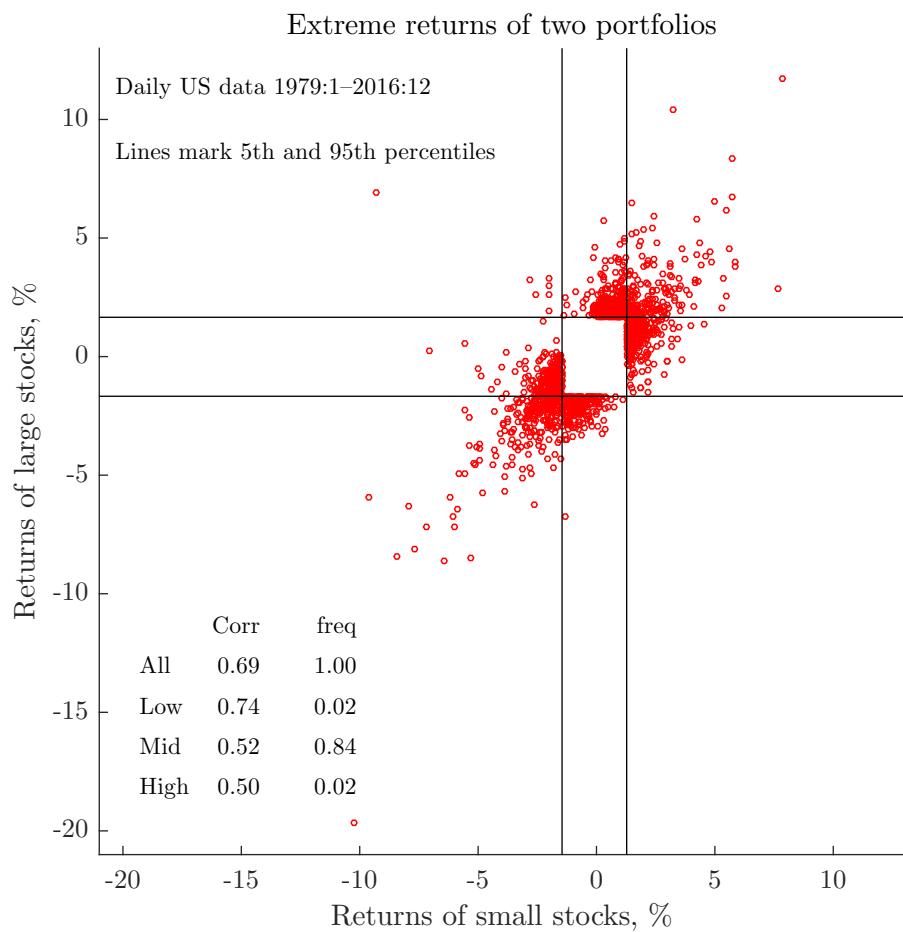


Figure 4.32: Correlation of two portfolios

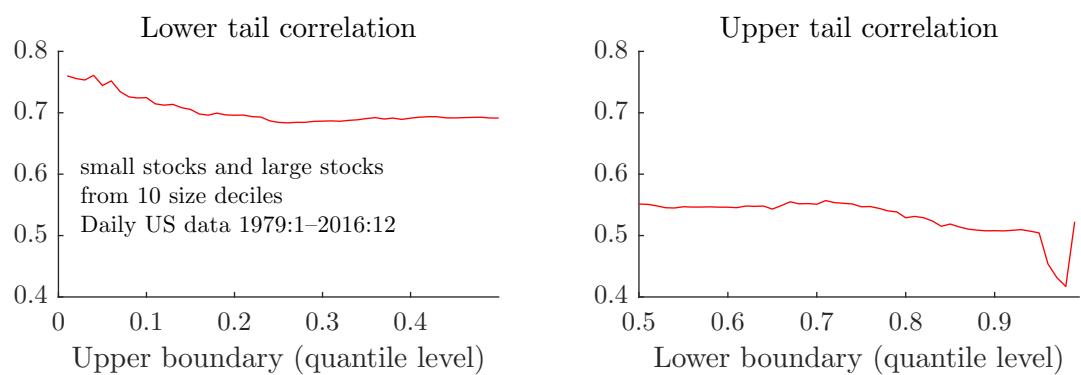


Figure 4.33: Correlation in the tails for two portfolios

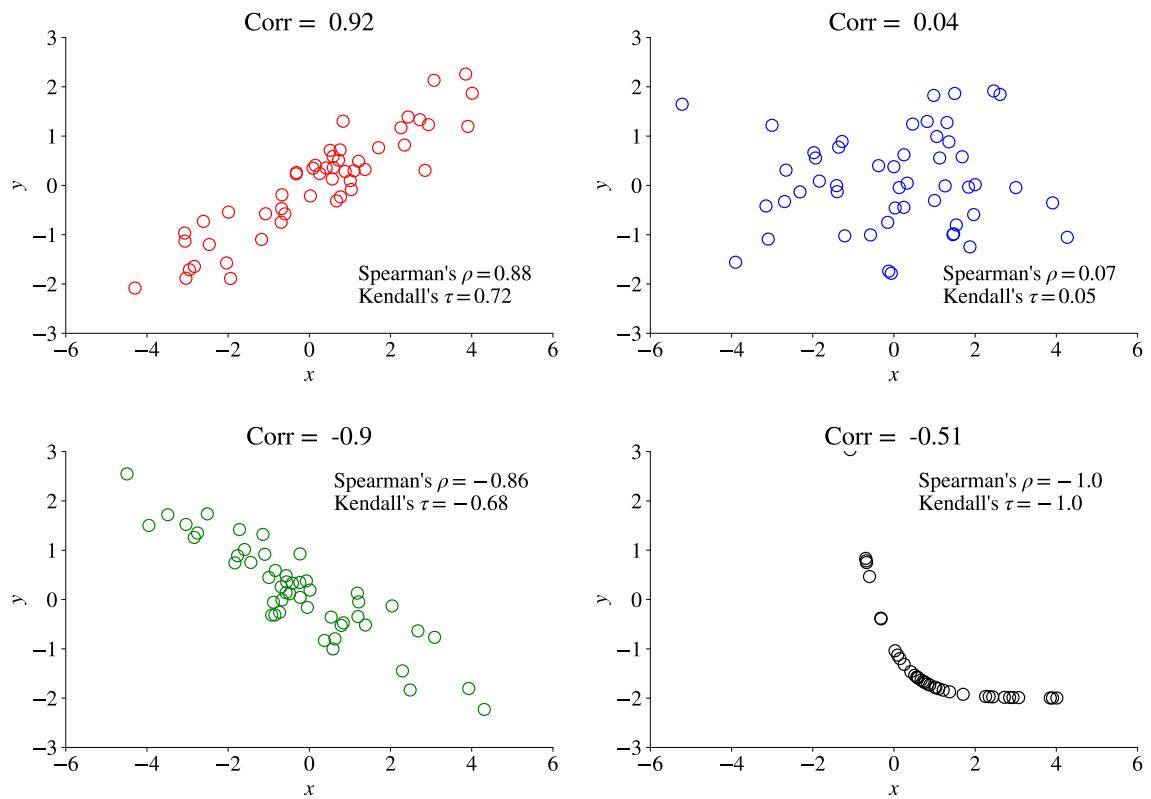


Figure 4.34: Illustration of correlation and rank correlation

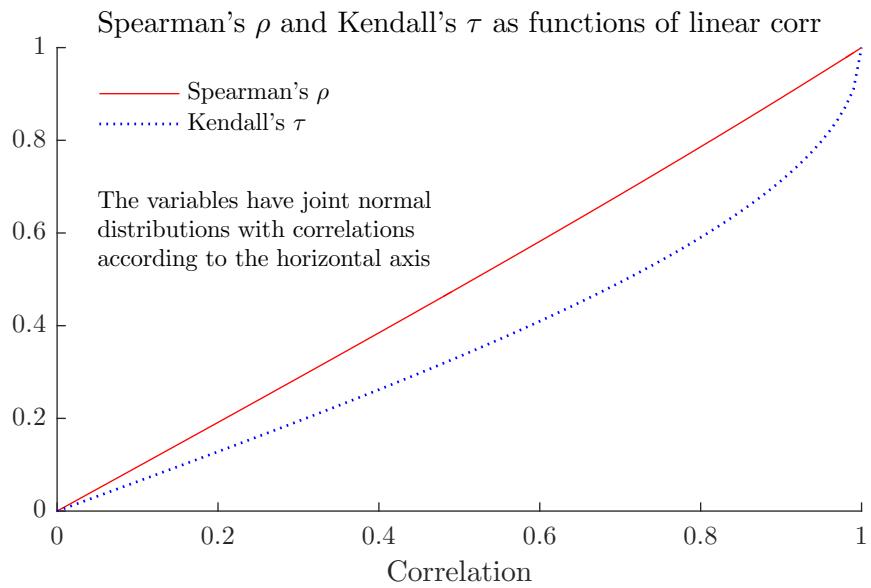


Figure 4.35: Spearman's rho and Kendall's tau if data has a bivariate normal distribution

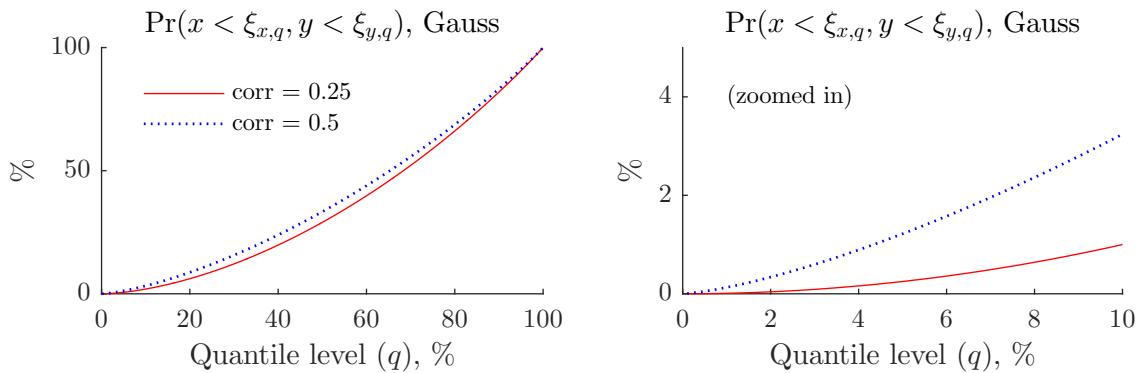


Figure 4.36: Probability of joint low returns, bivariate normal distribution

Bivariate normal distribution, pdf

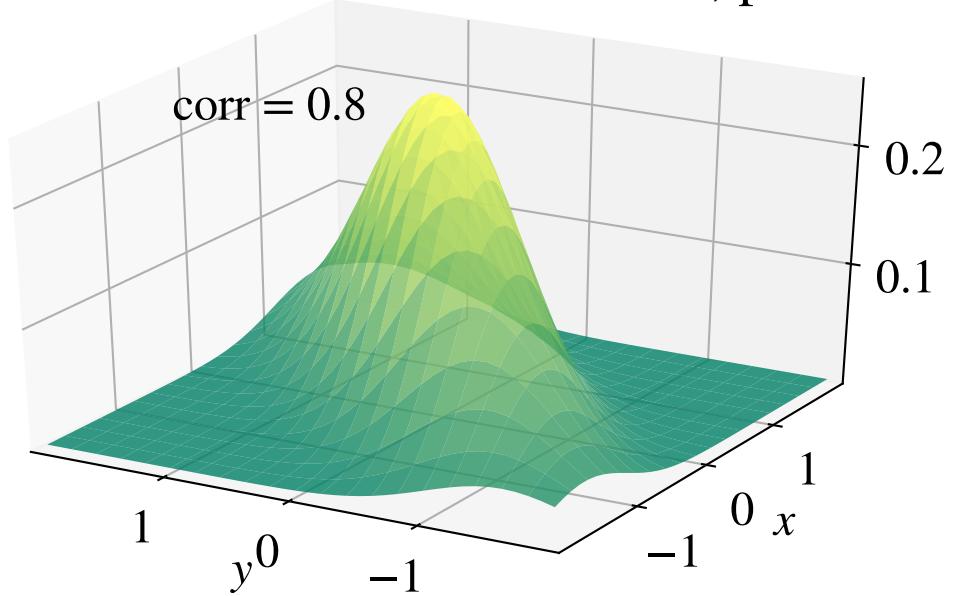


Figure 4.37: Bivariate normal distributions

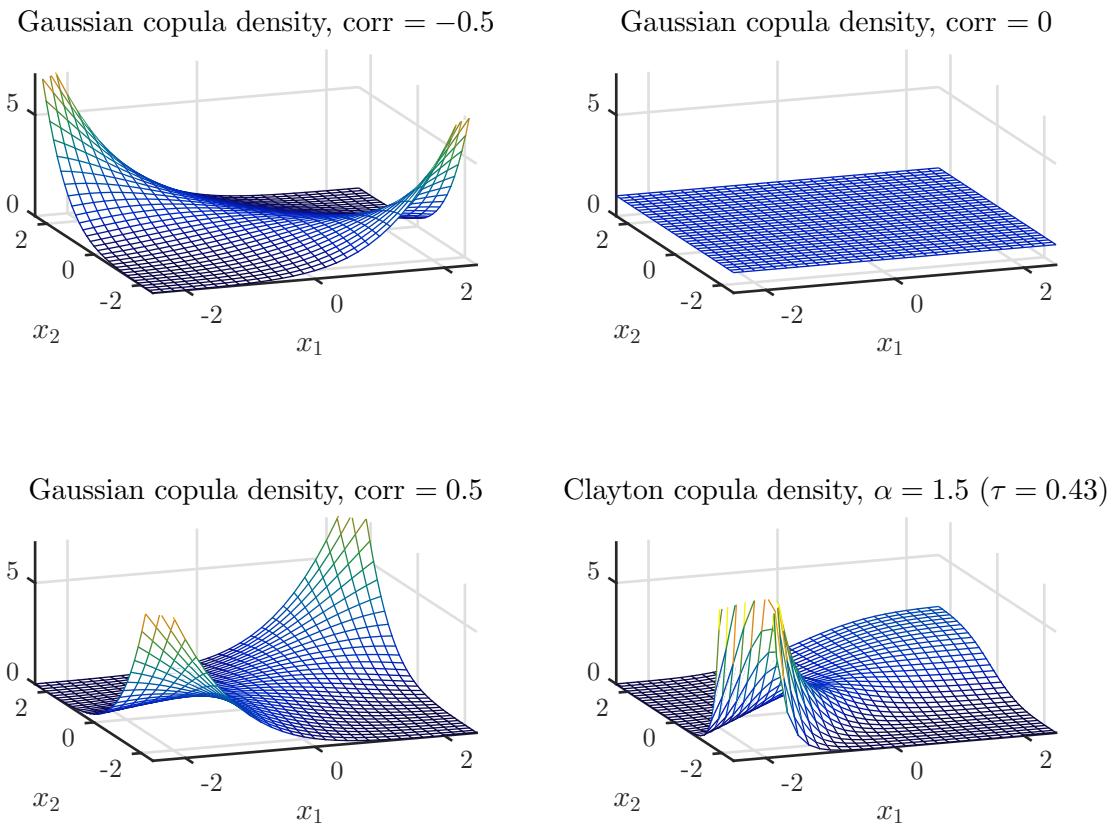
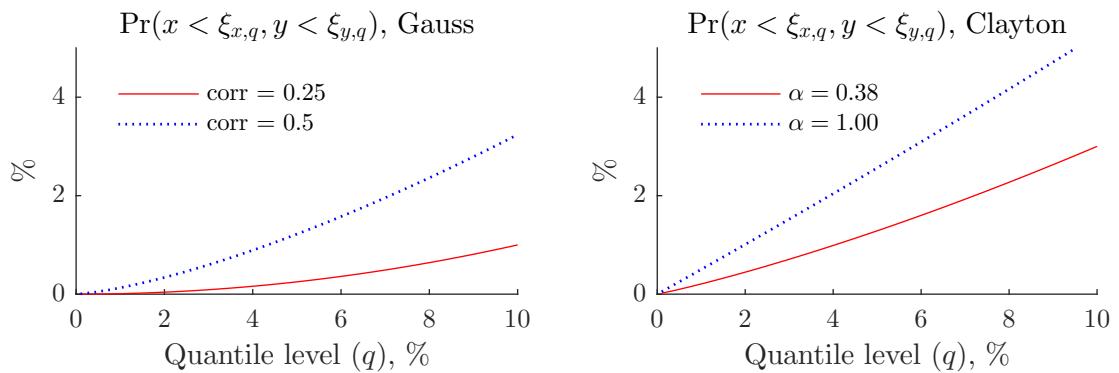


Figure 4.38: Copula densities (as functions of x_i)



The α values are calibrated to give correlations of 0.25 and 0.5.

Figure 4.39: Probability of joint low returns, Clayton copula

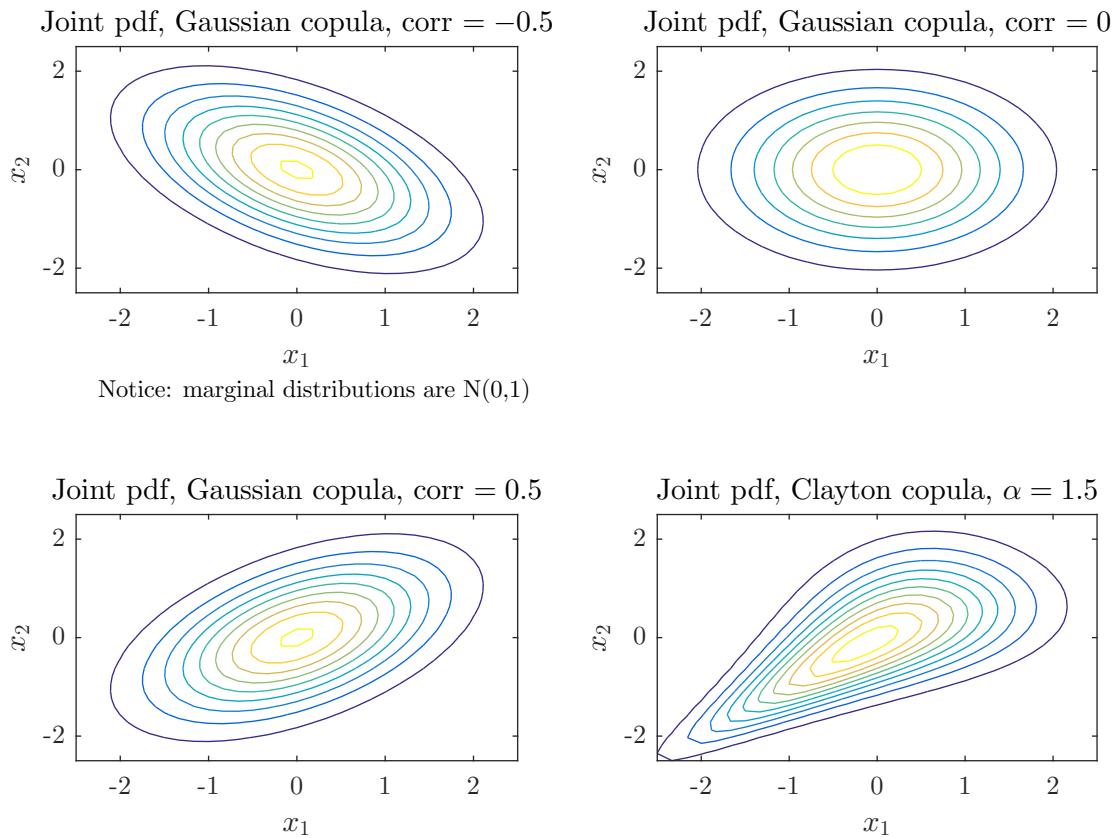


Figure 4.40: Contours of bivariate pdfs

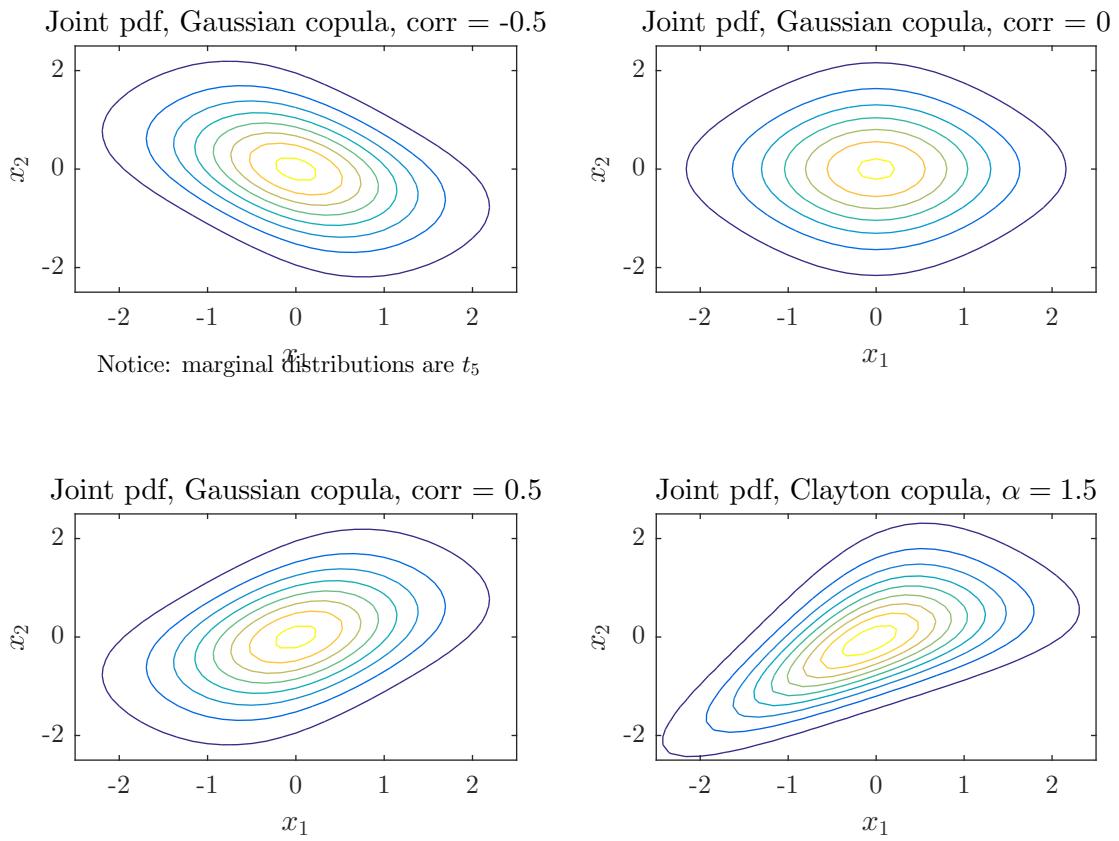


Figure 4.41: Contours of bivariate pdfs

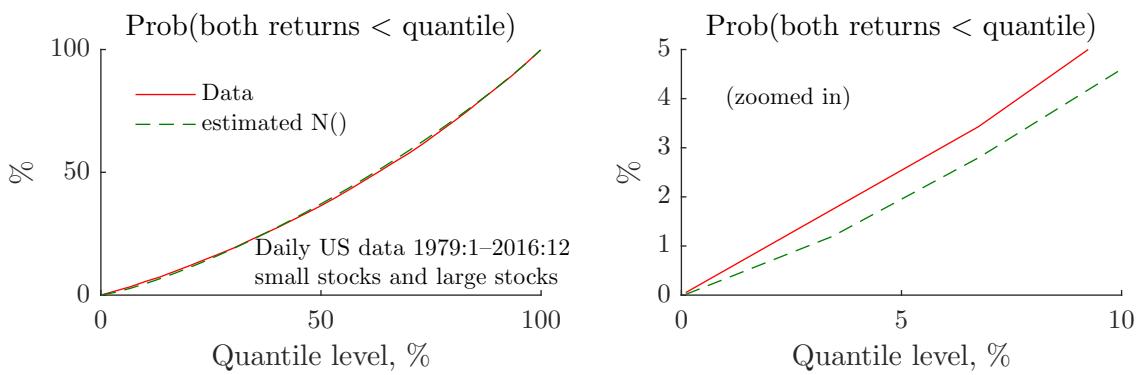


Figure 4.42: Probability of joint low returns

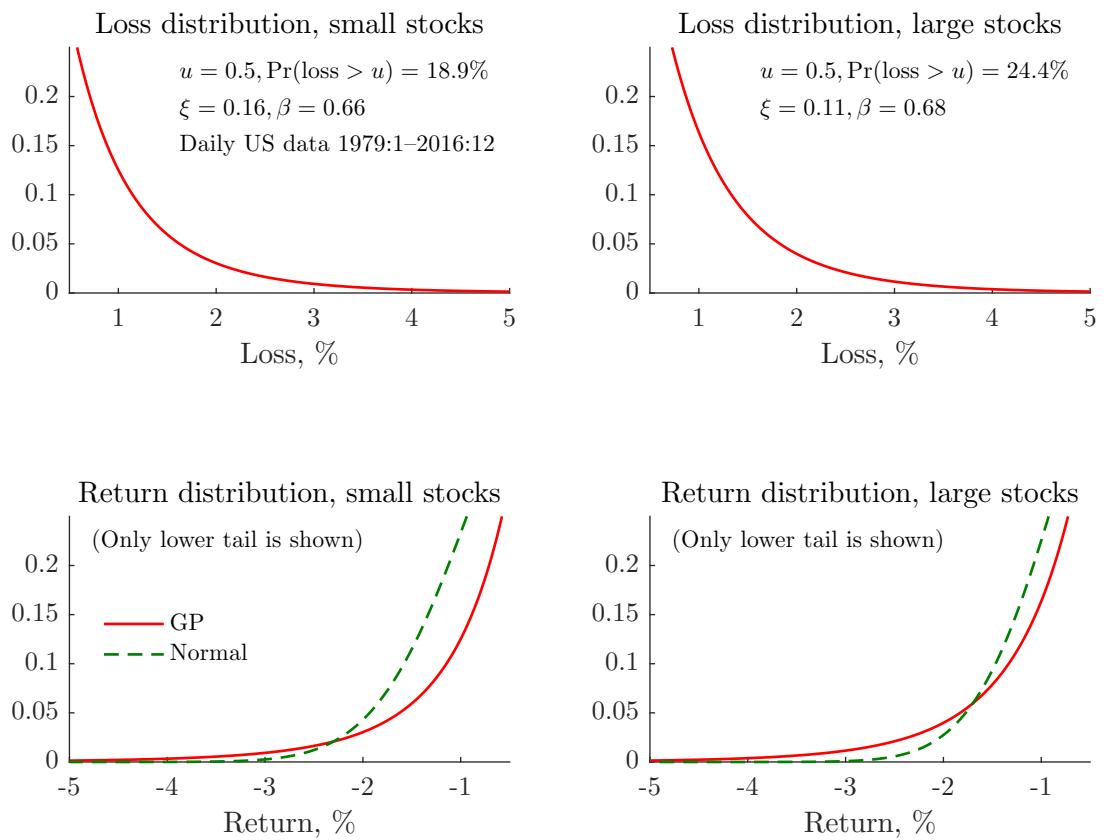


Figure 4.43: Estimation of marginal loss distributions

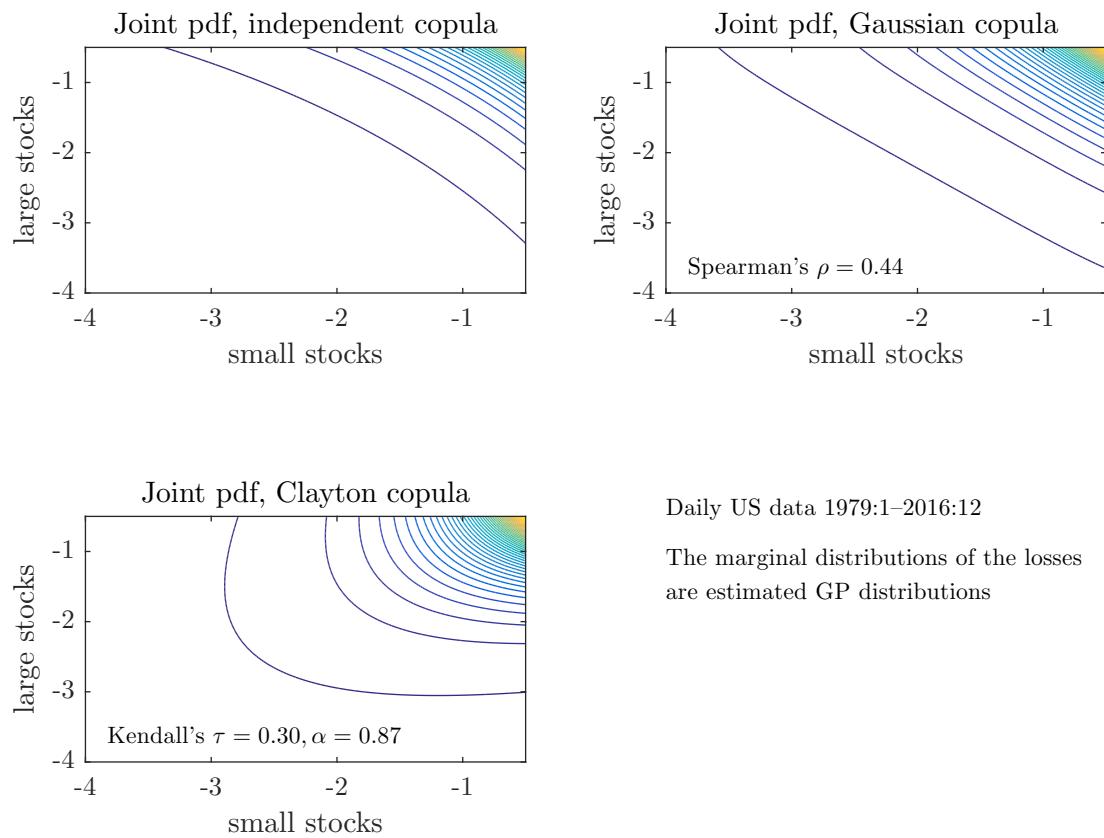


Figure 4.44: Joint pdfs with different copulas

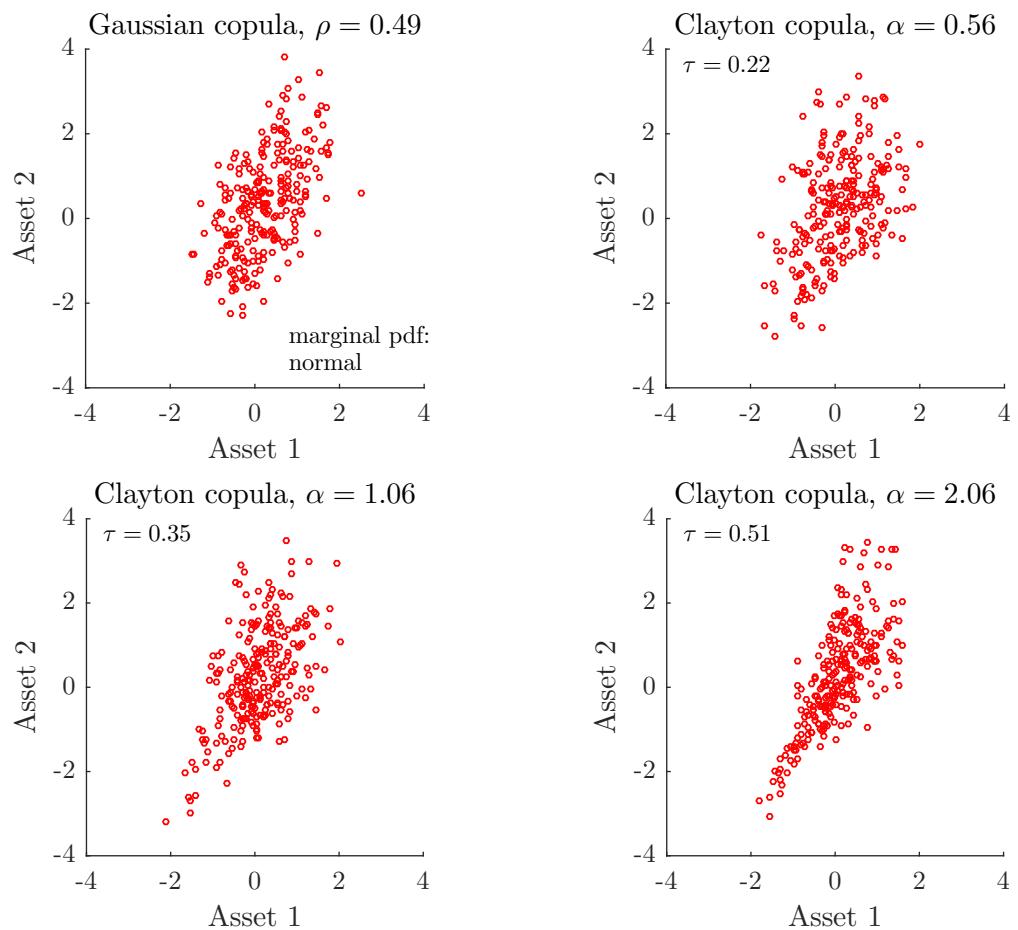


Figure 4.45: Example of scatter plots of two asset returns drawn from different copulas

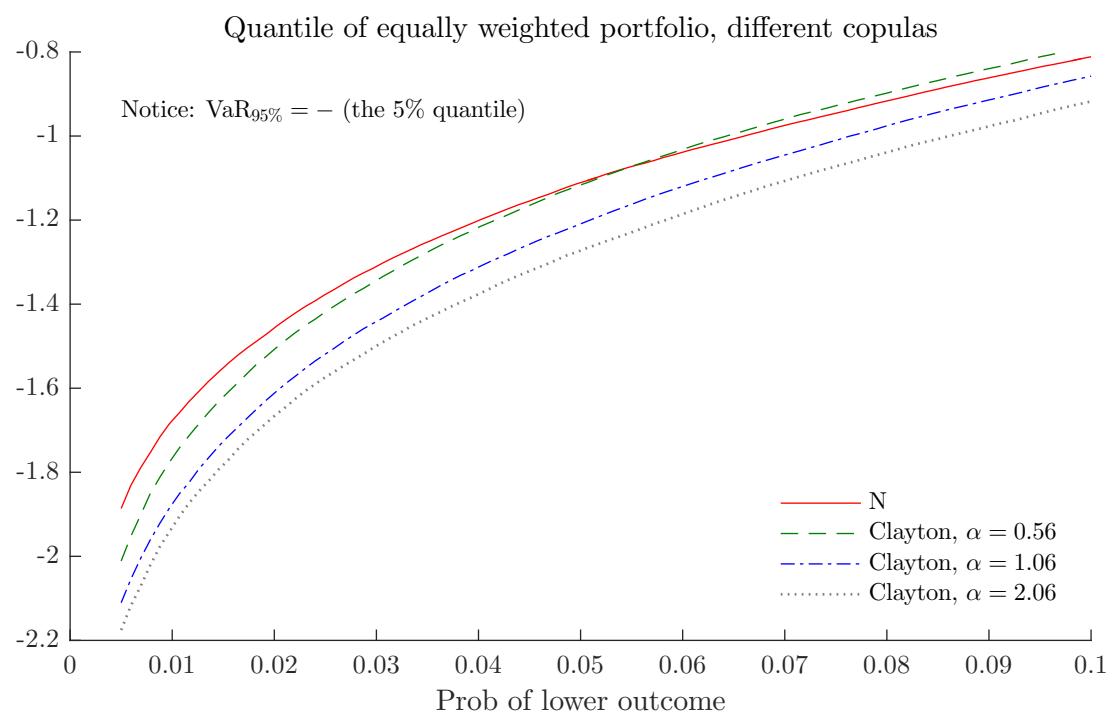


Figure 4.46: Quantiles of an equally weighted portfolio of two asset returns drawn from different copulas

Chapter 5

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; Elliot and Timmermann (2016)

5.1 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.

In contrast, if there are no important market imperfections, then predictability of excess returns should be thought of as predictable movements in risk premia. To see the latter, let R_{t+1}^e be the excess return on an asset. The canonical asset pricing equation says

$$\mathbb{E}_t m_{t+1} R_{t+1}^e = 0, \quad (5.1)$$

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

Remark 5.1 (*A consumption-based model*) Suppose we want to maximize the expected discounted sum of utility $\mathbb{E}_t \sum_{s=0}^{\infty} \beta^s u(c_{t+s})$. Let Q_t be the consumer price index in t . Then, we have

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

We can rewrite (5.1) (using $\text{Cov}(x, y) = \mathbb{E} xy - \mathbb{E} x \mathbb{E} y$) as

$$\mathbb{E}_t R_{t+1}^e = -\text{Cov}_t(m_{t+1}, R_{t+1}^e) / \mathbb{E}_t m_{t+1}. \quad (5.2)$$

This says that the expected excess return will vary if risk (the covariance) does. If we can model how these expected returns change over time, then we have a forecasting model for returns. (If the expectations are not too crazy, then the forecasting model may actually forecast future returns...)

Example 5.2 (*Epstein-Zin utility function*) *Epstein and Zin (1991)* define a certainty equivalent of future utility as $Z_t = [\mathbb{E}_t(U_{t+1}^{1-\gamma})]^{1/(1-\gamma)}$ 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,

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

(See *Söderlind (2006)* for a simple proof.) The point is that without predictability, the Epstein-Zin utility function has the same implications as the CRRA utility function. Establishing whether there is predictability is therefore a way to assess the importance of the theory.

Example 5.3 (*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, $\mathbb{E}_t(r_{t+1} - r_f) = x_t$. $\text{Cov}_t(u_{t+1}, \eta_{t+1})$ can be non-zero. For instance, with $\text{Cov}_t(u_{t+1}, \eta_{t+1}) < 0$, 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 $v_t = 1/(2\gamma) + x_t/[\gamma \text{Var}_t(u_{t+1})]$, so predictability does still not matter. Third, with a higher risk aversion and $\text{Cov}_t(u_{t+1}, \eta_{t+1}) < 0$, there is a positive hedging demand for the risky asset: it pays off (today) when the future investment opportunities are poor.

Example 5.4 (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 parameterised 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 5.5 (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. In short, mean reversion in the price level means negative autocorrelation of the returns—and vice versa. See Figure 5.1 for the impulse responses with respect to a piece of news, ε_t .

5.2 Autocorrelations

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

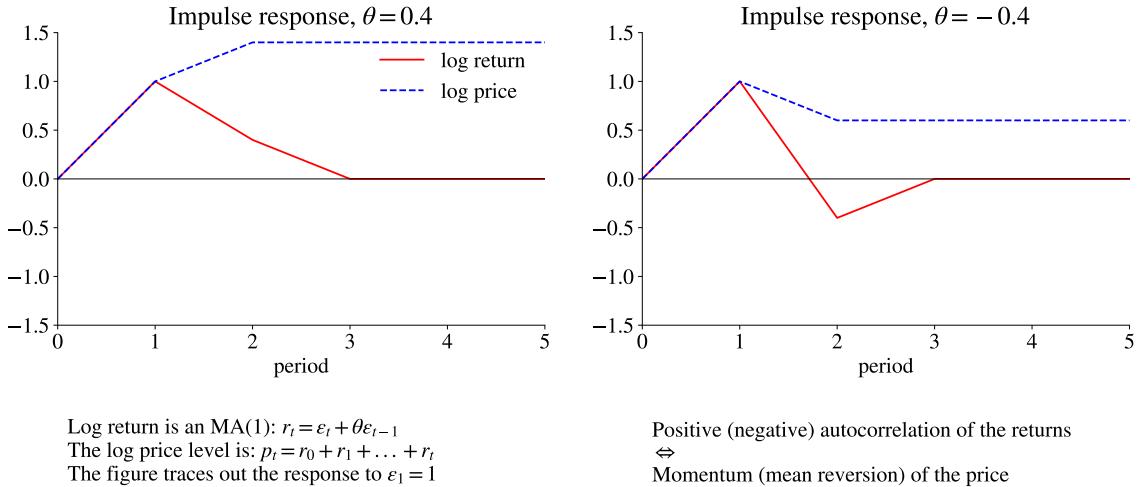


Figure 5.1: Impulse responses when price is random walk plus temporary component

5.2.1 Autocorrelation Coefficients and the Box-Pierce Test

The sampling properties of autocorrelations ($\hat{\rho}_s$) 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 we have

$$\sqrt{T} \begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right), \quad (5.3)$$

provided $(i, j) \neq 0$ and $i \neq j$. 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). To apply this on returns, the return horizon can be whatever (seconds, years,...), but it is important that the returns are non-overlapping (time aggregation can easily introduce spurious serial correlation).

Example 5.6 (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 -1.64 and another 5% above 1.64 , we can reject the null hypothesis at the 10% level if $\sqrt{T}|\hat{\rho}_1| > 1.64$. With $T = 100$, we therefore need $|\hat{\rho}_1| > 1.64/\sqrt{100} = 0.165$ for rejection, and with $T = 1000$ we need $|\hat{\rho}_1| > 1.64/\sqrt{1000} \approx 0.053$.

The *Box-Pierce test* follows directly from the result in (5.3), since it shows that $\sqrt{T}\hat{\rho}_i$

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

$$Q_L = T \sum_{s=1}^L \hat{\rho}_s^2 \xrightarrow{d} \chi_L^2. \quad (5.4)$$

Example 5.7 (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 (5.4), 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). See Figure 5.2 for an empirical illustrations.

The main problem with these tests is that the assumptions behind the results in (5.3) 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, a non-parametric test like the “runs test” can be used.

Remark 5.8 (Runs test*) A “runs test” is a non-parametric test of randomness. Let d_t be an indicator variable

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

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

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

(Warning: r indicates “runs,” not returns.) It is straightforward (but tedious) to show that, under the null hypothesis of randomness,

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

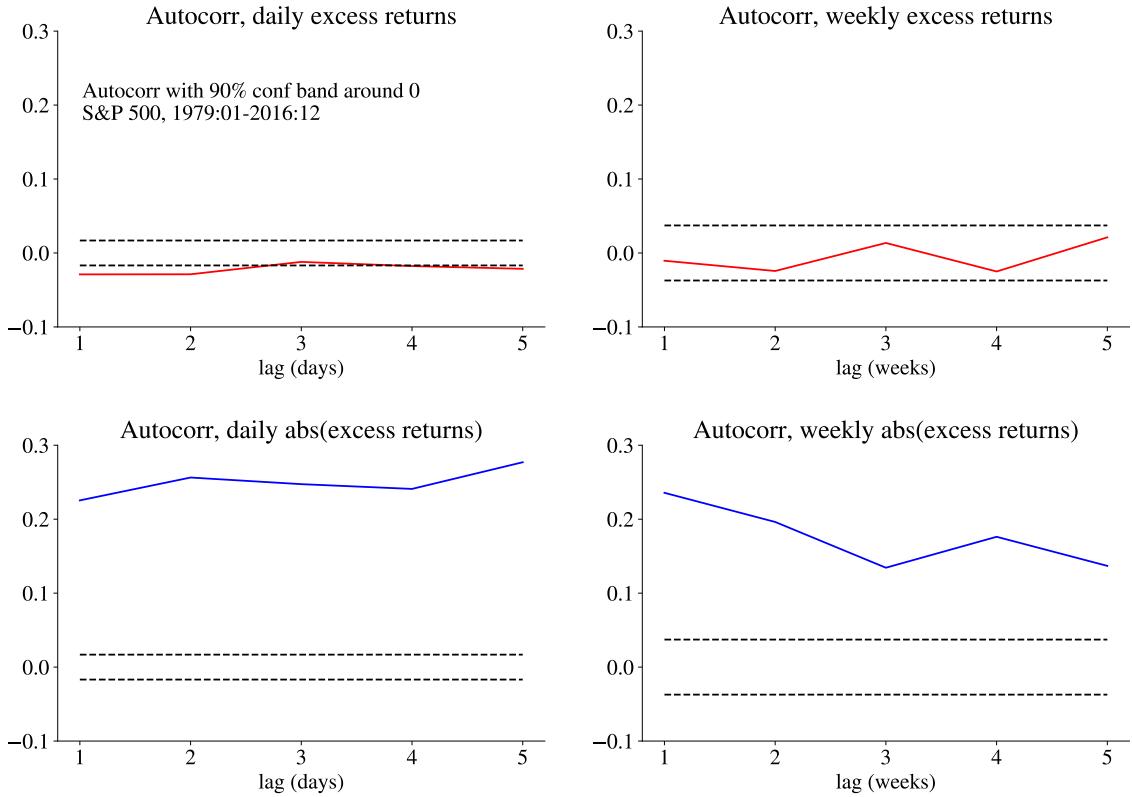


Figure 5.2: Predictability of US stock returns

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

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

The basic intuition of the test is that a positive autocorrelation would lead to too few runs ($r < \mathbb{E} r$): the y_t variable would stay on one side of the threshold q for long spells of time—and hence there would be few changes in x_t . Negative autocorrelation is just the opposite, since it tends to give a zigzag pattern around the mean. See Figure 5.4 for an example.

5.2.2 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 higher-order autocorrelation.

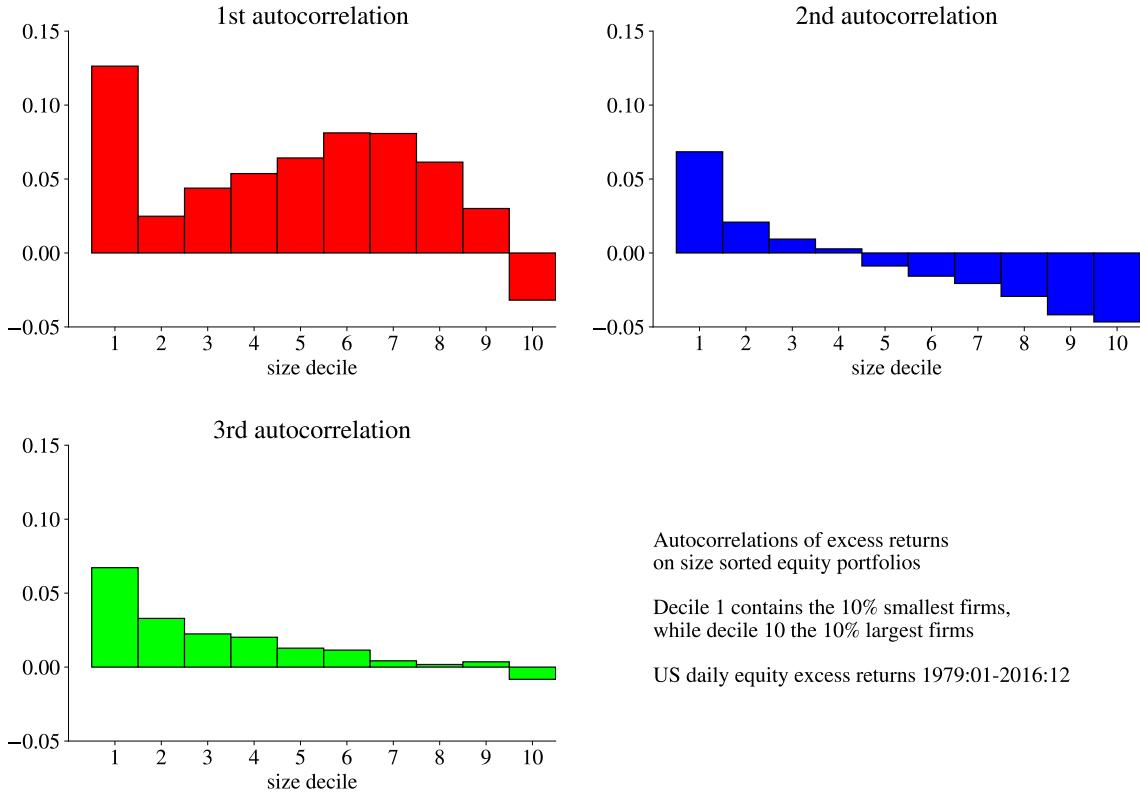


Figure 5.3: Predictability of US stock returns, size deciles

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

$$g_t(\beta) = \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ so } \bar{g}(\beta) = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ with } \beta = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}. \quad (5.5)$$

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.

Remark 5.9 (*Box-Pierce as an Application of GMM*) (5.5) is an exactly identified system

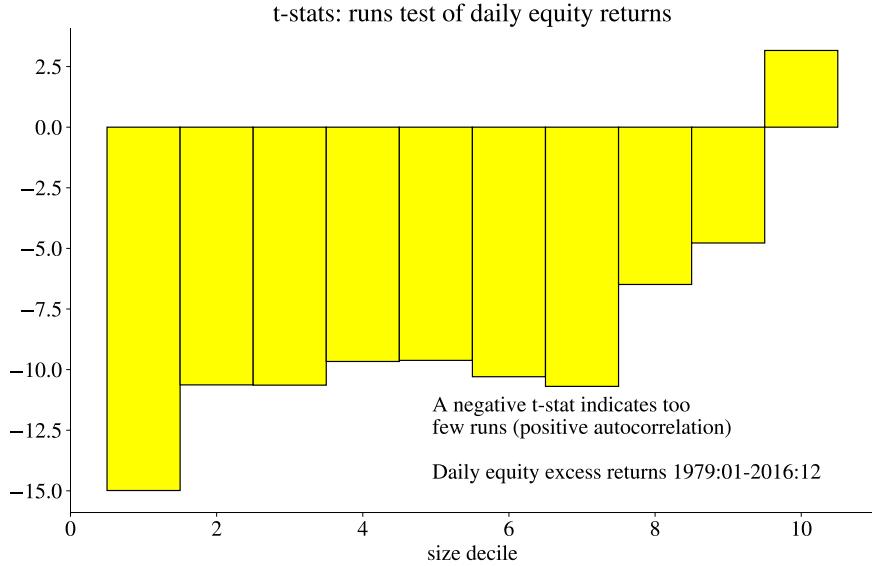


Figure 5.4: Runs test

so the weight matrix does not matter, so the asymptotic distribution is

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

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

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

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

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

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

$$S_0 = E g_t(\beta_0) g_t(\beta_0)'.$$

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

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

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

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

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

5.2.3 Autoregressions

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

$$R_t = c + a_1 R_{t-1} + a_2 R_{t-2} + \dots + a_p R_{t-p} + \varepsilon_t, \quad (5.6)$$

and then test if all the slope coefficients are zero with a χ^2 test. The return horizon can be whatever (seconds, years,...), but it is important that the returns are non-overlapping. See Figures 5.5–5.6 for illustrations.

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, we can estimate the covariance matrix of the parameters by an estimator like Newey-West. It can be noticed that when $R_t = c + a_1 R_{t-1} + \varepsilon_t$, then a_1 equals the first autocorrelation coefficient.

The autoregression can easily allow for the coefficients to depend on the market situation. For instance, consider an AR(1), but where the autoregression coefficient may be

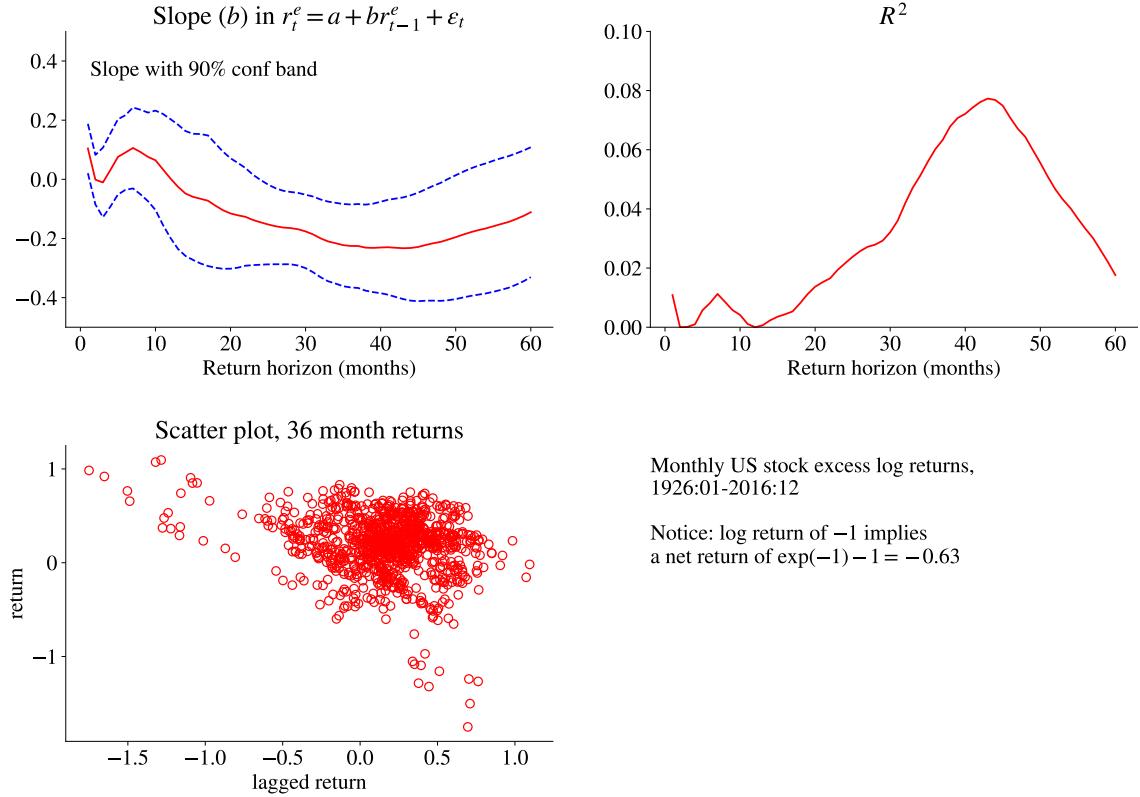


Figure 5.5: Predictability of US stock returns

different depending on the sign of last period's return

$$R_t = \alpha + \beta Q_{t-1} R_{t-1} + \gamma(1 - Q_{t-1}) R_{t-1} + \varepsilon_t, \text{ where} \quad (5.7)$$

$$Q_{t-1} = \begin{cases} 1 & \text{if } R_{t-1} < 0 \\ 0 & \text{else.} \end{cases}$$

See Figure 5.7 for an illustration.

Remark 5.10 (*Pitfall I in testing long-run returns*) Let the return in (5.6) be a two period return, $r_t = \tilde{r}_t + \tilde{r}_{t-1}$, where \tilde{r}_t is a one-period (log) return. An AR(1) on overlapping data would then be

$$\tilde{r}_t + \tilde{r}_{t-1} = c + a(\tilde{r}_{t-1} + \tilde{r}_{t-2}) + \varepsilon_t.$$

Even if the one-period returns are uncorrelated, a would tend to be positive and significant—since \tilde{r}_{t-1} shows up on both the left and right hand sides: the returns are overlapping.

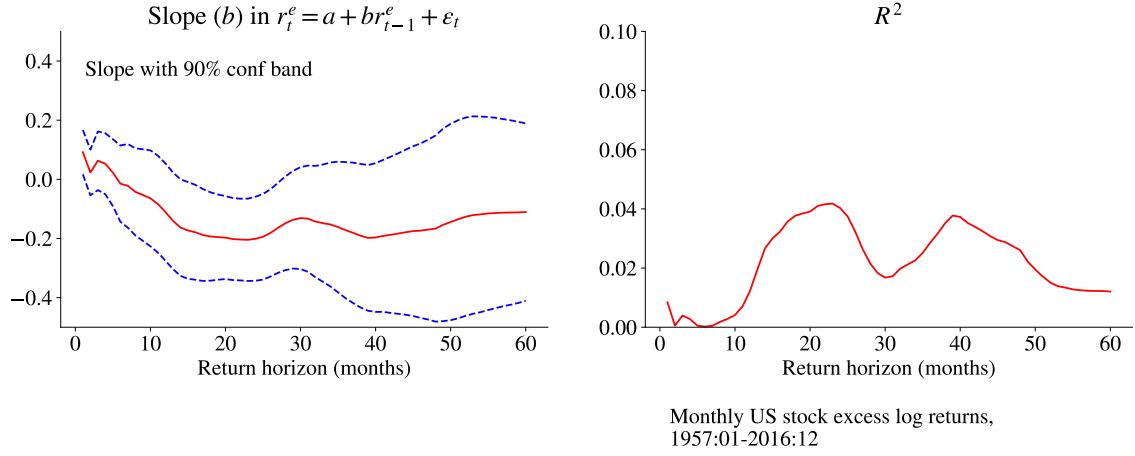


Figure 5.6: Predictability of US stock returns

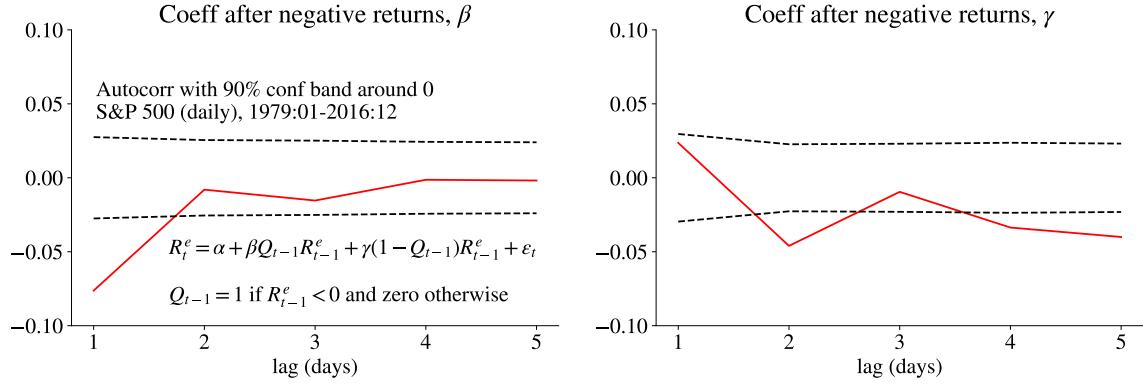


Figure 5.7: Predictability of US stock returns, results from a regression with interactive dummies

Instead, the correct specification is

$$\tilde{r}_t + \tilde{r}_{t-1} = c + a(\tilde{r}_{t-2} + \tilde{r}_{t-3}) + \varepsilon_t.$$

Remark 5.11 (*Pitfall 2 in testing long-run returns*) A less serious pitfall is to use all available returns on the left hand side, for instance, all daily two-day returns. Two successive observations are then

$$\begin{aligned}\tilde{r}_t + \tilde{r}_{t-1} &= c + a(\tilde{r}_{t-2} + \tilde{r}_{t-3}) + \varepsilon_t \\ \tilde{r}_{t+1} + \tilde{r}_t &= c + a(\tilde{r}_{t-1} + \tilde{r}_{t-2}) + \varepsilon_{t+1}\end{aligned}$$

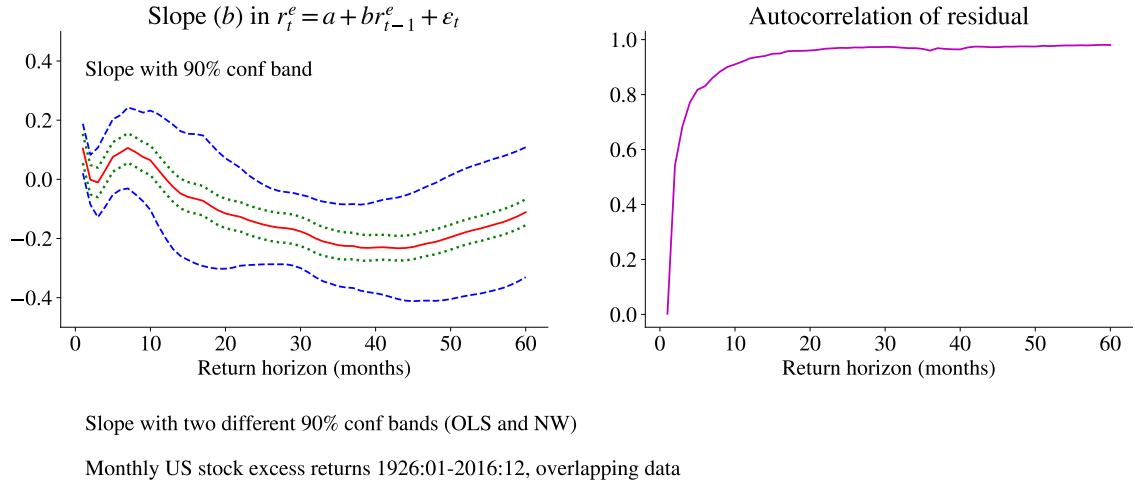


Figure 5.8: Slope coefficient, LS vs Newey-West standard errors

There is no problem with the point estimate of a , since the left and right hand side returns do not overlap. However, the residuals (ε_t and ε_{t+1}) are likely to be correlated which has to be handled in order to make correct inference. To see this, suppose $\tilde{r}_t = c/2 + u_t$ where u_t is iid. Clearly, the left and right hand sides are uncorrelated, so $a = 0$. With this we have

$$\begin{aligned}\tilde{r}_t + \tilde{r}_{t-1} &= c + \varepsilon_t, \text{ where } \varepsilon_t = u_t + u_{t-1} \\ \tilde{r}_{t+1} + \tilde{r}_t &= c + \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} = u_{t+1} + u_t.\end{aligned}$$

Since u_t shows up in both ε_t and ε_{t+1} , the latter are correlated. See Figure 5.8. This can be solved by using a Newey-West approach (or something similar), or by skipping every second observation (there is then no overlap of the residuals).

5.3 Multivariate (Auto-)correlations

There is no reason to restrict the prediction model to only use the lagged returns of the same asset. See Figure 5.9 for an illustration.

5.3.1 Momentum or Contrarian Strategy?

Reference: Lo and MacKinlay (1990)

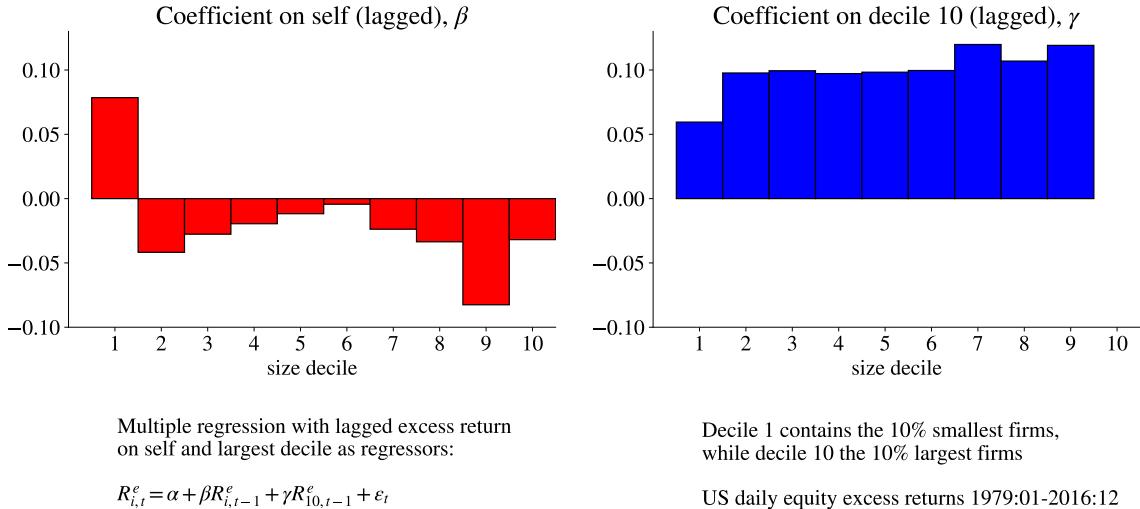


Figure 5.9: Coefficients from multiple prediction regressions

A momentum strategy invests in assets that have performed well recently—and often goes short in those that have underperformed. The performance is driven by both autocorrelation and spill-over effects from other assets. See 5.10 for an empirical illustration.

To disentangle the drivers of the return on a dynamic strategy, let there be N assets with returns R , with means and a cross autocovariance matrix

$$\begin{aligned} E R &= \mu \text{ and} \\ \Gamma(k) &= E[(R_t - \mu)(\tilde{R}_{t-k} - \mu)'], \end{aligned} \tag{5.8}$$

where \tilde{R}_{t-k} can be the returns in $t - k$ (so $\Gamma(k)$ is an autocovariance matrix) or instead the (time series) average returns over a period ending in $t - k$ (for instance, a moving average over 22 trading days). See Figure 5.11 for an empirical illustration.

Example 5.12 ($\Gamma(k)$ with two assets) We have

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

When $\tilde{R}_{t-k} = \tilde{R}_{t-k}$, then this is the autocovariance matrix for lag 1.

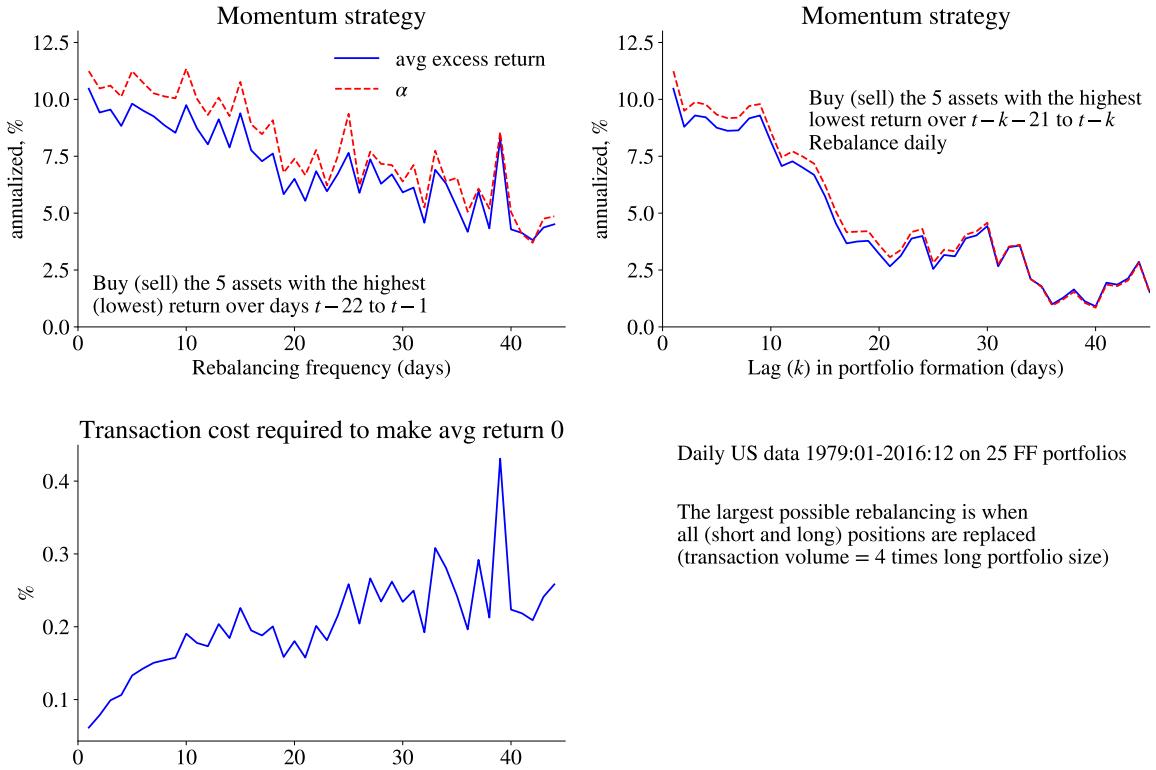


Figure 5.10: Performance of momentum investing

Define the equal weighted market portfolio return as

$$R_{mt} = \frac{1}{N} \sum_{i=1}^N R_{it}, \quad (5.9)$$

with the corresponding mean return $\mu_m = E R_{mt}$.

A *momentum strategy* could (for instance) use the $N \times 1$ vector or portfolio weights

$$w_t(k) = \frac{\tilde{R}_{t-k} - \tilde{R}_{mt-k}}{N}, \quad (5.10)$$

which says that $w_{it}(k)$ is positive for assets with a return above (the cross-sectional) average return k periods back. (To analyse a contrarian strategy, reverse the sign of (5.10).) Notice that the portfolio weights depend on $\tilde{R}_{t-k} - \tilde{R}_{mt-k}$, which can be just the returns in $t-k$ or perhaps a moving average of returns for a period ending in $t-k$. For instance, with daily data the weights for day t may depend on the returns over the last month.

Notice that the weights sum to zero, so this is a zero cost portfolio. However, the

weights differ from fixed weights (which would, for instance, be to 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.

The profit from this strategy is

$$\pi_t(k) = \sum_{i=1}^N \underbrace{\frac{\tilde{R}_{it-k} - \tilde{R}_{mt-k}}{N}}_{w_{it}} R_{it} = \sum_{i=1}^N \frac{\tilde{R}_{it-k} R_{it}}{N} - \tilde{R}_{mt-k} R_{mt}, \quad (5.11)$$

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

The expected profit is

$$E \pi_t(k) = \frac{N-1}{N^2} \text{tr } \Gamma(k) - \frac{1}{N^2} [\mathbf{1}' \Gamma(k) \mathbf{1} - \text{tr } \Gamma(k)] + \frac{1}{N} \sum_{i=1}^N (\mu_i - \mu_m)^2, \quad (5.12)$$

where the $\mathbf{1}' \Gamma(k) \mathbf{1}$ sums all the elements of $\Gamma(k)$ and $\text{tr } \Gamma(k)$ sums the elements along the main diagonal. (See below for a proof.)

With a random walk, $\Gamma(k) = 0$, (5.12) 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 (5.12) 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).

Notice that the *second term* of (5.12) 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, consider the case with only two assets and suppose we observe a higher lagged return on asset 1 than on asset 2. If this predicts a low return on asset 2 (since $\text{Cov}(R_{2,t}, R_{1,t-k}) < 0$), but asset 2 does not predict asset 1 (since $\text{Cov}(R_{1,t}, R_{2,t-k}) = 0$), then the momentum strategy will profit. The reason is that we have a negative portfolio weight of asset 2 (since it performed relatively worse than asset 1 in the previous period).

See Tables 5.1 and 5.2 for an empirical illustration.

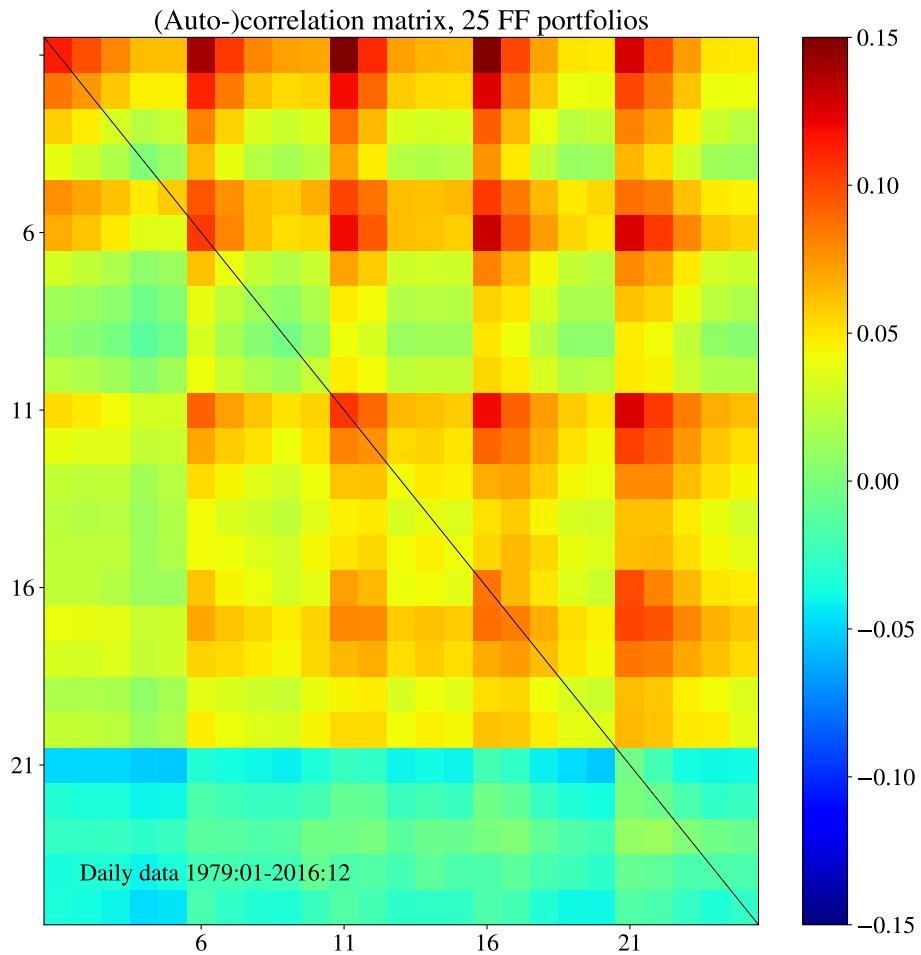


Figure 5.11: Illustration of the cross-autocorrelations, $\text{Corr}(R_t, R_{t-k})$, daily FF data. Dark colours indicate high correlations, light colours indicate low correlations.

Example 5.13 ((5.12) with 2 assets) With

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

then

$$\frac{N-1}{N^2} \text{tr } \Gamma(k) = \frac{2-1}{2^2} \times (0.1 + 0.1) = 0.05, \text{ and}$$

$$-\frac{1}{N^2} [\mathbf{1}' \Gamma(k) \mathbf{1} - \text{tr } \Gamma(k)] = -\frac{1}{2^2} (0.2 - 0.2) = 0$$

so the sum of the first two terms of (5.12) is positive (good for a momentum strategy).

Example 5.14 ((5.12) 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

$$\frac{N-1}{N^2} \text{tr } \Gamma(k) = \frac{2-1}{2^2} \times 0 = 0, \text{ and}$$

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

so the sum of the first two terms of (5.12) 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).

	Portfolio 1	Portfolio 2	Portfolio 3
1	0.58	14.63	9.67
2	0.47	11.91	8.18
3	0.49	12.40	8.78
4	0.45	11.38	8.12
5	0.41	10.41	7.83

Table 5.1: Returns on different momentum portfolios, annualized %. The rows are for different formation lags (days). Portfolio 1 follows Lo and MacKinlay (1990), except that the portfolio weights depend on the average return over the previous month. Portfolio 2 applies a static scaling of the portfolio weights to get an average long (short) exposure of 1. Portfolio 3 instead scales the weights in each period. Daily US data 1979:01-2016:12 on 25 FF portfolios.

Proof. (of (5.12)) Take expectations of (5.11) and use the fact that $\text{E } xy = \text{Cov}(x, y) + \text{E } x \text{ E } y$ to get

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

(using the fact that $\text{E } \tilde{R}_{it-k} = \text{E } R_{it}$ and $\text{E } \tilde{R}_{mt-k} = \text{E } R_{mt}$). Define the $N \times N$ cross-covariance matrix $\Gamma(k) = \text{Cov}(R_t, \tilde{R}_{t-k})$ and recall that $R_{mt} = \mathbf{1}' R_t / N$ (and $\tilde{R}_{mt} =$

	auto cov	Cross cov	means	sum
1	1.45	-0.90	0.03	0.58
2	0.93	-0.49	0.03	0.47
3	1.18	-0.72	0.03	0.49
4	0.85	-0.43	0.03	0.45
5	0.75	-0.36	0.03	0.41

Table 5.2: Contributions to the average returns on a momentum portfolio, annualized %. The rows are for different formation lags (days). The strategy follows Lo and MacKinlay (1990), except that the portfolio weights depend on the average return over the previous month. Daily US data 1979:01-2016:12 on 25 FF portfolios.

$\mathbf{1}'\tilde{R}_t/N$). We can then rewrite the terms as

$$\begin{aligned}\frac{1}{N} \sum_{i=1}^N \text{Cov}(R_{it}, \tilde{R}_{it-k}) &= \frac{1}{N} \text{tr } \Gamma(k) \\ \text{Cov}(R_{mt}, \tilde{R}_{mt-k}) &= \mathbf{1}' \Gamma(k) \mathbf{1} / N^2 \\ \frac{1}{N} \sum_{i=1}^N \mu_i^2 - \mu_m^2 &= \frac{1}{N} \sum_{i=1}^N (\mu_i - \mu_m)^2.\end{aligned}$$

Combine to rewrite $E \pi_t$ as

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

which can be rearranged as (5.12). ■

5.4 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 short-run returns on other assets have been used to predict short-run returns.

See Figure 5.12 for an illustration.

5.4.1 Prices and Dividends

Reference: Campbell and Shiller (1988), Campbell, Lo, and MacKinlay (1997) 7 and Cochrane (2005) 20.1.

Recall that the asset price P_t , gross return R_{t+1} and dividends are related according

to

$$P_t = \frac{D_{t+1} + P_{t+1}}{R_{t+1}}. \quad (5.13)$$

(This is an identity, since it defines the gross return.) Recursively solving this equation forward gives an expression of the price (or price/dividend ratio) in terms of the present value of future dividends, where the discounting is made by the actual returns. (This is also an identity.) See Appendix for details. We now log-linearise this present value expression in order to tie it more closely to the (typically linear) econometrics methods for detecting predictability. The result is

$$p_t - d_t \approx \sum_{s=0}^{\infty} \rho^s [(d_{t+1+s} - d_{t+s}) - \tilde{r}_{t+1+s}], \quad (5.14)$$

where p_t is the log price, d_t the log dividend and \tilde{r}_{t+1+s} is a one-period log return. Also, $\rho = 1/(1 + \overline{D/P})$ where $\overline{D/P}$ is a steady state dividend-price ratio ($\rho = 1/1.04 \approx 0.96$ if $\overline{D/P}$ is 4%) and where See Appendix for details. Clearly, a high price-dividend ratio must imply future dividend growth and/or low future returns.

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

$$r_{t+q} = \alpha + \beta_q (d_t - p_t) + \varepsilon_{t+q}, \quad (5.15)$$

where r_{t+q} is the log return between t and $t + q$. 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 5.12 for an illustration.

By comparing with (5.14), we see that the dividend-ratio in (5.15) is only asked to predict a finite (unweighted) sum of future returns and not dividend growth. We should therefore expect (5.15) to work particularly well if the horizon is long (high q) and if dividends are stable over time, which seems to be the case.

5.4.2 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

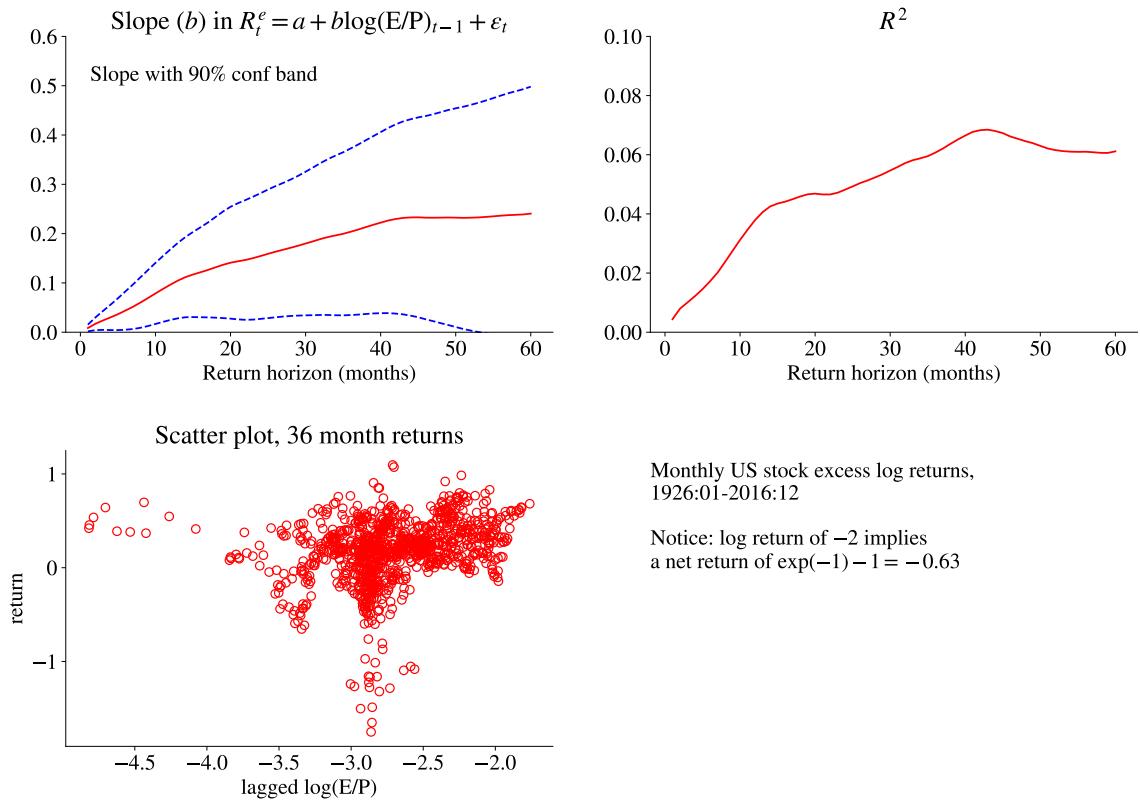


Figure 5.12: Predictability of US stock returns

let the return be forecastable by some state variable ($E_t d_{t+1+s} - E_t d_{t+s} = 0$ and $E_t r_{t+1} = r + x_t$), but in such a way that there is little autocorrelation in returns. By taking expectations of (5.14) we see that price-dividend will then reflect expected future returns and therefore be useful for forecasting.

5.5 Spurious Regressions and In-Sample Overfitting

References: Ferson, Sarkissian, and Simin (2003)

5.5.1 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

$$R_{t+1} = \alpha + \delta Z_t + v_{t+1}, \quad (5.16)$$

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

$$R_{t+1} = \mu + Z_t^* + u_{t+1}, \quad (5.17)$$

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

$$\begin{bmatrix} Z_t \\ Z_t^* \end{bmatrix} = \begin{bmatrix} \rho & 0 \\ 0 & \rho^* \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ Z_{t-1}^* \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \varepsilon_t^* \end{bmatrix}, \text{ with } \text{Cov} \left(\begin{bmatrix} \varepsilon_t \\ \varepsilon_t^* \end{bmatrix} \right) = \Sigma. \quad (5.18)$$

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. See Table 5.3 for an illustration.

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

Table 5.3: Standard error of OLS slope (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t$, ξ_t is iid $N()$. $x_t = \kappa x_{t-1} + \eta_t$, η_t is iid $N()$. NW uses 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

Under the null hypothesis of $\delta = 0$, this autocorrelation 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 (5.17), 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.

See *Table 5.3* for an illustration. The table shows 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 (5.18) 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.

5.6 Model Selection

Selecting a good prediction model is often very different from constructing a model to test a theoretical hypothesis or to establish economic causality. In particular, theory plays a somewhat smaller role (just to help identifying a set of reasonable predictors) and there is a greater emphasis on having a small model. The focus on small models is driven by a considerable amount of evidence suggesting that large prediction models often perform poorly out-of-sample.

This section summarises some standard approaches to keeping the model small, while still providing a good in-sample fit. They can be applied to the full sample or data, or on recursive/moving data windows.

5.6.1 Traditional Model Selection

Remember that R^2 can never decrease by adding more regressors, so it is not really a good guide in selecting a model (unless you have already decided on the number of pre-

dictors, for instance, only one). To avoid overfitting, we “punish” models with too many parameters by using the adjusted R^2 , defined as

$$\bar{R}^2 = 1 - (1 - R^2) \frac{T - 1}{T - k}, \quad (5.19)$$

where T is the sample size and k is the number of regressors (including the constant). This measure includes trade-off between fit and the number of regressors (per data point). Notice that \bar{R}^2 can be negative (while $0 \leq R^2 \leq 1$). Clearly, the model must include a constant for R^2 (and therefore \bar{R}^2) to make sense. Alternatively, apply Akaike’s Information Criterion (AIC) and the Bayesian information criterion (BIC). They are

$$AIC = \ln \sigma^2 + 2 \frac{k}{T} \quad (5.20)$$

$$BIC = \ln \sigma^2 + \frac{k}{T} \ln T, \quad (5.21)$$

where σ^2 is the variance of the fitted residuals.

These measures also involve trade-offs between fit (low σ^2) and number of parameters (k , including the intercept). Choose the model with the *highest* \bar{R}^2 or *lowest* AIC or BIC. It can be shown (by using $R^2 = 1 - \sigma^2 / \text{Var}(y_t)$) that AIC and BIC can be rewritten as

$$AIC = \ln \text{Var}(y_t) + \ln(1 - R^2) + 2 \frac{k}{T} \quad (5.22)$$

$$BIC = \ln \text{Var}(y_t) + \ln(1 - R^2) + \frac{k}{T} \ln T. \quad (5.23)$$

This shows that both are decreasing in R^2 (which is good), but increasing in the number of regressors per data point (k / T). It therefore leads to a similar trade-off as in \bar{R}^2 . Recall that the model should always include a constant.

Example 5.15 (*Empirical application of model selection*) See Table 5.4 for an empirical example showing a number of possible model specifications. The dependent variable is the monthly realized variance of S&P 500 returns (calculated from daily returns). The possible regressors are lags of the dependent variable, the VIX index and the S&P 500 returns. Similarly, Table 5.5 for the the best specification according to AIC. Notice that AIC tend to favour fairly large models with many regressors.

5.6.2 Sequential Model Selection

Reference: Hastie, Tibshirani, and Friedman (2001) 3

	1	2	3	4	5	6	7
RV _{t-1}	0.74 (11.02)						0.21 (1.26)
RV _{t-2}		0.58 (10.86)					0.15 (1.97)
VIX _{t-1}			0.92 (14.24)				0.89 (8.52)
VIX _{t-2}				0.68 (14.76)			-0.39 (-2.38)
R _{t-1}					-0.79 (-3.21)		-0.03 (-0.21)
R _{t-2}						-0.44 (-2.82)	-0.05 (-0.76)
constant	4.05 (4.63)	6.49 (8.71)	-2.67 (-2.59)	2.11 (2.71)	16.02 (16.83)	15.77 (16.55)	0.39 (0.45)
R2	0.54	0.34	0.63	0.34	0.15	0.05	0.65
obs	318.00	318.00	318.00	318.00	318.00	318.00	318.00

Table 5.4: Regression of monthly realized S&P 500 return volatility 1990:2–2016:12. Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

If there are k potential regressors, then there are $2^k - 1$ different models. If the list of models is not too long, then we can try them all and use the AIC and BIC in (5.20)–(5.21), see Table 5.5. Otherwise, we need some type of sequential approach.

Example 5.16 (3 potential regressors) If the three potential regressors are 1, x_1 and x_2 , then the list of models has $2^3 - 1 = 7$ possibilities: (1); (x_1) ; (x_2) ; $(1, x_1)$; $(1, x_2)$; (x_1, x_2) ; $(1, x_1, x_2)$.

A forward stepwise selection is as follows

- (1) start with an intercept
- (2) add the variable that improves the fit the most
- (3) repeat (2) until the fit does not improve much.

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

$$RSS(\beta) = \sum_{t=1}^T (y_t - x_t' \beta)^2. \quad (5.25)$$

In step (2) we would then add the variable that gives the lowest RSS (when added to the previous selection). In step (3), it is often recommended that we stop adding regressors

	1	2	3	4
RV _{t-1}	0.23 (1.52)	0.21 (1.29)	0.23 (1.50)	0.22 (1.47)
RV _{t-2}	0.14 (1.90)	0.15 (1.99)		0.14 (1.87)
VIX _{t-1}	0.89 (8.36)	0.91 (7.87)	0.90 (8.41)	0.87 (8.49)
VIX _{t-2}	-0.39 (-3.05)	-0.41 (-3.48)	-0.26 (-2.66)	-0.37 (-2.10)
R _{t-1}				-0.03 (-0.24)
R _{t-2}		-0.05 (-0.76)		
constant	0.13 (0.13)	0.33 (0.37)	-0.55 (-0.69)	0.20 (0.21)
R2	0.65	0.65	0.65	0.65
obs	318.00	318.00	318.00	318.00

Table 5.5: Best four regressions of monthly realized S&P 500 return volatility according to AIC, 1990:2–2016:12. Ordered from best (1) to fourth best (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

when

$$\frac{RSS(\hat{\beta}_{\text{old}}) - RSS(\hat{\beta}_{\text{new}})}{RSS(\hat{\beta}_{\text{new}})/(T - k - 1)} < c_{1,T-k-1}, \quad (5.26)$$

where k is the number of coefficients in $\hat{\beta}_{\text{old}}$ (including the intercept) so there are $k + 1$ coefficients in $\hat{\beta}_{\text{new}}$ and $c_{1,T-k-1}$ is the 90% or 95% critical value of an $F_{1,T-k-1}$ distribution. For instance, the 90% critical value of $F_{1,100}$ equals 2.76.

As an alternative to the RSS based rule in (5.25)–(5.26), we could instead use t-stats: in step (2) add the variable with the highest $|t\text{-stat}|$ and in step (3) stop adding variables when that $|t\text{-stat}|$ is lower than 1.64 (or 1.96).

Example 5.17 (Forward stepwise selection) Applying the forward step selection approach (based on t-stats) to the regression discussed in Example 5.15 gives a sequence of larger and larger models shown in Table 5.6.

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

	1	2	3	4
RV _{t-1}			0.23 (1.50)	0.23 (1.52)
RV _{t-2}				0.14 (1.90)
VIX _{t-1}	0.92 (14.24)	1.09 (10.73)	0.90 (8.41)	0.89 (8.36)
VIX _{t-2}		-0.21 (-2.55)	-0.26 (-2.66)	-0.39 (-3.05)
R _{t-1}				
R _{t-2}				
constant	-2.67 (-2.59)	-1.89 (-2.40)	-0.55 (-0.69)	0.13 (0.13)
R2	0.63	0.64	0.65	0.65
obs	318.00	318.00	318.00	318.00

Table 5.6: Best four regression of monthly realized S&P 500 return volatility according to a forward step selection (based on t-stats), 1990:2–2016:12. Ordered from smallest (1) to fourth smallest (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

following constrained optimization problem

$$\min_b \sum_{t=1}^T (y_t - \alpha - x_t' b)^2 \text{ subject to } \sum_{i=1}^{k-1} |b_i| \leq t, \quad (5.27)$$

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

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

The lasso method can be used as a model selection technique by estimating a sequence of models with increasingly higher t values. With a sufficiently low t , only one coefficient is non-zero—for a somewhat higher t value, two coefficients are non-zero and so on. (If we solve (5.27) with brute force, then this might involve some tweaking of the sequence of the t values. However, the lars algorithm does this automatically.) Once the L (five,

	1	2	3	4
RV _{t-1}		0.16 (1.21)	0.18 (1.38)	0.23 (1.44)
RV _{t-2}				
VIX _{t-1}	0.92 (14.24)	0.76 (7.33)	0.69 (5.86)	0.87 (8.33)
VIX _{t-2}				-0.24 (-1.57)
R _{t-1}			-0.22 (-2.62)	-0.04 (-0.33)
R _{t-2}				
constant	-2.67 (-2.59)	-1.91 (-2.85)	-0.68 (-1.02)	-0.46 (-0.64)
R2	0.63	0.63	0.64	0.65
obs	318.00	318.00	318.00	318.00

Table 5.7: Regression of monthly realized S&P 500 return volatility with model selection done by lasso, but then estimated with OLS, 1990:2–2016:12. Ordered from smallest (1) to fourth smallest (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

say) smallest specifications are found, we could re-estimate each of them with OLS. (This is the lars-OLS hybrid discussed in Efron, Hasti, Johnstone, and Tibshirani (2004).)

Example 5.18 (*lasso regression*) Applying the lasso approach to the regression discussed in Example 5.15 gives a sequence of larger and larger models. Re-estimating the four smallest of those models with OLS gives the results in Table 5.7.

Remark 5.19 (*Ridge regression**) The ridge regression solves $\min_b \sum_{t=1}^T (y_t - \alpha - x_t' b)^2 + \lambda \sum_{i=1}^{k-1} b_i^2$, where $\lambda > 0$, so it forms a compromise between OLS and zero coefficients. This is easiest to see if y_t and x_t are demeaned so $\alpha = 0$. Then, the first order conditions for minimization are $\sum_{t=1}^T x_t (y_t - x_t' \tilde{b}) - \lambda \tilde{b} = 0$, so $\tilde{b} = (\frac{1}{T} \sum_{t=1}^T x_t x_t' + \lambda)^{-1} \frac{1}{T} \sum_{t=1}^T x_t y_t$. Notice that $\lambda = 0$ gives OLS, while $\lambda = \infty$ gives $\tilde{b} = \mathbf{0}$.

Remark 5.20 (*Application of the lasso/lars algorithms*) These algorithms often standardize x_t to have zero means and unit standard deviations, and y_t to have zero means. In some cases, they also calculate $b_i \sqrt{T}$ instead of b_i .

5.7 Forecast Averaging

Reference: Elliot and Timmermann (2016) 14

Averaging across forecasts have often proved to be a good way of producing a superior forecast.

There are two main cases: (1) when we have access to the forecasts and also the data/model that produced them and (2) when we have access to the forecasts only. We discuss them in reverse order.

Suppose we have access to K different forecasts (\hat{R}_t^i for $i = 1$ to K) of the return R_t . All these forecasts are made in period $t - h$ (with $h \geq 1$). We form a weighted average as

$$R_t^* = \sum_{i=1}^K w_i \hat{R}_t^i, \text{ with } \sum_{i=1}^K w_i = 1. \quad (5.28)$$

For instance, w be chosen as to minimize the forecast error variance or the MSE over the sample up to and including $t - h$. In practice, it seems difficult to beat an unweighted average or an unweighted average after having pruned the most extreme forecasts (“trimmed mean”).

Instead, suppose we have access also to the models and data that produces the various forecasts. It can then be argued that the proper way to proceed is to pool all the data and apply the model selection techniques. However, the unweighted average across forecasts often perform reasonably well.

5.8 Out-of-Sample Forecasting Performance

References: Goyal and Welch (2008), and Campbell and Thompson (2008)

The idea of out-of-sample forecasting is to replicate real life forecasting. The prediction equation is estimated on data up to and including $t - 1$, and then a forecast is made for period t . The forecasting performance of the equation is then compared to some benchmark prediction model like the historical average (also estimated on data up to and including $t - 1$). See Figure 5.13 for an illustration. Then, the sample is extended with one period (t) and a forecast is made for $t + 1$. This continues until the sample is exhausted.

Goyal and Welch (2008) find that the evidence of predictability of equity returns disappears when out-of-sample forecasts are considered.

In contrast, Campbell and Thompson (2008) claim that there is still some out-of-sample predictability, provided we put restrictions on the estimated models. They first report that only few variables (earnings price ratio, T-bill rate and the inflation rate) have

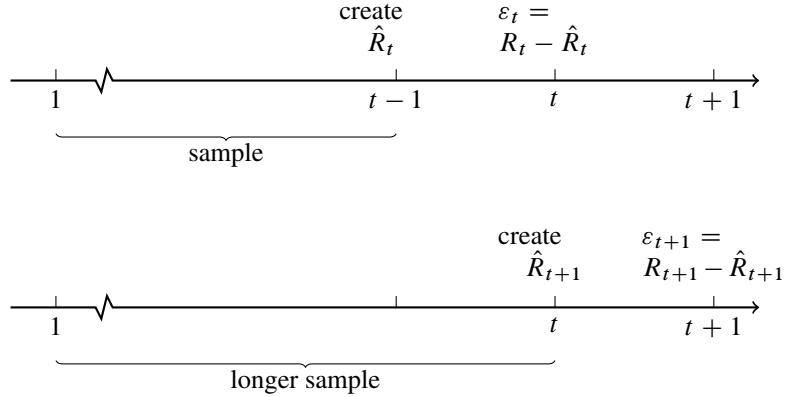


Figure 5.13: Out-of-sample forecasting

significant predictive power for one-month stock returns in the full sample (1871–2003 or early 1920s–2003, depending on predictor). The comparison is done in terms of the MSE and an “out-of-sample R^2 ”

$$R_{OS}^2 = 1 - \sum_{t=s}^T \varepsilon_t^2 / \sum_{t=s}^T e_t^2, \quad (5.29)$$

where s is the first period with an out-of-sample forecast, $\varepsilon_t = R_t - \hat{R}_t$ is the forecast based on the prediction model (estimated on data up to and including $t-1$) and $e_t = R_t - \tilde{R}_t$ is the prediction from some benchmark model (also estimated on data up to and including $t-1$). The paper uses the historical average (also estimated on data up to and including $t-1$) as the benchmark prediction. The evidence shows that the out-of-sample forecasting performance is very weak—as claimed by Goyal and Welch (2008).

Campbell and Thompson (2008) argue 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 Figures 5.14–5.15 for an illustrations. The evidence suggests that the in-sample long-run predictability vanishes out-of-sample. It also suggests that there is still some short-run predictability for small firm stocks.

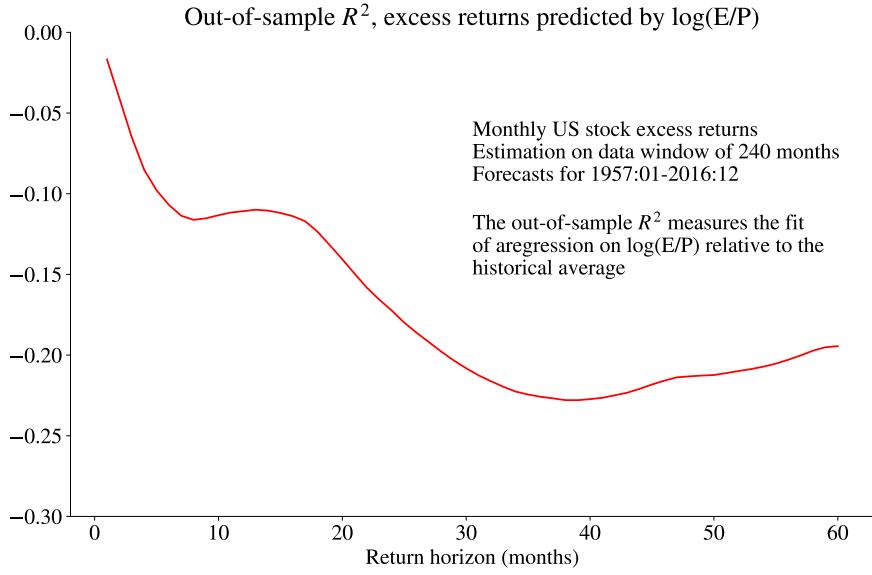


Figure 5.14: Predictability of US stock returns, out-of-sample

5.9 Evaluating Forecasting Performance

Further reading: Diebold (2001) 11; Stekler (1991); Diebold and Mariano (1995); Clark and West (2007)

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

To set up tests of the forecasting performance, let ε_t be the forecast error in period t

$$\varepsilon_t = R_t - \hat{R}_t, \quad (5.30)$$

where \hat{R}_t is the forecast (made in $t - h$) and R_t the actual outcome. (Warning: some authors prefer to work with $\hat{R}_t - R_t$ as the forecast error instead.)

Quite often, we compare a forecast method (or forecasting institute) with a benchmark forecast like a “no change,” a random walk or the historical average. The idea of such a comparison is to study if the resources employed in creating the forecast really bring value added compared to a very simple (and inexpensive) forecast.

Ultimately, the ranking of forecasting methods should be done based on the true ben-

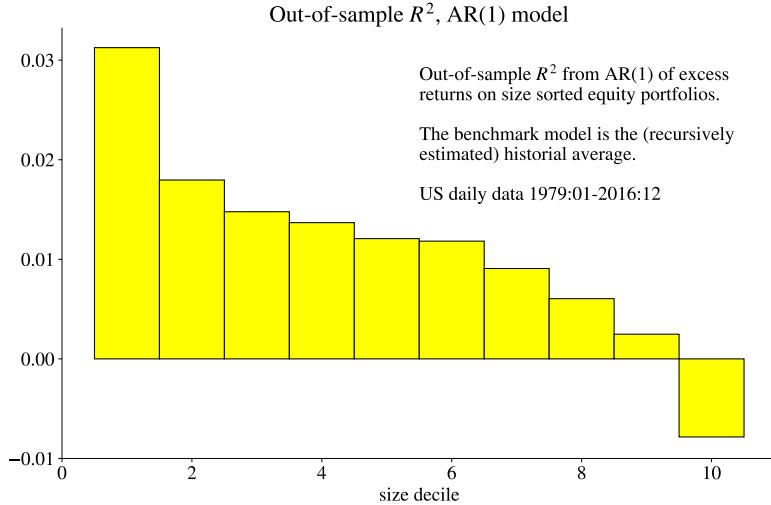


Figure 5.15: Short-run predictability of US stock returns, out-of-sample.

efits/costs of forecast errors—which may differ between organizations. For instance, a forecasting agency has a reputation (and eventually customers) to lose, while an investor has more immediate pecuniary concerns. Unless the relation between the forecast error and the losses are immediately understood, the ranking of two forecast methods is typically done based on a number of standard criteria. Several of those criteria are inspired by basic statistics.

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

$$R_t = \beta_0 + \beta_1 x_{t-h} + \varepsilon_t \quad (5.31)$$

to minimize the sum of squared residuals. This will, among other things, give a zero mean of the fitted residuals and also a zero correlation between the fitted residual and the regressor. As usual, rational forecasts should have forecast errors that cannot be predicted (by past regressors or forecast errors).

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 (i) whether the forecast error, e_t , has a zero mean; (ii) the mean squared (or absolute value) of the forecast error ; (iii) the fraction of times the squared (or absolute value) of the forecast error is lower than some threshold; (iv) the profit from investing by following a forecasting model; (v) if the forecast errors are autocorrelated or correlated

with past information.

Remark 5.21 (*Autocorrelation of forecast errors**) *An efficient h -step-ahead forecast error has a zero correlation with the forecast error h (and more) periods earlier. For instance, with $h = 2$, let $e_{t+2,t} = y_{t+2} - E_t y_{t+2}$ be the error of forecasting y_{t+2} using the information in period t . It should be uncorrelated with $e_{t,t-2} = y_t - E_{t-2} y_t$, since the latter is known when the forecast $E_t y_{t+2}$ is formed.*

To perform formal tests of forecasting performance a Diebold and Mariano (1995) test is typically performed. It is an application of GMM. To implement it, consider two different forecasts. For instance, the first forecast could come from a naive forecasting model (for instance, no change) that you hope to beat (forecast errors e_t) and the other is your estimated model (forecast errors ε_t). To test the different aspects discussed before, let $\delta(x)$ be an indicator function that is one if x is true and zero otherwise, and let R_t^e and R_t^ε denote the returns from following trading strategies based on the different forecasts. Then, we could consider, for instance, the following moment conditions

$$g_t = e_t - \varepsilon_t, \text{ or} \quad (5.32)$$

$$g_t = e_t^2 - \varepsilon_t^2 \text{ or } g_t = |e_t| - |\varepsilon_t|, \text{ or} \quad (5.33)$$

$$g_t = \delta[\text{sign}(\hat{R}_t) \neq \text{sign}(R_t)] - \delta[\text{sign}(\hat{R}_t) \neq \text{sign}(R_t)], \text{ or} \quad (5.34)$$

$$g_t = R_t^e - R_t^\varepsilon, \text{ or} \quad (5.35)$$

$$g_t = e_t e_{t-1} - \varepsilon_t \varepsilon_{t-1} \text{ or } g_t = e_t x_{t-h} - \varepsilon_t x_{t-h}. \quad (5.36)$$

The different moment conditions correspond to the different aspects of the forecasts discussed above. For instance, (5.32) is for testing if the two methods have the same average forecast error, while (5.33) tests the MSE, which is an application of the Mariano-Diebold approach. In contrast, (5.34) tests if the e model forecasts the wrong sign of the return more often than the ε model does. Finally, (5.35) compares the returns of a trading strategy (not specified here) that depends on the forecasts and (5.36) tests if the e_t errors are more predictable than the ε_t errors.

From the usual properties of GMM, we have typically have

$$\sqrt{T} \bar{g} \xrightarrow{d} N(0, S_0), \quad (5.37)$$

where $\bar{g} = \sum_{t=1}^T g_t / T$ is the average moment condition and S_0 is the variance of $\sqrt{T} \bar{g}$. S_0 can be estimated by, for instance, a Newey-West approach. It is especially important to

handle autocorrelations in the forecast errors when we are forecasting multi-period returns using overlapping data (for instance, monthly data on annual returns). This can be used to construct a t-test. The tests are typically two-sided (are the forecast errors different?), but it sometimes makes sense to use a one-sided test, in particular, when testing nested models (are the errors from the larger model “smaller” than the errors from the smaller model?).

However, when the models behind e and ε are *nested* (say, e is generated by a special case of the model that generates ε), then the asymptotic distribution is non-normal so other critical values must be applied (see [Clark and McCracken \(2001\)](#)). This is, for instance, the case when model behind e includes just an intercept and model behind ε has an intercept and a slope coefficient of some predictor x . If applied to returns, the model behind e would just pick up the historical average return, while the model behind ε would also capture the predictive changes related to x . The basic reason for the non-normal behaviour is that, even under the null hypothesis of equal performance, the average $e_t^2 - \varepsilon_t^2$ is likely to negative since ε_t^2 is affected by the noise caused by estimating too many parameters. [Clark and West \(2007\)](#) suggest another way of handling this problem. In particular, they suggest replacing the squared forecast errors in [\(5.33\)](#) with

$$g_t = e_t^2 - [\varepsilon_t^2 - (\hat{R}_t^e - \hat{R}_t^\varepsilon)^2], \quad (5.38)$$

and then use [\(5.37\)](#). This approach adjusts for the fact that the model behind ε is affected by noise caused by the estimation of the extra parameters. (This logic assumes that the smaller model is the true one, so the larger model includes parameters that ought to be set to zero.)

Since $R_t^e = \hat{R}_t^e + e_t$ (and similarly for ε_t), [\(5.38\)](#) can be rewritten in terms of the forecast errors only

$$g_t = e_t^2 - [\varepsilon_t^2 - (e_t - \varepsilon_t)^2] \quad (5.39)$$

$$= 2e_t(e_t - \varepsilon_t). \quad (5.40)$$

(Recall that e_t is the error from the smaller model, while ε_t is from the larger model.) The simulation evidence in [Clark and West \(2007\)](#) suggests that using [\(5.40\)](#) or applying a bootstrap to [\(5.33\)](#) have similar properties. The bootstrap approach can also be readily applied to the other evaluation criteria [\(5.32\)–\(5.36\)](#).

Remark 5.22 (*Empirical results on predicting annual S&P 500 returns*) [Tables 5.8–5.9](#) and [Figure 5.16](#) summarize the results. The combined model seems to do slightly better

	AR(1)		E/P		Combination	
	mean	t-stat	mean	t-stat	mean	t-stat
MSE in-sample	295.78		279.30			
R^2_{oos}	-0.05		-0.05		-0.02	
$e - \varepsilon$	0.24	1.88	-1.22	-1.09	-0.49	-0.85
$e^2 - \varepsilon^2$	-14.86	-1.61	-15.14	-0.59	-6.26	-0.50
$ e - \varepsilon $	-0.25	-1.58	-0.62	-0.80	-0.22	-0.57
wrong sign	0.00		-0.04	-1.48		
$2e(e - \varepsilon)$	-13.85	-1.56	18.49	0.71	2.32	0.19

Table 5.8: Mariano-Diebold (and Clark-West) tests of forecasting 1-year S&P returns with different models. The total sample is 1946–2015, but the forecasts as made for 1971–2015. The e forecasts are the historical average returns while the ε forecasts are out-of-sample and based on the different regressions. The ‘wrong sign’ indicates the wrong sign of the forecast and takes the difference between the e model and the ε model. Estimation is done on an expanding data window. The std use a NW approach with 1 lag (year).

	5th percentile	95th percentile
$e^2 - \varepsilon^2$	-2.18	0.30
$ e - \varepsilon $	-2.51	0.22
wrong sign	-1.93	1.02
$2e(e - \varepsilon)$	-1.74	1.71

Table 5.9: Bootstrapped percentiles of the Mariano-Diebold (and Clark-West) tests of the E/P model in Table 5.8. The simulations are done under the null hypothesis by randomly drawing (with replacement) the returns from the prediction sample.

than the two individual models. The build-up in the oos MSE shows some jumps, but the ranking of the three methods do not change dramatically over time. Notice also that the bootstrapped confidence bands in 5.9 appear very asymmetric, in spite of being simulated under the null hypothesis. Only the Clark-West has a symmetric confidence band.

For instance, Leitch and Tanner (1991) analyse 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.

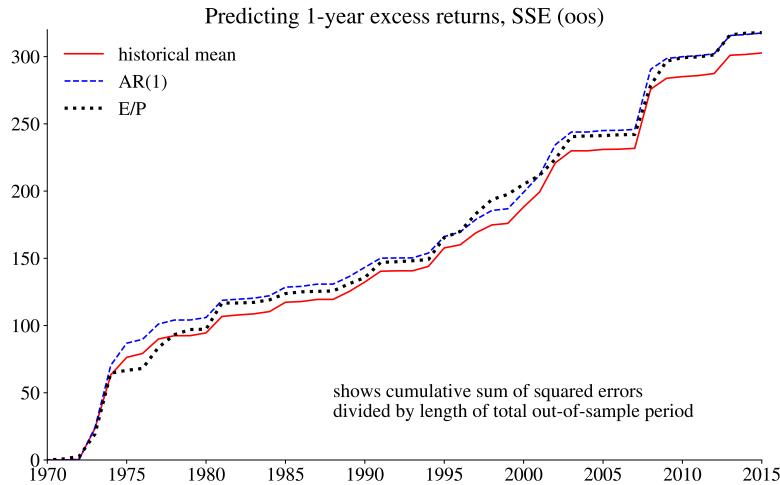


Figure 5.16: Accumulation of the (oos) MSE from three different forecasting models

5.10 Appendix: Prices and Dividends

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

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t}, \text{ so } P_t = \frac{D_{t+1} + P_{t+1}}{R_{t+1}}. \quad (5.41)$$

Substituting for P_{t+1} (and then P_{t+2}, \dots) gives

$$P_t = \frac{D_{t+1}}{R_{t+1}} + \frac{D_{t+2}}{R_{t+1}R_{t+2}} + \frac{D_{t+3}}{R_{t+1}R_{t+2}R_{t+3}} + \dots \quad (5.42)$$

$$= \sum_{j=1}^{\infty} \frac{D_{t+j}}{\prod_{k=1}^j R_{t+k}}, \quad (5.43)$$

provided the discounted value of P_{t+j} goes to zero as $j \rightarrow \infty$. 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 analyse the price-dividend ratio. Dividing (5.42)

and (5.43) by D_t gives

$$\frac{P_t}{D_t} = \frac{1}{R_{t+1}} \frac{D_{t+1}}{D_t} + \frac{1}{R_{t+1}R_{t+2}} \frac{D_{t+2}}{D_{t+1}} \frac{D_{t+1}}{D_t} + \frac{1}{R_{t+1}R_{t+2}R_{t+3}} \frac{D_{t+3}}{D_{t+2}} \frac{D_{t+2}}{D_{t+1}} \frac{D_{t+1}}{D_t} + \dots \quad (5.44)$$

$$= \sum_{j=1}^{\infty} \prod_{k=1}^j \frac{D_{t+k}/D_{t+k-1}}{R_{t+k}}. \quad (5.45)$$

As with (5.43) 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, (5.43), and that the price-dividend ratio should predict a discounted sum of future changes in dividends.

Proof. (of (5.14)—slow version) Rewrite (5.41) as

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \frac{P_{t+1}}{P_t} \left(1 + \frac{D_{t+1}}{P_{t+1}} \right) \text{ or in logs}$$

$$\tilde{r}_{t+1} = p_{t+1} - p_t + \ln [1 + \exp(d_{t+1} - p_{t+1})].$$

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

$$\ln [1 + \exp(d_{t+1} - p_{t+1})] \approx \ln [1 + \exp(\overline{d - p})] + \frac{\exp(\overline{d - p})}{1 + \exp(\overline{d - p})} [d_{t+1} - p_{t+1} - (\overline{d - p})]$$

$$\approx \text{constant} + (1 - \rho) (d_{t+1} - p_{t+1}),$$

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

$$\begin{aligned} \tilde{r}_{t+1} &\approx p_{t+1} - p_t + (1 - \rho) (d_{t+1} - p_{t+1}) \\ &= \rho p_{t+1} - p_t + (1 - \rho) d_{t+1}, \end{aligned}$$

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

$$\begin{aligned} \tilde{r}_{t+1} &\approx \rho (p_{t+1} - d_{t+1}) - (p_t - d_t) + (d_{t+1} - d_t), \text{ or} \\ p_t - d_t &\approx \rho (p_{t+1} - d_{t+1}) + (d_{t+1} - d_t) - \tilde{r}_{t+1} \end{aligned}$$

This is a (forward looking, unstable) difference equation, which we can solve recursively

forward. Provided $\lim_{s \rightarrow \infty} \rho^s(p_{t+s} - d_{t+s}) = 0$, the solution is (5.14). (Trying to solve for the log price level instead of the log price-dividend ratio is problematic since the condition $\lim_{s \rightarrow \infty} \rho^s p_{t+s} = 0$ may not be satisfied.) ■

Chapter 6

Predicting Asset Returns: Nonparametric Estimation

6.1 Basics of Kernel Regressions

Reference: Campbell, Lo, and MacKinlay (1997) 12.3; Härdle (1990); Pagan and Ullah (1999); Mittelhammer, Judge, and Miller (2000) 21

6.1.1 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

$$y_t = b(x_t) + \varepsilon_t, \quad (6.1)$$

where ε_t is uncorrelated over time and where $E \varepsilon_t = 0$ and $E(\varepsilon_t | x_t) = 0$. The function $b()$ is unknown and possibly non-linear. In comparison, in a linear regression we have $b(x_t) = \beta x_t$.

One possibility of estimating such a function is to approximate $b(x)$ by a polynomial (or some other basis). This will give quick estimates, but the results are “global” in the sense that the value of $b(x)$ at a particular x value ($x = 1.9$, say) will depend on all the data points—and potentially very strongly so. The approach in this section is more “local” by down weighting information from data points where x_t is far from x .

As a starting point, suppose we want to estimate $b(x)$ at $x = 1.9$. If our sample has 3 observations (say, $t = 3, 27$, and 99) with $x_t = 1.9$, then it would be straightforward to average over these three observations to estimate $b(1.9)$ as $(y_3 + y_{27} + y_{99})/3$. This makes sense, since the average of the error terms $(\varepsilon_3, \varepsilon_{27}, \varepsilon_{99})$ is likely to be close to zero.

Unfortunately, we seldom have repeated observations of this type. Moreover, it seems

to be a waste to disregard data points where x_t is close, but not equal, to x . Instead, we may try to estimate the value of $b(x)$ by averaging over (y) observations where x_t is close to x (here 1.9). The general form of this type of estimator is

$$\hat{b}(x) = \frac{\sum_{t=1}^T w(x_t - x)y_t}{\sum_{t=1}^T w(x_t - x)}, \quad (6.2)$$

where $w(x_t - x)/\sum_{t=1}^T w(x_t - x)$ is the weight on data in t (in practice, y_t). This weight is non-negative and (weakly) decreasing in the distance of x_t from x . Note that the denominator makes the weights sum to unity. The basic assumption behind (6.2) is that the $b(x)$ function is smooth so local averaging (around x) makes sense.

As an example of a $w(\cdot)$ function, it could be to give equal weight to the k values of x_t which are closest to x and zero weight to all other observations (this is the “ k -nearest neighbor” estimator, see Härdle (1990) 3.2). As another example, the weight function could be defined so that it trades off the expected squared errors, $E[y_t - \hat{b}(x)]^2$, and the expected squared acceleration, $E[d^2\hat{b}(x)/dx^2]^2$. This defines a cubic spline (often used in macroeconomics when $x_t = t$, and is then called the Hodrick-Prescott filter).

Remark 6.1 (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 centred moving average (over $\pm k$ data points) of the first column (y).

6.1.2 Kernel Regression

A *Kernel regression* uses a pdf as the weight function, $w(x_t - x) = K[(x_t - x)/h]$, 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 flat function for x_t over $x - h/2$ to $x + h/2$ (and zero outside this interval). In this case, the weighting function is

$$w(x_t - x) = \delta\left(\left|\frac{x_t - x}{h}\right| \leq 1/2\right), \text{ where} \quad (6.3)$$

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

This weighting function puts a constant weight on all data point in the interval $x \pm h/2$ and zero on all other data points. It can be noticed that if we divide the weighting function by h , (which would cancel in (6.2)), then we get the density function of a variable that is

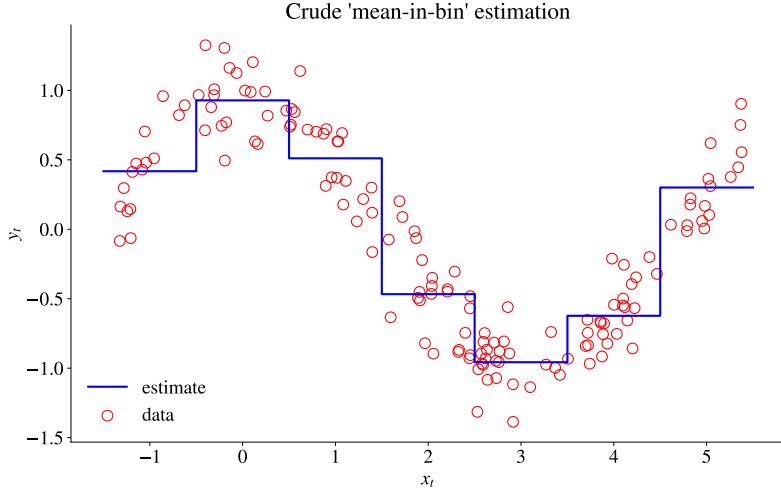


Figure 6.1: Example of a crude estimation

uniformly distributed over $x - h/2$ to $x + h/2$.

However, we can gain efficiency and get a smoother (across x values) estimate by using a density function that tapers off more smoothly. With an $N(0, 1)$ kernel applied to $(x_t - x)/h$, we get the following weights at a point x

$$w(x_t - x) = \frac{1}{\sqrt{2\pi}} \exp \left[-\left(\frac{x_t - x}{h} \right)^2 / 2 \right]. \quad (6.4)$$

Clearly, by dividing by h (which would cancel in (6.2)) we get an $N(x, h^2)$ pdf for x_t . See Figure 6.5 for an example. When $h \rightarrow 0$, then no averaging is done ($\hat{b}(x)$ evaluated at $x = x_t$ is just y_t). In contrast, as $h \rightarrow \infty$, $\hat{b}(x)$ becomes the sample average of y_t so we have global averaging. Clearly, some value of h in between is needed.

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. See Figure 6.5 for an illustration. Special corrections might be needed if there are a lot of observations stacked close to the boundary of the support of x (see Härdle (1990) 4.4).

Example 6.2 (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 $x = 1.9$.

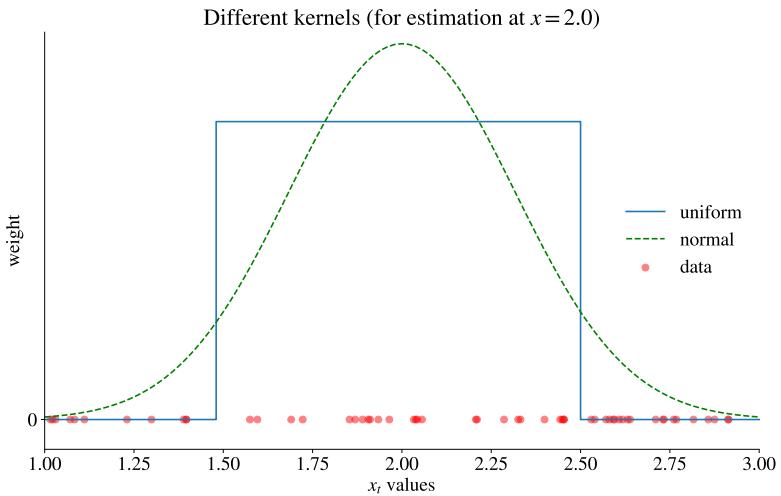


Figure 6.2: Different weighting functions for non-parametric regression

With $h = 1$, the numerator in (6.4) is

$$\begin{aligned} \sum_{t=1}^T w(x_t - x) y_t &= \left(e^{-(1.5-1.9)^2/2} \times 5 + e^{-(2-1.9)^2/2} \times 4 + e^{-(2.5-1.9)^2/2} \times 3.5 \right) / \sqrt{2\pi} \\ &\approx (0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5) / \sqrt{2\pi} \\ &= 11.52 / \sqrt{2\pi}. \end{aligned}$$

The denominator is

$$\begin{aligned} \sum_{t=1}^T w(x_t - x) &= \left(e^{-(1.5-1.9)^2/2} + e^{-(2-1.9)^2/2} + e^{-(2.5-1.9)^2/2} \right) / \sqrt{2\pi} \\ &\approx 2.75 / \sqrt{2\pi}. \end{aligned}$$

The estimate at $x = 1.9$ is therefore

$$\hat{b}(1.9) \approx 11.52 / 2.75 \approx 4.19.$$

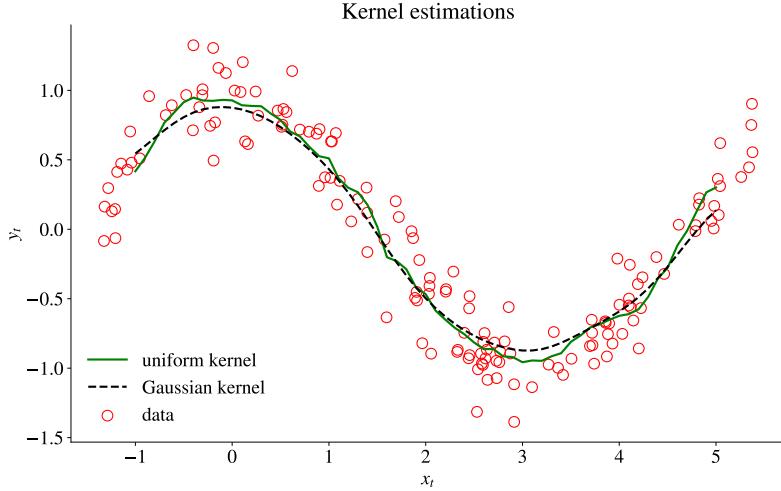


Figure 6.3: Example of kernel estimations

6.1.3 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \quad (6.5)$$

where ε_t is uncorrelated over time and where $E \varepsilon_t = 0$ and $E(\varepsilon_t | x_t, z_t) = 0$.

This makes the estimation problem more data demanding. To see why, suppose we use a uniform density function as weighting function (see in (6.3)). However, with two regressors, the interval becomes a rectangle. With as little as a 20 intervals of each of x and z , we get 400 bins, so we need a large sample to have a reasonable number of observations in every bin.

In any case, the most common way to implement the kernel regressor is to let

$$\hat{b}(x, z) = \frac{\sum_{t=1}^T w(x_t - x)v(z_t - z)y_t}{\sum_{t=1}^T w(x_t - x)v(z_t - z)}, \quad (6.6)$$

where $w()$ and $v()$ are two kernels like in (6.4) and where we may allow the bandwidth (h) to be different for x_t and z_t (and depend on the variance of x_t and y_t). In this case, the weight of the observation (x_t, z_t) is proportional to $w(x_t - x)v(z_t - z)$, which is high if both x_t and z_t are close to x and z respectively. See Figure 6.7.

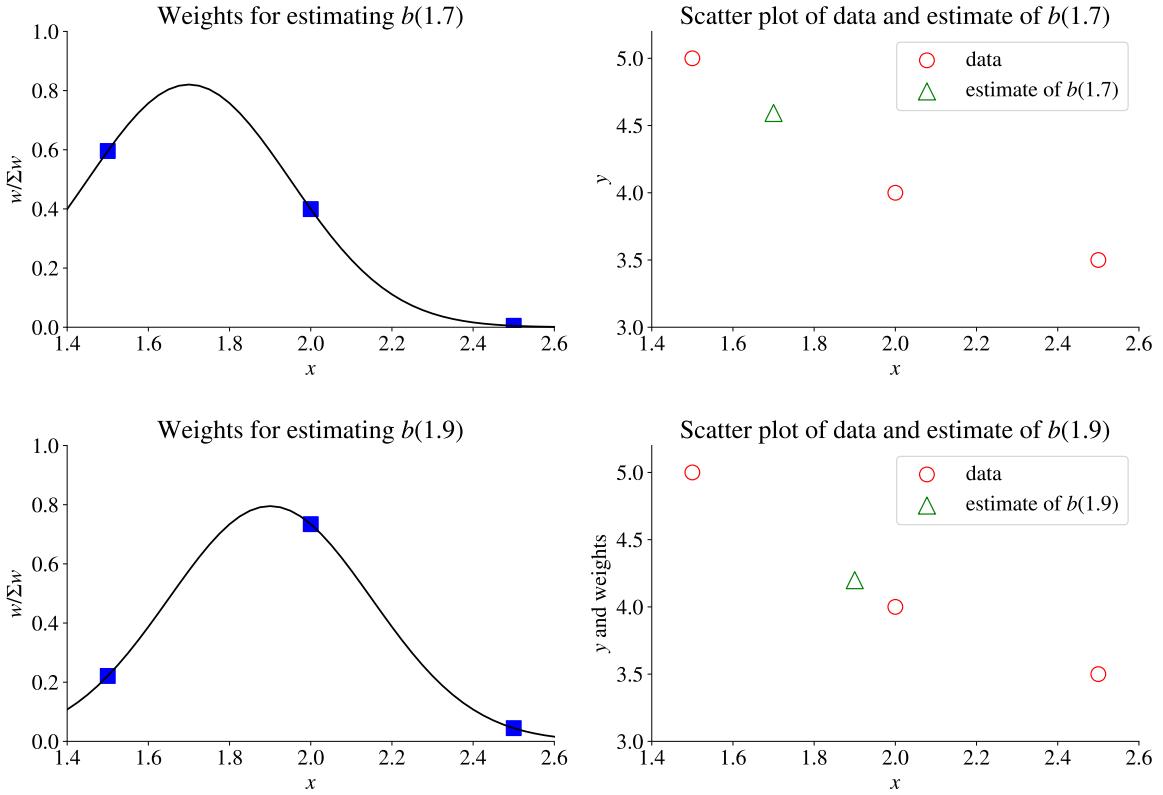


Figure 6.4: Example of kernel regression with three data points

6.2 Distribution of the Kernel Regression and Choice of Bandwidth

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 of the estimator at the value x are approximately (for general kernel functions)

$$\text{MSE}(x) = \text{Var}[\hat{b}(x)] + \text{Bias}[\hat{b}(x)]^2, \text{ with} \quad (6.7)$$

$$\text{Var}[\hat{b}(x)] = \frac{1}{Th} \frac{\sigma^2(x)}{f(x)} \times \int_{-\infty}^{\infty} K(u)^2 du \quad (6.8)$$

$$\text{Bias}[\hat{b}(x)] = h^2 \times \left[\frac{1}{2} \frac{d^2 b(x)}{dx^2} + \frac{df(x)}{dx} \frac{1}{f(x)} \frac{db(x)}{dx} \right] \times \int_{-\infty}^{\infty} K(u) u^2 du. \quad (6.9)$$

In these expressions, $\sigma^2(x)$ is the variance of the residuals in (6.1) which may depend on the x value, $f(x)$ the marginal density of x and $K(u)$ the kernel (pdf) used as a weighting

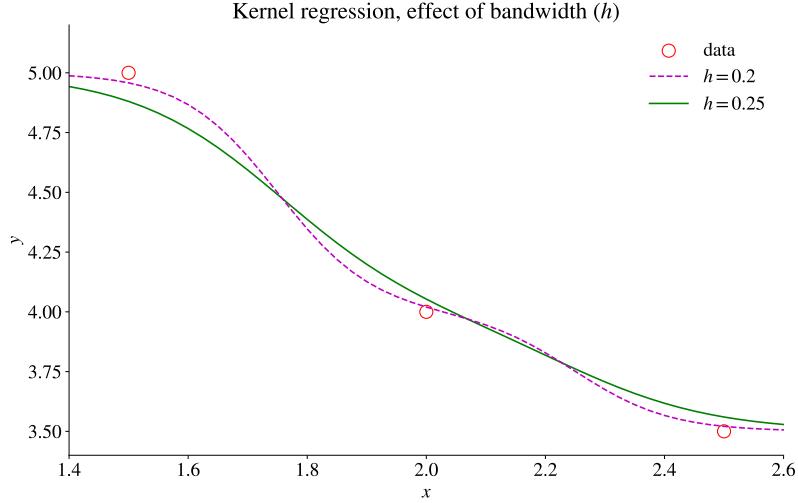


Figure 6.5: Example of kernel regression with three data points

function for $u_t = (x_t - x)/h$. The remaining terms are functions of the true regression function $b(x)$.

As a comparison, a linear regression has $\hat{b}(x) = z'\hat{\gamma}$ where $z = [1, x]$, so the variance of the fitted value is

$$\text{Var}(z'\hat{\gamma}) = z'V(\hat{\gamma})z, \quad (6.10)$$

where $V(\hat{\gamma})$ is the variance-covariance matrix of $\hat{\gamma}$. (Notice that this is different from the variance of a forecast error, since the latter also includes the variance of the residual.)

With a Gaussian kernel these expressions can be simplified to

$$\text{Var}[\hat{b}(x)] = \frac{1}{Th} \frac{\sigma^2(x)}{f(x)} \times \frac{1}{2\sqrt{\pi}} \quad (6.11)$$

$$\text{Bias}[\hat{b}(x)] = h^2 \times \left[\frac{1}{2} \frac{d^2 b(x)}{dx^2} + \frac{df(x)}{dx} \frac{1}{f(x)} \frac{db(x)}{dx} \right]. \quad (6.12)$$

Proof. (of (6.11)–(6.12)) 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 using the $N(0, 1)$ pdf for the variable $u_t = (x_t - x)/h$.) Use in (6.8)–(6.9). ■

Equations (6.8) and (6.11) show that smaller h increases the variance (we effectively use fewer data points to estimate $b(x)$) but decreases the bias of the estimator (it becomes

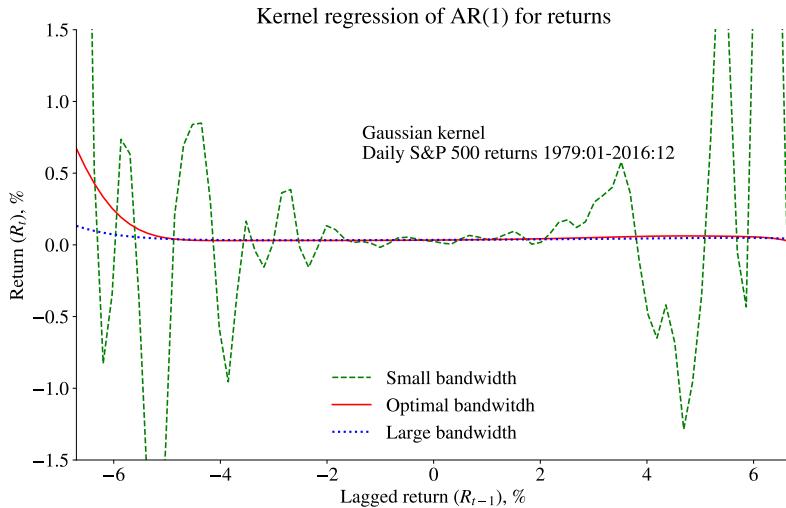


Figure 6.6: Non-parametric regression, importance of bandwidth

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). It is clear that the choice of h has a major importance on the estimation results. See Figure 6.6 for an example.

The variance is also 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.

Equations (6.9) and (6.12) show that the bias increases (in magnitude) with the curvature of the $b(x)$ function (that is, $(d^2 b(x)/dx^2)^2$). 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 .

Remark 6.3 (Rule of thumb value of h) In a simplified case, we can find the h value that minimizes the MSE by an analytical approach. Use (6.12) to construct the $MSE = \text{Var}(b) + \text{bias}(b)^2$. To simplify, assume the distribution of x is uniform, so $f(x) = 1/(x_{\max} - x_{\min})$ and $df(x)/dx = 0$. In addition, run the regression $y = \alpha + \beta x + \gamma x^2 + \varepsilon$

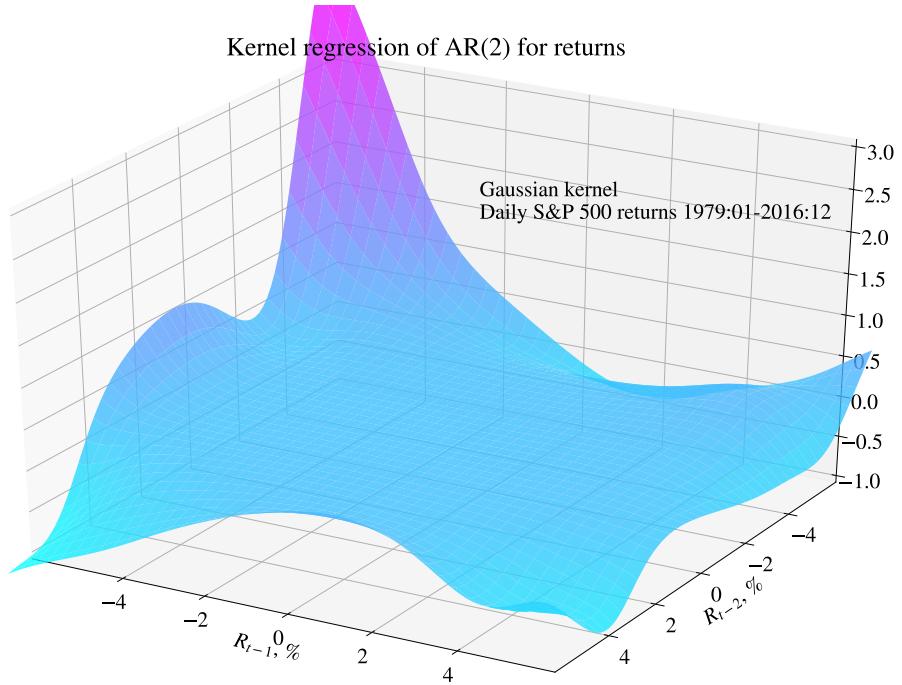


Figure 6.7: Non-parametric regression with two regressors

as an approximation of $b(x)$. With this we have $d^2b(x)/dx^2 \approx 2\gamma$ and we approximate σ^2 by the variance of the fitted residuals, σ_ε^2 . Combining, we have

$$MSE = \frac{1}{Th^2} \sigma_\varepsilon^2 (x_{\max} - x_{\min}) \frac{1}{2\sqrt{\pi}} + h^4 \gamma^2.$$

Minimizing with respect to h gives the first order condition

$$\begin{aligned} 0 &= -\frac{1}{Th^2} \sigma_\varepsilon^2 (x_{\max} - x_{\min}) \frac{1}{2\sqrt{\pi}} + 4h^3 \gamma^2, \text{ so} \\ h^5 &= \frac{1}{T\gamma^2} \sigma_\varepsilon^2 (x_{\max} - x_{\min}) \frac{1}{8\sqrt{\pi}}, \text{ or} \\ h &= T^{-1/5} |\gamma|^{-2/5} \sigma_\varepsilon^{2/5} (x_{\max} - x_{\min})^{1/5} \times 0.6. \end{aligned}$$

In practice, replace $x_{\max} - x_{\min}$ by the difference between the 90th and 10th percentiles of x .

A good (but computationally intensive) approach to choose h is by the leave-one-out cross-validation technique. This approach would, for instance, choose h to minimize the

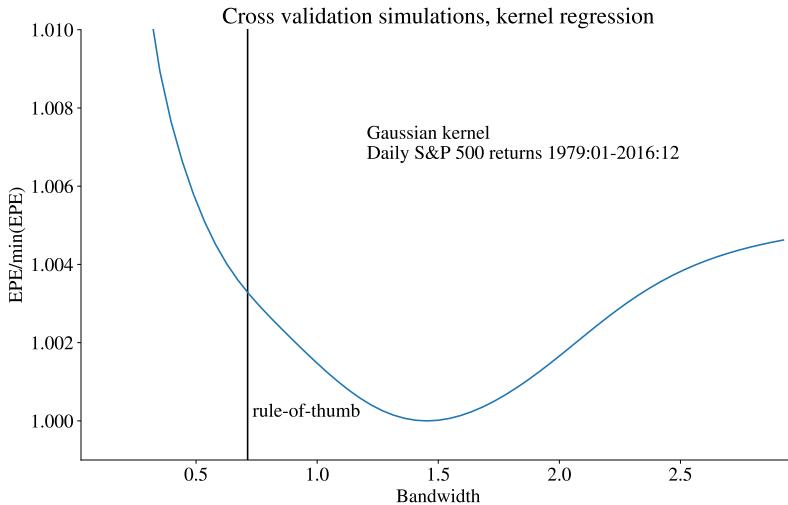


Figure 6.8: Cross-validation

expected (or average) prediction error

$$\text{EPE}(h) = \sum_{t=1}^T [y_t - \hat{b}_{-t}(x_t, h)]^2 / T, \quad (6.13)$$

where $\hat{b}_{-t}(x_t, h)$ is the fitted value of the regression function evaluated at $x = x_t$. Notice that the regression function $\hat{b}_{-t}(x_t, h)$ is estimated (using the bandwidth h) on a sample that excludes observation (y_t, x_t) . This means that each prediction is out-of-sample. To calculate (6.13) we clearly need to make T estimations, that is, we have to estimate $\hat{b}_{-t}(x_t, h)$ for each t . Then we repeat this for different values of h to find the minimum. See Figure 6.8 for an example.

Remark 6.4 (*EPE calculations*) Step 1: pick a value for h

Step 2: estimate the $b(x)$ function on all data, but exclude $t = 1$, then calculate $\hat{b}_{-1}(x_1)$ and the error $y_1 - \hat{b}_{-1}(x_1)$

Step 3: redo Step 2, but now exclude $t = 2$ and. calculate the error $y_2 - \hat{b}_{-2}(x_2)$. Repeat this for $t = 3, 4, \dots, T$. Calculate the EPE as in (6.13).

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 6.5 (*Speed and fast Fourier transforms*) A a fast Fourier transform can help speeding up the calculation of the kernel estimator.

If the observations are independent, then it can be shown (see Härdle (1990) 4.2,

Pagan and Ullah (1999) 3.3–6, and also (6.12)) that, with a Gaussian kernel, the estimator at point x is asymptotically normally distributed

$$\sqrt{Th}[\hat{b}(x) - b(x)] \xrightarrow{d} N\left(0, \frac{1}{2\sqrt{\pi}} \frac{\sigma^2(x)}{f(x)}\right), \quad (6.14)$$

where $f(x)$ is the density of x and $\sigma^2(x)$ is the variance of the residuals in (6.1) which could also depend on the x value. (A similar expression for the distribution holds for other choices of the kernel.) This expression assumes that the asymptotic bias is zero, which is guaranteed if h is decreased (as T increases) slightly faster than $T^{-1/5}$ (for instance, suppose $h = T^{-1.1/5}h_0$, where h_0 is a constant). 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 $1.06 \text{ Std}(x_t)T^{-1/5}$.

Remark 6.6 (Asymptotic bias) *The condition that h decreases faster than $T^{-1/5}$ ensures that the bias of $\sqrt{Th}\hat{b}(x)$ vanishes as $T \rightarrow \infty$. This is seen by noticing that the bias of $\hat{b}(x)$ is proportional to h^2 (see (6.12)). Multiplying by \sqrt{Th} gives the bias of $\sqrt{Th}\hat{b}(x)$ as being proportional to $T^{1/2}h^{5/2}$. With $h = T^{-1.1/5}h_0$, this bias is proportional to $T^{1/2}(T^{-1.1/5}h_0)^{5/2}$, that is, to $T^{-0.05}h_0^{5/2}$ which is decreases to zero as T increases.*

To estimate $\sigma^2(x)$ in (6.14), we may assume that it does not depend on x , so we just estimate the variance of the fitted residuals. (Clearly, this requires estimating $\hat{b}(x)$ at every point $x = x_t$ in the sample, not just a small grid of x values.) Alternatively, we use a non-parametric regression of the squared fitted residuals on x_t

$$\hat{\varepsilon}_t^2 = \sigma^2(x_t), \text{ where } \hat{\varepsilon}_t = y_t - \hat{b}(x_t), \quad (6.15)$$

where $\hat{b}(x_t)$ are the fitted values from the non-parametric regression (6.1). To draw confidence bands, it is typically assumed that the asymptotic bias is zero ($E\hat{b}(x) = b(x)$).

See Figure 6.9 for an example where the width of the confidence band varies across x values—mostly because the sample contains few observations of extreme x values as shown in Figure 6.10. In particular, compare with the confidence bands of a linear regression in Figure 6.11, which do account for the lack of data points with extreme x values.

6.3 Local Linear Regressions

Notice that (6.2) solves the problem $\min_{\alpha_x} \sum_{t=1}^T w(x_t - x)(y_t - \alpha_x)^2$ for each value of x . For a given value of x , α_x is a constant—but it can vary across x values. The first order

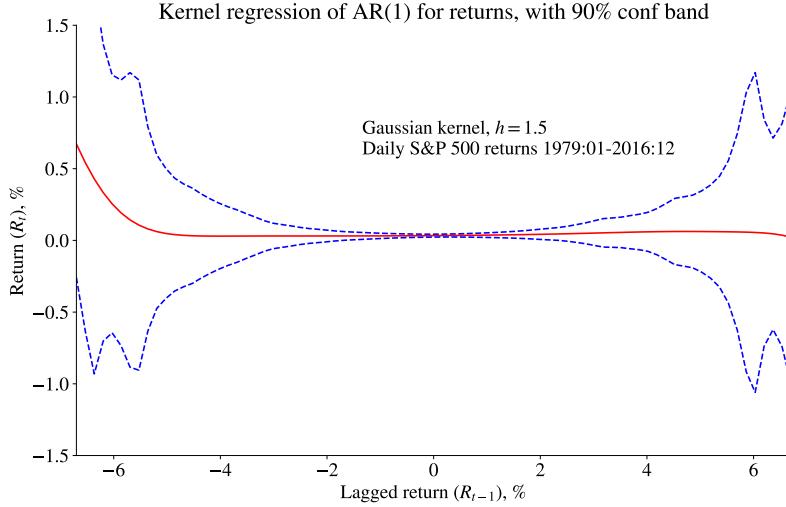


Figure 6.9: Non-parametric regression with confidence bands

condition (at a given x value) is $\sum_{t=1}^T w(x_t - x)(y_t - \alpha_x) = 0$, so the solution is as in (6.2), that is, $\hat{\alpha}_x = \hat{b}(x)$. This can be interpreted as a “local constant” regression model: for each x it is just a constant.

This can be extended to solving a problem like

$$\min_{\alpha_x, \beta_x} \sum_{t=1}^T w(x_t - x)[y_t - \alpha_x - \beta_x(x_t - x)]^2, \quad (6.16)$$

which defines the local linear estimator. (Yes, the convention is to use $x_t - x$ as the regressor, but this could easily be changed.) The first order conditions are similar to the usual normal equations for LS (except that data point t has the weight $w(x_t - x)$ and that we use $x_t - x$ as the regressor). In fact, if we let $z_t = [1, x_t - x]'$ and collect the coefficients in $\theta_x = [\alpha_x, \beta_x]'$, then the first order conditions can be written

$$\sum_{t=1}^T w(x_t - x) z_t y_t = \sum_{t=1}^T w(x_t - x) z_t z_t' \hat{\theta}_x. \quad (6.17)$$

It is straightforward to solve these, but perhaps even easier if we create $\tilde{z}_t = \sqrt{w(x_t - x)} z_t$ and $\tilde{y}_t = \sqrt{w(x_t - x)} y_t$, because (6.17) is then the same as the first order conditions for a regression of \tilde{y}_t on \tilde{z}_t (without a constant). (An extension to a quadratic or higher function seems straightforward.)

Clearly, solving (6.17) gives one $\hat{\theta}_x$ vector for each x value that we consider. Once we have the estimates, the fitted value at the value x is just $\hat{\alpha}_x$ (since the regression function is $y_t = \alpha_x + \beta_x(x_t - x) + \varepsilon_t$ and we evaluate it at $x_t = x$.)

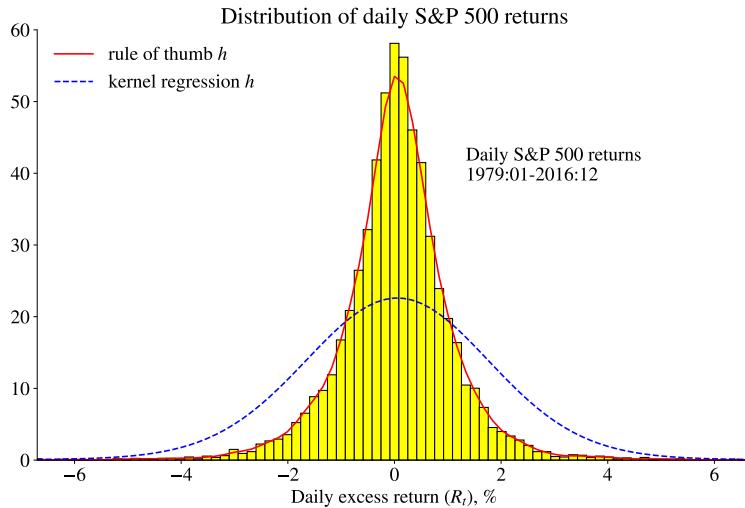


Figure 6.10: Distribution of daily stock returns

It can be shown that the local-linear estimator has the same asymptotic variance as the kernel regression, and that the bias only includes the $d^2b(x)/dx^2$ term (not the linear term). The latter means that the bias does not depend on the pdf of the regressor ($f(x)$), which is an advantage.

The bandwidth parameter (which only shows up in the calculations of the weights, $w(x_t - x)$) can be chosen by a leave-one-out cross validation approach or use the same rule of thumb choice as in Remark 6.3.

Remark 6.7 (*Rule of thumb value of h*) Since Remark 6.3 effectively disregards the linear term in the bias (by assuming $df(x)/dx = 0$), it actually solves the same problem as for the local linear regression. The optimal h values is thus the same.

See Figures 6.12 – 6.13 for an empirical illustration.

Figure 6.14 shows a bootstrapped confidence band. The bootstrap simulations account for the (non-linear) autocorrelation (the returns are generated recursively using the estimated regression function) and the residuals have regressor-dependent heteroskedasticity. The latter is achieved by first estimating (by a non-parametric approach) how the squared fitted residuals depend on the regressor (lagged return). Then standardized fitted residuals are calculated. In the simulations, the fitted residuals are drawn (with replacement) and then scaled up by the regressor dependent volatility.

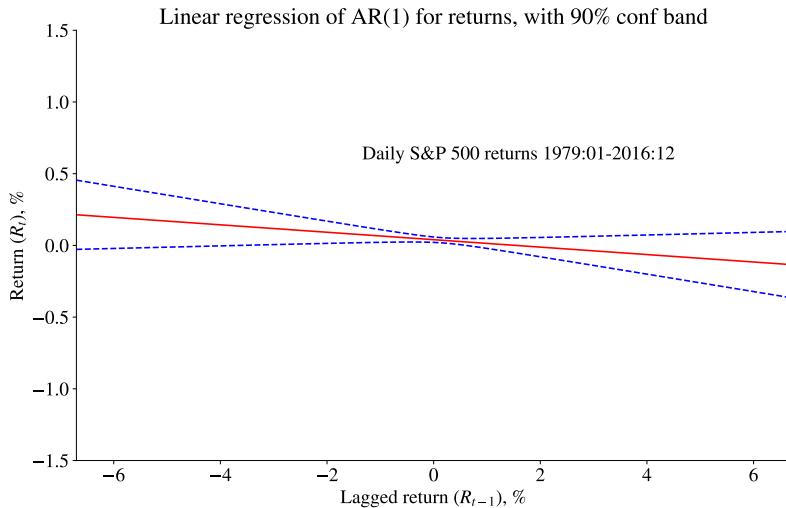


Figure 6.11: Linear regression with confidence bands

6.4 Applications of Kernel Regressions

6.4.1 “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.

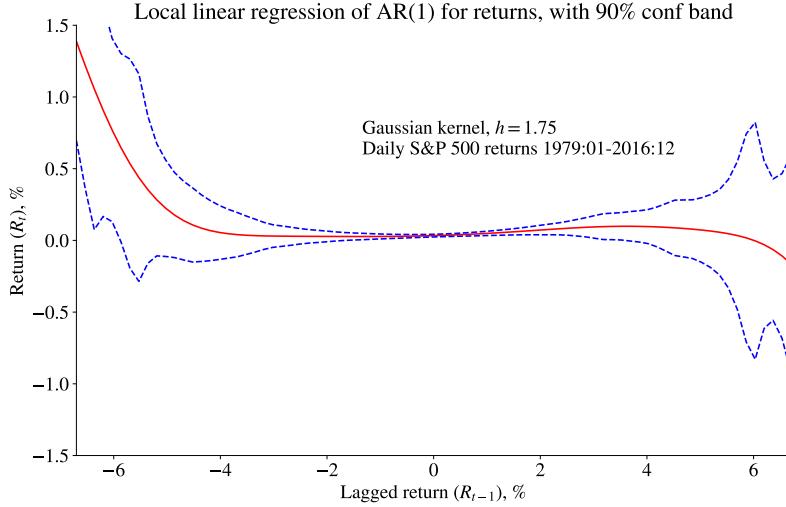


Figure 6.12: Non-parametric local linear regression with confidence bands

1. First, the call option price, H_{it} , is estimated as a multivariate kernel regression

$$H_{it} = b(S_t, X, \tau, r_{\tau t}, \delta_{\tau t}) + \varepsilon_{it}, \quad (6.18)$$

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 $\delta_{\tau t}$ separately.

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

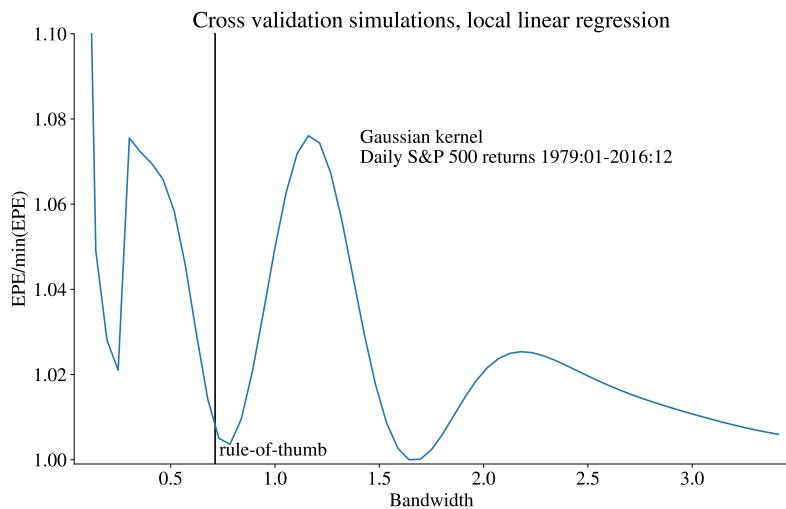


Figure 6.13: Cross-validation

their sample.

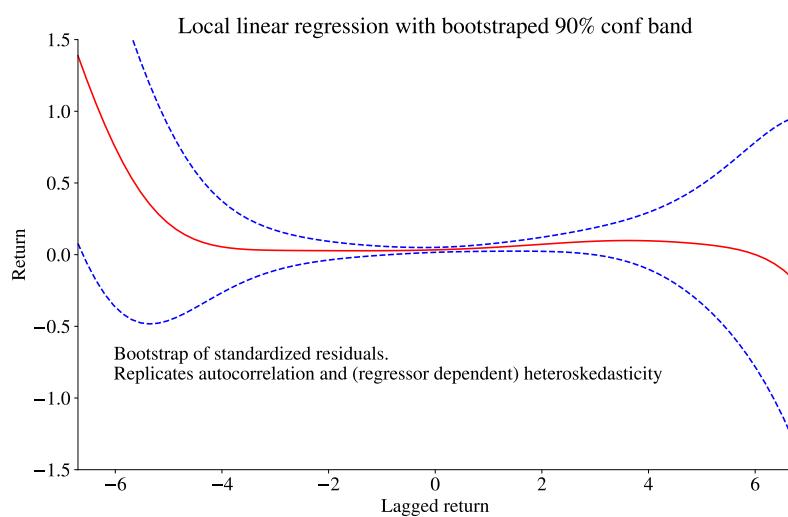


Figure 6.14: Non-parametric local linear regression with bootstrapped confidence bands

Chapter 7

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)

7.1 Heteroskedasticity

7.1.1 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 be

$$\sigma_t^2 = \frac{1}{q} \sum_{s=1}^q u_{t-s}^2 = (u_{t-1}^2 + u_{t-2}^2 + \dots + u_{t-q}^2)/q. \quad (7.1)$$

Notice that σ_t^2 depends on lagged information, and could therefore be thought of as the prediction (made in $t - 1$) of the volatility in t . Unfortunately, this method can produce quite abrupt changes in the estimate. See Figures 7.1–7.4 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. Let the weight for lag s be $(1 - \lambda)\lambda^s$ where $0 < \lambda < 1$, so

$$\sigma_t^2 = (1 - \lambda) \sum_{s=1}^{\infty} \lambda^{s-1} u_{t-s}^2 = (1 - \lambda)(u_{t-1}^2 + \lambda u_{t-2}^2 + \lambda^2 u_{t-3}^2 + \dots), \quad (7.2)$$

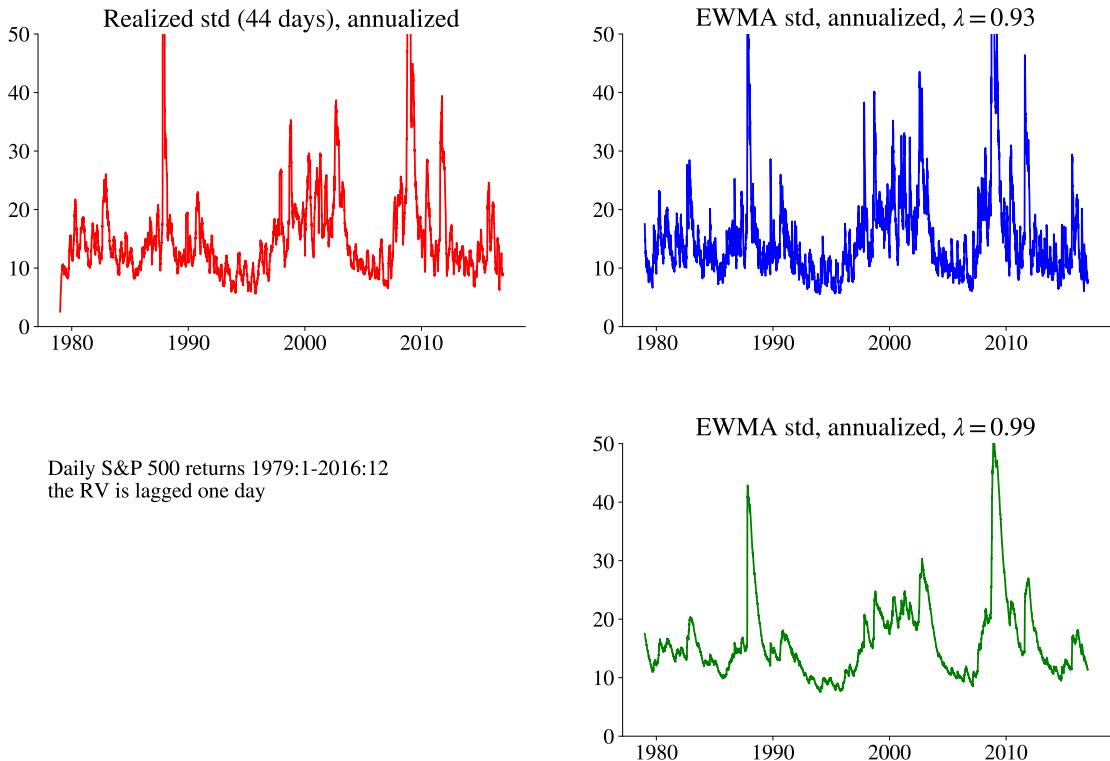


Figure 7.1: Standard deviation

which can also be calculated in a recursive fashion as

$$\sigma_t^2 = (1 - \lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2. \quad (7.3)$$

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 $\sum_{t=1}^T (u_t^2 - \sigma_t^2)^2$. See Figure 7.1 for an empirical example and 7.2 for an illustration of the weights. (They clearly sum to one.) It is straightforward to show that the weight for $t - s - 1$ where $s = -\ln(2)/\ln(\lambda)$ is half the weight for the first term ($t - 1$). For instance, with $\lambda = 0.94$ it takes 11 periods to halve the weight, but with $\lambda = 0.99$ it takes 68 periods.

Remark 7.1 (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).*

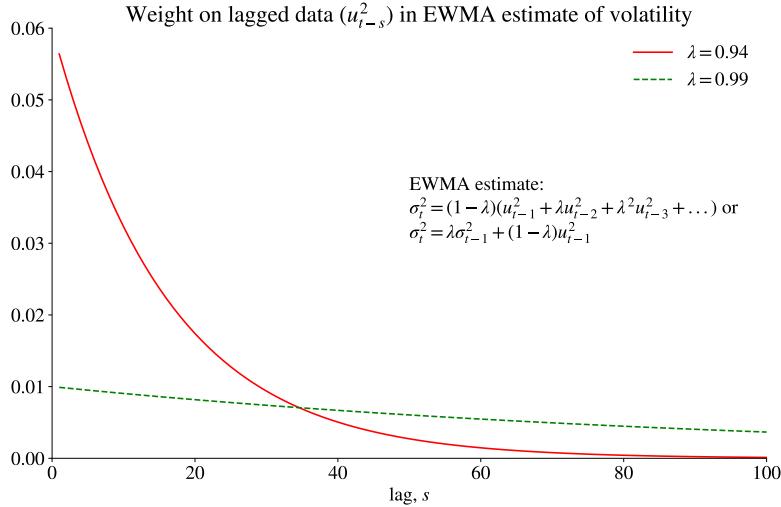


Figure 7.2: Weights on old data in the EWMA approach to estimate volatility

See Figure 7.5 for an example.

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

$$\sigma_{ij,t} = \frac{1}{q} \sum_{s=1}^q u_{i,t-s} u_{j,t-s}, \quad (7.4)$$

as well as the EWMA

$$\sigma_{ij,t} = (1 - \lambda) u_{i,t-1} u_{j,t-1} + \lambda \sigma_{ij,t-1}. \quad (7.5)$$

By combining with the estimates of the variances, it is straightforward to estimate correlations. See Figures 7.6–7.7 for illustrations.

7.1.2 Variance and Volatility Swaps

Many financial instruments are indirectly affected by volatility. This is particularly true for options. There are also instruments that are directly related to volatility, for instance, *variance swaps*. Such a contract has a zero price in inception (in t) and the payoff at expiration (in $t + m$) is

$$\text{Variance swap payoff}_{t+m} = \text{realized variance}_{t+m} - \text{variance swap rate}_t, \quad (7.6)$$

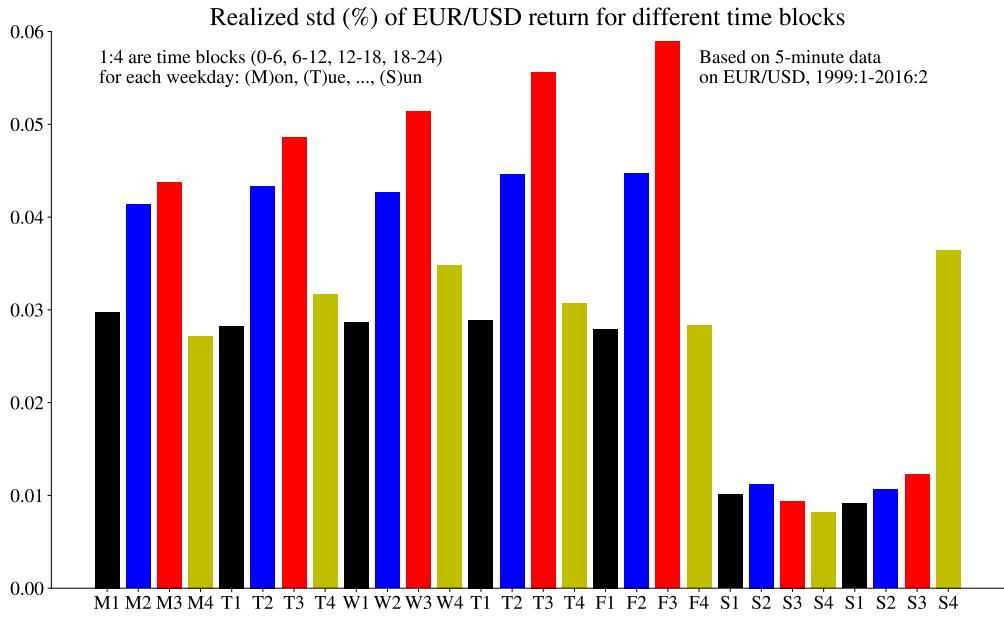


Figure 7.3: Standard deviation of EUR/USD exchange rate changes

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

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

$$RV_{t+m} = \frac{252}{m} \sum_{s=1}^m R_{t+s}^2, \quad (7.7)$$

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.)

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

$$\text{Volatility swap payoff}_{t+m} = \sqrt{\text{realized variance}_{t+m}} - \text{volatility swap rate}_t, \quad (7.8)$$

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 $t + m$) turns out to be higher than what was thought in when the futures contract was bought (here, t). In a way, we can think of the VIX futures

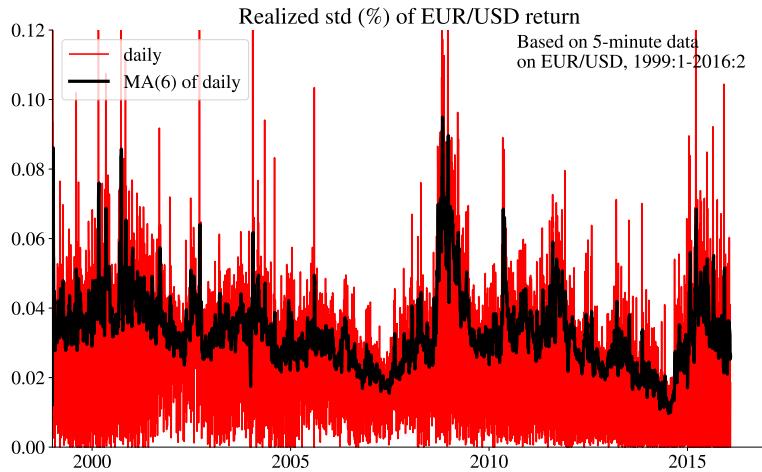


Figure 7.4: Standard deviation of exchange rate changes

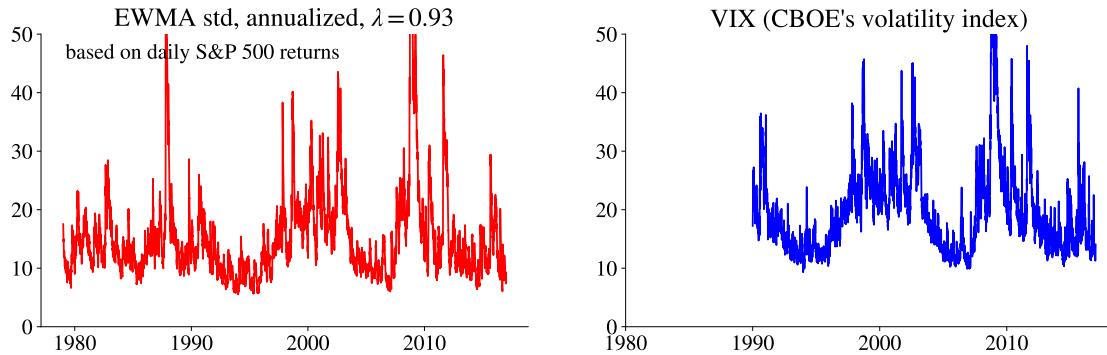


Figure 7.5: Different estimates of US equity market volatility

as a futures on a volatility swap (between $t + m$ and a month later).

Since VIX² is a good approximation of variance swap rate for a 30-day contract, the return can be approximated as

$$\text{Return of a variance swap}_{t+m} = (RV_{t+m} - VIX_t^2)/VIX_t^2. \quad (7.9)$$

Figures 7.8 and 7.9 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

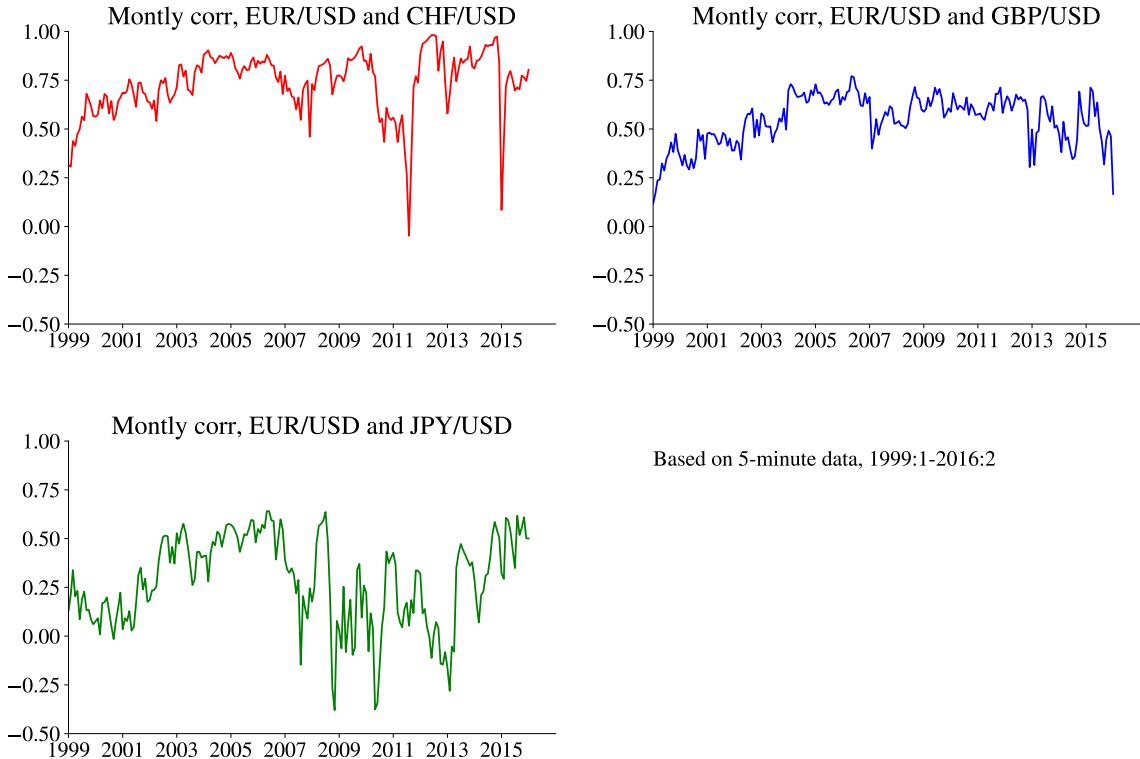


Figure 7.6: Correlation of exchange rate changes

swaps (which has been the specialty 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.

7.1.3 Forecasting Realized Volatility

Implied volatility from options (iv) should contain information about future volatility and 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 7.8.

Table 7.1 shows that the iv (here represented by VIX) is close to be an unbiased 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

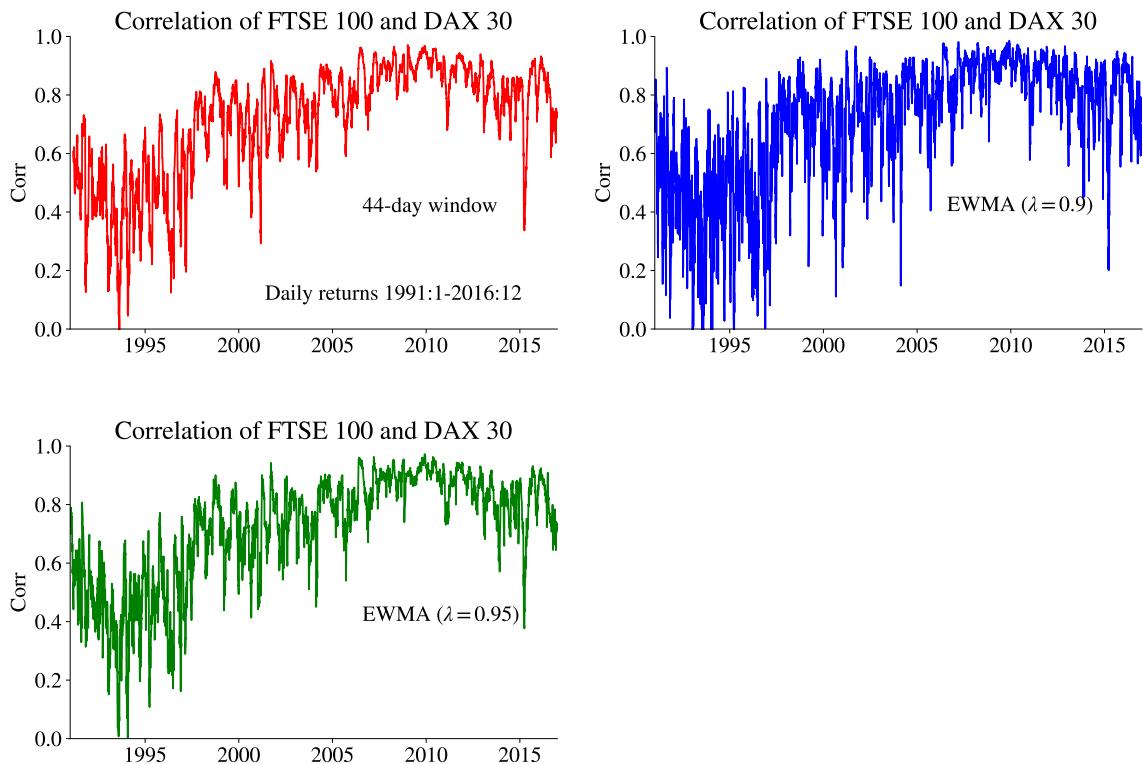


Figure 7.7: Time-varying correlations (realized and EWMA)

realized volatility, the forecast powers seems to improve.

Remark 7.2 (*Restricting the predicted volatility to be positive*) A linear regression (like those in Table 7.1) can produce negative volatility forecasts. An easy way to get around that is to specify the regression in terms on the log volatility.

Remark 7.3 (*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, $z = \ln[(1 + \rho)/(1 - \rho)]/2$, where ρ is the correlation coefficient. The correlation coefficient can then be calculated by the inverse transformation $\rho = [\exp(2z) - 1]/[\exp(2z) + 1]$.

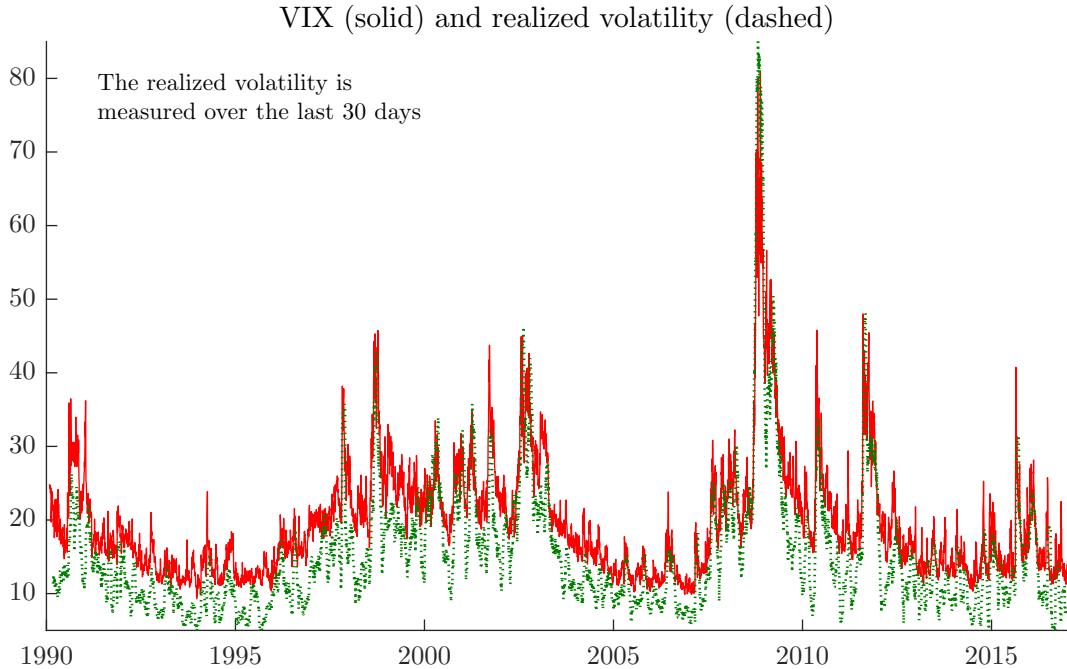


Figure 7.8: VIX and realized volatility (variance)

7.1.4 Heteroskedastic Residuals in a Regression

Suppose we have a regression model

$$y_t = x_t' b + u_t, \text{ where} \quad (7.10)$$

$$\mathbb{E} u_t = 0 \text{ and } \text{Cov}(x_{it}, u_t) = 0.$$

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

In case the residuals actually are heteroskedastic, least squares (LS) is nevertheless a useful estimator: it is still consistent (we get the correct values as the sample becomes really large)—and it is reasonably efficient (in terms of the variance of the estimates). However, the standard expression for the standard errors (of the coefficients) is (except in a special case, see below) not correct. This is illustrated in Table 7.4.

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

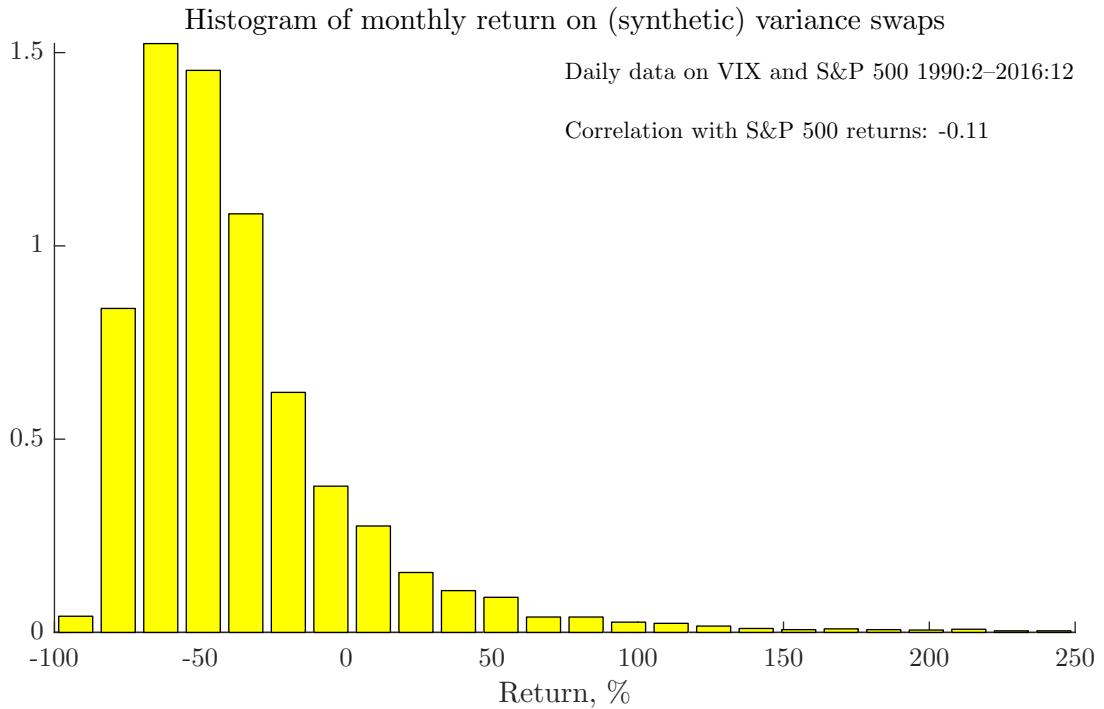


Figure 7.9: Distribution of return from investing in variance swaps

get the correct standard errors (provided the distribution used in the likelihood function is correct). Second, we could stick to OLS, but use another expression for the variance of the coefficients: a “heteroskedasticity consistent covariance matrix,” among which “White’s covariance matrix” is the most common.

To 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 below)—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, run a regression of squared fitted residuals on w_t

$$\hat{u}_t^2 = w_t' \gamma + v_t, \quad (7.11)$$

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

Example 7.4 (*White’s test*) If the regressors include $(1, x_{1t}, x_{2t})$ then w_t in (7.11) is the

	(1)	(2)	(3)
lagged RV	0.73 (10.25)	0.22 (1.91)	
lagged VIX		0.90 (12.13)	0.68 (8.94)
constant	4.12 (4.28)	-2.26 (-1.76)	-1.21 (-1.45)
R2	0.54	0.60	0.61
obs	6675.00	6695.00	6675.00

Table 7.1: Regression of 22-day realized S&P return volatility 1990:2–2016:12. All daily observations are used, so the residuals are likely to be autocorrelated. Numbers in parentheses are t-stats, based on Newey-West with 30 lags.

	Corr(EUR,CHF)	Corr(EUR,GBP)	Corr(EUR,JPY)
lagged Corr	0.83 (12.46)	0.81 (23.29)	0.84 (22.52)
constant	0.18 (2.89)	0.12 (5.06)	0.05 (3.01)
R2	0.70	0.67	0.71
obs	204.00	204.00	204.00

Table 7.2: Regression of monthly realized correlations 1999:1-2016:2. The Fisher transformation has been applied to all correlations. 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.

vector $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$.

If we reject the null hypothesis in White's test, then we either have to model both the regression equation and the volatility process simultaneously (for instance, using MLE) or adjust the OLS standard errors.

Remark 7.5 (White's covariance matrix) Recall that the sample moment conditions for OLS are

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

where we have k regressors in x_t . For the asymptotic distribution we need covariance matrix of $\sqrt{T}\bar{g}(\beta)$ and the Jacobian of the moment conditions with respect to the param-

	log RV(EUR)	log RV(GBP)	log RV(CHF)	log RV(JPY)
lagged log RV	0.77 (64.58)	0.68 (30.61)	0.77 (49.16)	0.72 (49.78)
constant	-0.56 (-26.34)	-0.62 (-22.15)	-0.57 (-20.98)	-0.59 (-24.58)
D(Tue)	0.41 (18.02)	0.42 (16.02)	0.37 (15.53)	0.36 (13.11)
D(Wed)	0.34 (15.62)	0.34 (13.87)	0.32 (14.73)	0.32 (12.35)
D(Thu)	0.35 (15.55)	0.38 (15.06)	0.27 (12.59)	0.29 (11.26)
D(Fri)	0.30 (11.43)	0.34 (12.25)	0.27 (10.71)	0.30 (9.98)
R2	0.60	0.47	0.60	0.53
obs	4454.00	4454.00	4454.00	4454.00

Table 7.3: Regression of daily log realized variance 1999:1-2016:2. 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.

eters (b). Let $u_t = y_t - x_t' b$ and notice that

$$S_0 = \text{Cov} \left[\frac{\sqrt{T}}{T} \sum_{t=1}^T x_t u_t \right] = \text{Cov}(x_t u_t),$$

where we have assumed that there is no autocorrelation, but we allow for heteroskedasticity since we are using the covariance matrix of $x_t u_t$ (not imposing that x_t and u_t are unrelated). We can estimate this as

$$\hat{S} = \frac{1}{T} \sum_{t=1}^T x_t x_t' \hat{u}_t^2.$$

Moreover, the Jacobian is

$$D_0 = \text{plim} \left(-\frac{1}{T} \sum_{t=1}^T x_t x_t' \right) = -\Sigma_{xx}.$$

Combining gives

$$\sqrt{T}(\hat{b} - b_0) \xrightarrow{d} N(0, \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1}).$$

In practice, we use \hat{S} instead of S_0 .

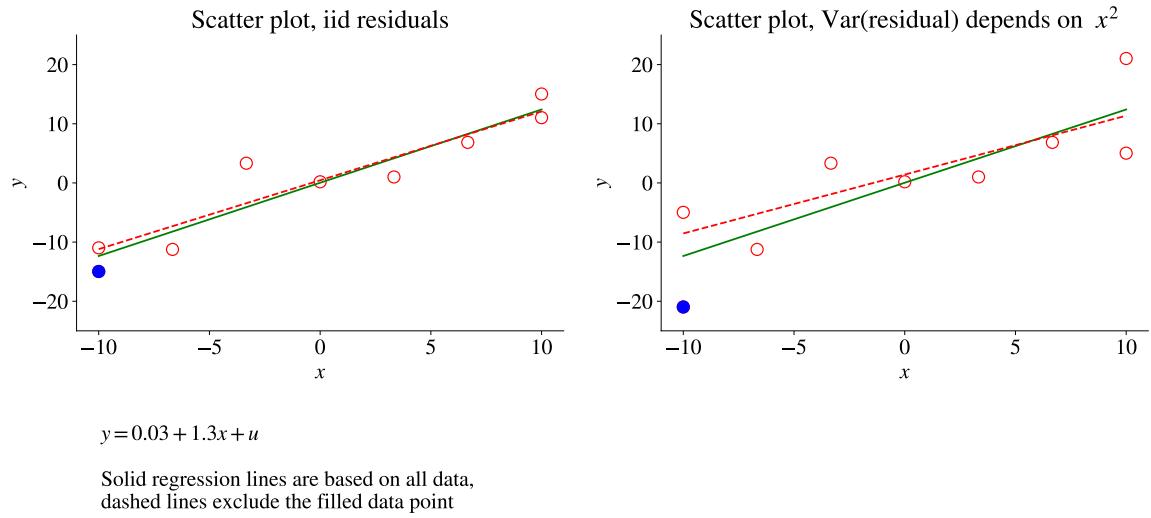


Figure 7.10: Effect of heteroskedasticity on uncertainty about regression line

7.1.5 Autoregressive Conditional Heteroskedasticity (ARCH)

Autoregressive heteroskedasticity is a special form of heteroskedasticity—and it is often found in financial data which shows volatility clustering.

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^2)

$$u_t^2 = \omega + a_1 u_{t-1}^2 + \dots + a_q u_{t-q}^2 + v_t, \quad (7.12)$$

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, $TR^2/(1-R^2) \sim \chi_q^2$.) This test can also be applied to the fitted residuals from a regression like (7.10). However, in this case, it is not obvious that ARCH effects makes the standard expression for the LS covariance matrix invalid (use White's test (7.11) for this).

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

$$\mathbb{E} \begin{bmatrix} u_{t-1}^2 \\ \vdots \\ u_{t-q}^2 \end{bmatrix} (u_t^2 - a_0) = \mathbf{0}_{q \times 1}. \quad (7.13)$$

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

Table 7.4: Standard error of OLS slope (Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1 + \gamma|z_t| + \alpha|x_t|)^2$, where z_t is iid $N(0, 1)$ and independent of x_t . Sample length: 200. Number of simulations: 25000. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

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.

7.2 ARCH Models

Consider the regression model

$$y_t = x_t' b + u_t, \text{ where} \quad (7.14)$$

$E u_t = 0$ and $\text{Cov}(x_{it}, u_t) = 0$.

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

7.2.1 Properties of ARCH(1)

In the ARCH(1) model the residual in the regression equation (7.14) can be written

$$u_t = v_t \sigma_t, \text{ with} \quad (7.15)$$

$v_t \sim \text{iid}$ with $E v_t = 0$ and $\text{Var}(v_t) = 1$,

and the conditional variance is generated by

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha u_{t-1}^2, \text{ with} \\ \omega &> 0 \text{ and } 0 \leq \alpha < 1.\end{aligned}\tag{7.16}$$

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 . 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. See Figure 7.11 for an illustration.

It is straightforward to show (proofs are in an Appendix) that the forecast (made in t) of volatility in $t+s$ is

$$E_t \sigma_{t+s}^2 = \bar{\sigma}^2 + \alpha^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1-\alpha},\tag{7.17}$$

where $\bar{\sigma}^2$ is the unconditional variance and we recall that σ_{t+1}^2 is known in t . The forecast of the variance is just like in an AR(1) process with α as the AR parameter. Also, the conditional variance of u_{t+s} is clearly equal to the expected value of σ_{t+s}^2

$$\text{Var}_t(u_{t+s}) = E_t \sigma_{t+s}^2.\tag{7.18}$$

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}^2 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

$$u_{t+2} = v_{t+2} \sigma_{t+2} = v_{t+2} \sqrt{\omega + \alpha v_{t+1}^2 \sigma_{t+1}^2},\tag{7.19}$$

which is a nonlinear function of v_{t+2} and v_{t+1} (which are standard normal) and it depends on σ_{t+2} which is not known in t . This gives u_{t+2} 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

$$\frac{E u_t^4}{(E u_t^2)^2} = \begin{cases} 3 \frac{1-\alpha^2}{1-3\alpha^2} \geq 3 & \text{if denominator is positive} \\ \infty & \text{otherwise.} \end{cases}\tag{7.20}$$

As a comparison, the kurtosis of a normal distribution is 3 (you also get this by setting

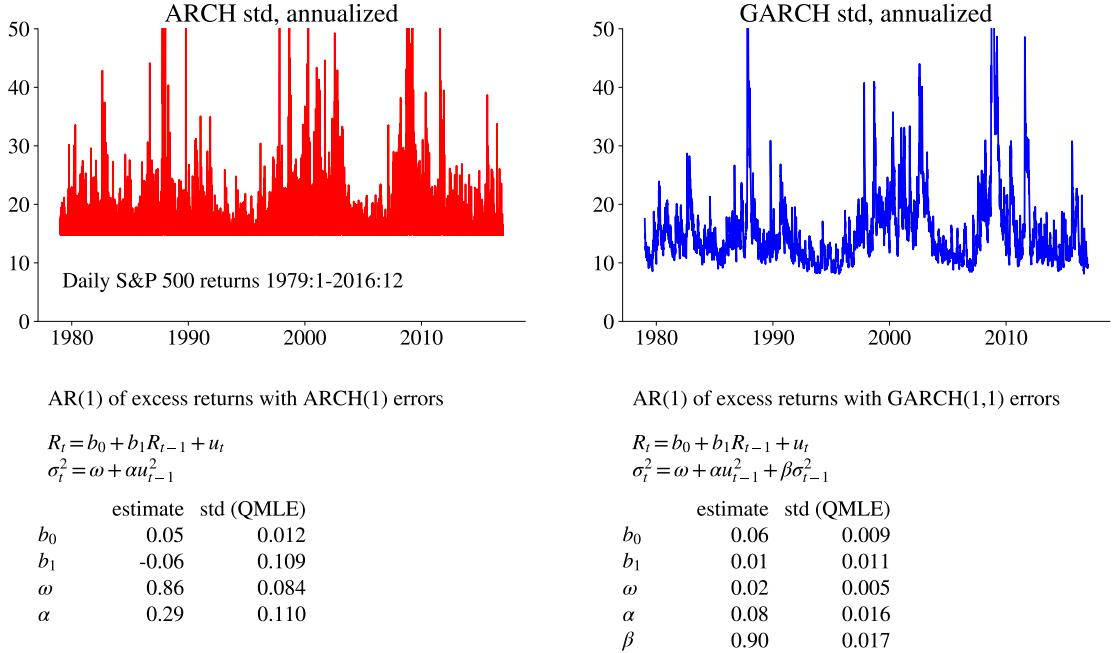


Figure 7.11: ARCH and GARCH estimates

$\alpha = 0$). This means that we can expect u_t to have fat tails, but that the standardized residuals u_t/σ_t perhaps look more normally distributed. See Figure 7.13 for an illustration (although based on a GARCH model).

Example 7.6 (Kurtosis) With $\alpha = 1/3$, the kurtosis is 4, at $\alpha = 0.5$ it is 9 and at $\alpha = 0.6$ it is infinite. With $\alpha = 0$, it is 3.

7.2.2 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 (7.14)–(7.16) by ML or GMM.

Remark 7.7 (*Likelihood function of $u_t \sim N(\mu, \sigma^2)$*) The pdf of an $u_t \sim N(0, \sigma^2)$ is

$$pdf(x_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{u_t^2}{\sigma^2}\right),$$

so the log-likelihood is

$$\ln \mathcal{L}_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^2 - \frac{1}{2} \frac{u_t^2}{\sigma^2}.$$

If u_t and u_s are independent (uncorrelated if jointly 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.

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

$$\begin{aligned} \ln \mathcal{L} &= -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^T \ln \sigma_t^2 - \frac{1}{2} \sum_{t=1}^T \frac{u_t^2}{\sigma_t^2}, \text{ if} \\ v_t &\text{ is iid } N(0, 1). \end{aligned} \quad (7.21)$$

By plugging in (7.14) for u_t and (7.16) 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 (7.16). The MLE is consistent.

Remark 7.8 (Coding the ARCH(1) ML estimation) A straightforward way of coding the estimation problem (7.14)–(7.16) and (7.21) 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 (7.14), 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 (7.14).

Second, loop over the sample (first $t = 1$, then $t = 2$, etc.) and calculate \hat{u}_t from (7.14) and σ_t^2 from (7.16). Plug in these numbers in (7.21) 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 7.9 (Imposing parameter constraints on ARCH(1)) To impose the restrictions in (7.16), iterate over values of $(b, \tilde{\omega}, \tilde{\alpha})$, but use $\omega = \tilde{\omega}^2$ and $\alpha = \exp(\tilde{\alpha})/[1 + \exp(\tilde{\alpha})]$ inside the likelihood function. (It is sometimes useful to notice that the inverse is $\tilde{\alpha} = \ln[\alpha/(1 - \alpha)]$).

It is often found that the fitted normalized residuals, \hat{u}_t/σ_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 a 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

$$\mathbb{E} \begin{bmatrix} x_t u_t \\ u_t^2 - \sigma_t^2 \\ u_{t-1}^2 (u_t^2 - \sigma_t^2) \end{bmatrix} = \mathbf{0}_{(k+2) \times 1}, \quad (7.22)$$

where u_t and σ_t^2 are given by (7.14) and (7.16).

It is straightforward to add more lags to (7.16). For instance, an ARCH(p) would be

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \dots + \alpha_p u_{t-p}^2. \quad (7.23)$$

We then have to add more moment conditions to (7.22), 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 $\sum_{j=1}^p \alpha_j \leq 1$.

7.3 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

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2, \text{ with} \\ \omega &> 0; \alpha, \beta \geq 0; \text{ and } \alpha + \beta < 1, \end{aligned} \quad (7.24)$$

combined with (7.14) and (7.15). See Figure 7.11 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.

The forecast (made in t) of the future conditional variance (σ_{t+s}^2) is

$$\mathbb{E}_t \sigma_{t+s}^2 = \bar{\sigma}^2 + (\alpha + \beta)^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2), \text{ with } \bar{\sigma}^2 = \frac{\omega}{1 - \alpha - \beta}, \quad (7.25)$$

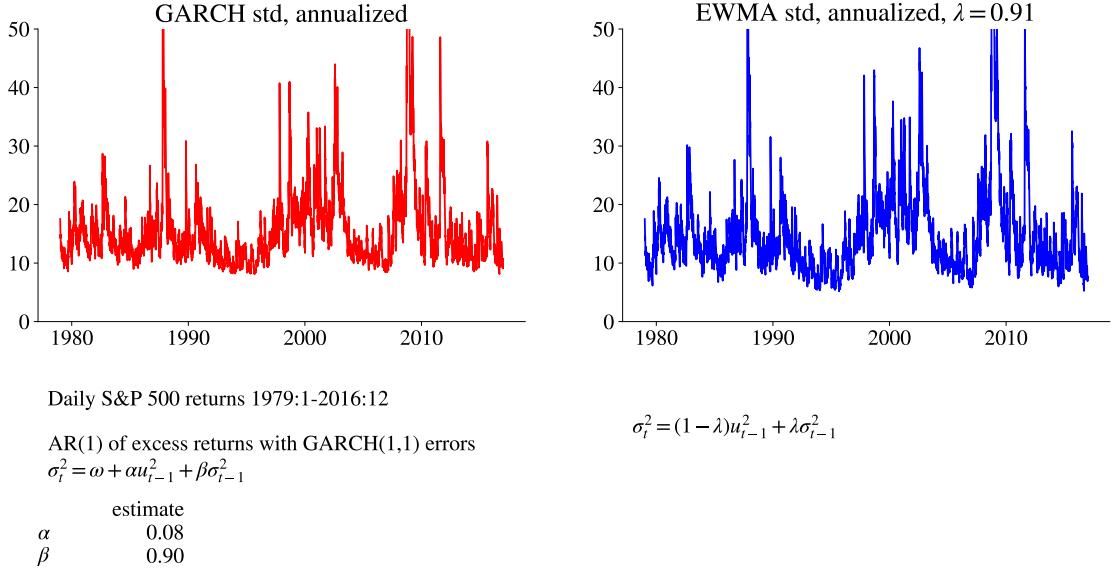


Figure 7.12: Conditional standard deviation, estimated by GARCH(1,1) model

where $\bar{\sigma}^2$ is the unconditional variance. (Recall that σ_{t+1}^2 is known in t .) This has the same form as in the ARCH(1) model (7.17), but where the sum of α and β is like an AR(1) parameter. As for the ARCH model, the conditional variance of u_{t+s} is equal to the expected value of σ_{t+s}^2

$$\text{Var}_t(u_{t+s}) = E_t \sigma_{t+s}^2. \quad (7.26)$$

Assuming that u_t has no autocorrelation, it follows directly from (7.25) that the expected variance of a longer time period ($u_{t+1} + u_{t+2} + \dots + u_{t+K}$) is

$$\begin{aligned} \text{Var}_t(\sum_{s=1}^K u_{t+s}) &= E_t \sum_{s=1}^K \sigma_{t+s}^2 \\ &= K\bar{\sigma}^2 + \frac{1 - (\alpha + \beta)^K}{1 - (\alpha + \beta)} (\sigma_{t+1}^2 - \bar{\sigma}^2). \end{aligned} \quad (7.27)$$

This is useful for portfolio choice and asset pricing when the horizon is longer than one period (day, perhaps).

See Figures 7.12–7.13 for illustrations.

Remark 7.10 (EWMA) *The GARCH(1,1) has many similarities with the exponential moving average estimator of volatility*

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

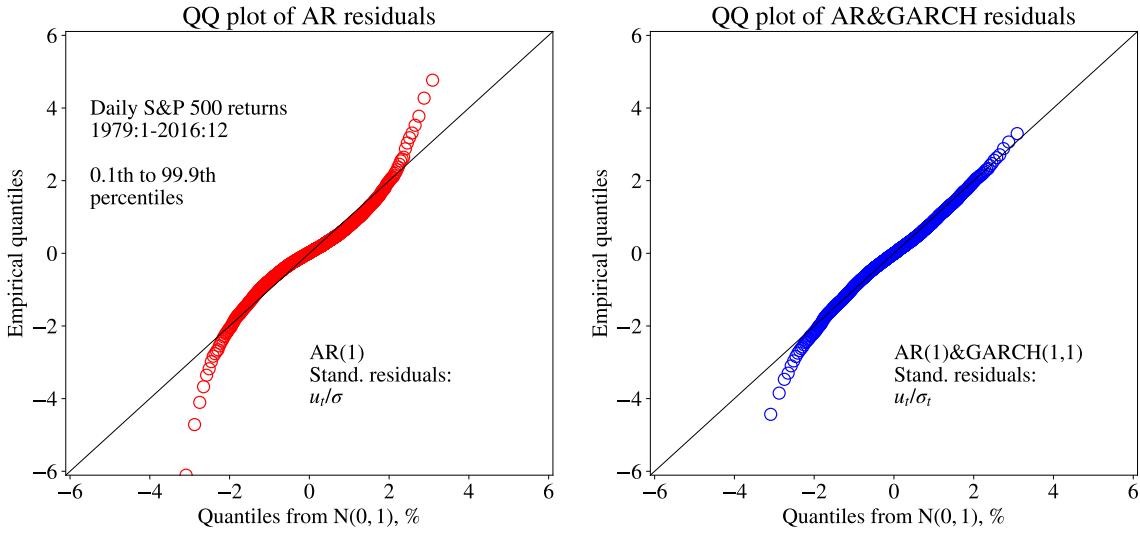


Figure 7.13: QQ-plot of residuals

This methods is commonly used by practitioners. For instance, the RISK Metrics uses this method with $\lambda = 0.94$ for daily data. Clearly, λ plays the same type of role as β in (7.24) 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 7.12 for a comparison.

The kurtosis of the process is

$$\frac{\mathbb{E} u_t^4}{(\mathbb{E} u_t^2)^2} = \begin{cases} 3 \frac{1-(\alpha+\beta)^2}{1-2\alpha^2-(\alpha+\beta)^2} \geq 3 & \text{if denominator is positive} \\ \infty & \text{otherwise.} \end{cases} \quad (7.28)$$

If $\alpha = 0$, then the variance becomes deterministic and the distribution normal (with a kurtosis of 3).

The GARCH(1,1) corresponds to an ARCH(∞) with geometrically declining weights, which is seen by solving (7.24) recursively by substituting for σ_{t-1}^2 (and then σ_{t-2}^2 , σ_{t-3}^2 , ...)

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha \sum_{j=0}^{\infty} \beta^j u_{t-1-j}^2. \quad (7.29)$$

This suggests that a GARCH(1,1) might be a reasonable approximation of a high-order ARCH.

To estimate the model consisting of (7.14), (7.15) and (7.24) we can still use the

likelihood function (7.21) 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 (7.14). It is also possible to assume another distribution than $N(0, 1)$.

Remark 7.11 (*Imposing parameter constraints on GARCH(1,1)*) To impose the restrictions in (7.24), 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 (7.24))

$$E \begin{bmatrix} x_t u_t \\ u_t^2 - \sigma_t^2 \\ u_{t-1}^2(u_t^2 - \sigma_t^2) \\ u_{t-2}^2(u_t^2 - \sigma_t^2) \end{bmatrix} = \mathbf{0}_{(k+3) \times 1}, \text{ where } u_t = y_t - x_t' b. \quad (7.30)$$

7.4 Value at Risk

The value at risk (as fraction of the investment) at the α level (say, $\alpha = 0.95$) is $\text{VaR}_\alpha = -\text{cdf}^{-1}(1 - \alpha)$, where $\text{cdf}^{-1}()$ is the inverse of the cdf. This means that $\text{cdf}^{-1}(1 - \alpha)$ is the $1 - \alpha$ quantile of the return distribution. See Figure 7.14 for an illustration. When the return has an $N(\mu, \sigma^2)$ distribution, then $\text{VaR}_{95\%} = -(\mu - 1.64\sigma)$. See Figures 7.15–7.17 for an example of time-varying VaR, based on a GARCH model.

7.4.1 Backtesting a VaR Model

Backtesting a VaR model amounts to checking if (historical) data fits with the VaR numbers. For instance, we first find the $\text{VaR}_{95\%}$ and then calculate what fraction of returns that is actually below (the negative of) this number. If the model is correct it should be 5%. We then repeat this for $\text{VaR}_{96\%}$: only 4% of the returns should be below (the negative of) this number. Figures 7.16–7.17 show results from backtesting a VaR model where the volatility follows a GARCH process (to capture the time varying volatility) with normally distributed shocks. The evidence suggests that this model works relatively well except for very high confidence levels.

Remark 7.12 (*Bernoulli and binomial distributions*) In a Bernoulli distribution, the random variable X can only take two values: 1 or 0, with probability p and $1-p$ respectively.

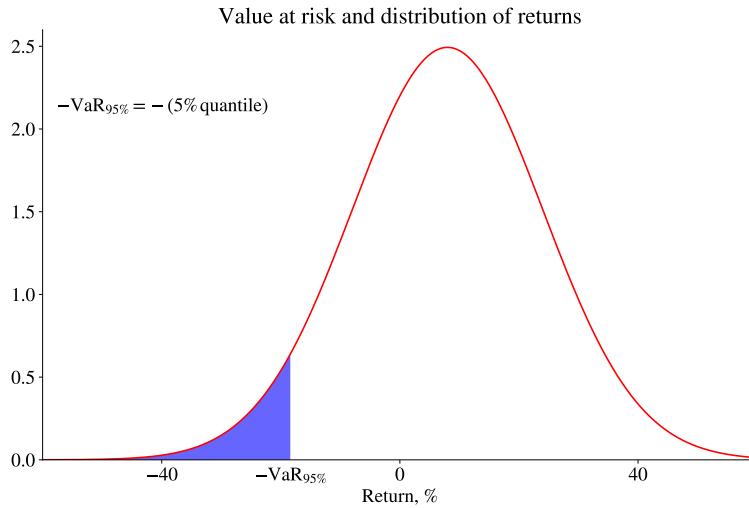


Figure 7.14: Value at risk

This gives $E(X) = p$ and $\text{Var}(X) = p(1 - p)$. After n independent trials, the number of successes (y) has a binomial distribution with $E(y) = np$ and $\text{Var}(y) = np(1 - p)$.

To perform a statistical back test of a VaR model, define a variable that is one if the loss is greater than the VaR

$$d_t = \begin{cases} 1 & \text{if } R_t < -\text{VaR}_\alpha \\ 0 & \text{otherwise} \end{cases} \quad (7.31)$$

By using the properties of a binomial distribution, we can notice that

$$\sum_{t=1}^T d_t / T \xrightarrow{d} N[1 - \alpha, \alpha(1 - \alpha)/T]. \quad (7.32)$$

(This follows from $\sum_{t=1}^T d_t \xrightarrow{d} N[(1 - \alpha)T, T\alpha(1 - \alpha)]$.) This expression can be used for testing the null hypothesis that $R_t < -\text{VaR}_\alpha$ for the fraction $1 - \alpha$ of the observations. See Figure 7.17.

Remark 7.13 (GMM for backtesting the VaR) Alternatively, we could use GMM with the moment condition

$$g_t = d_t - (1 - \alpha) = 0.$$

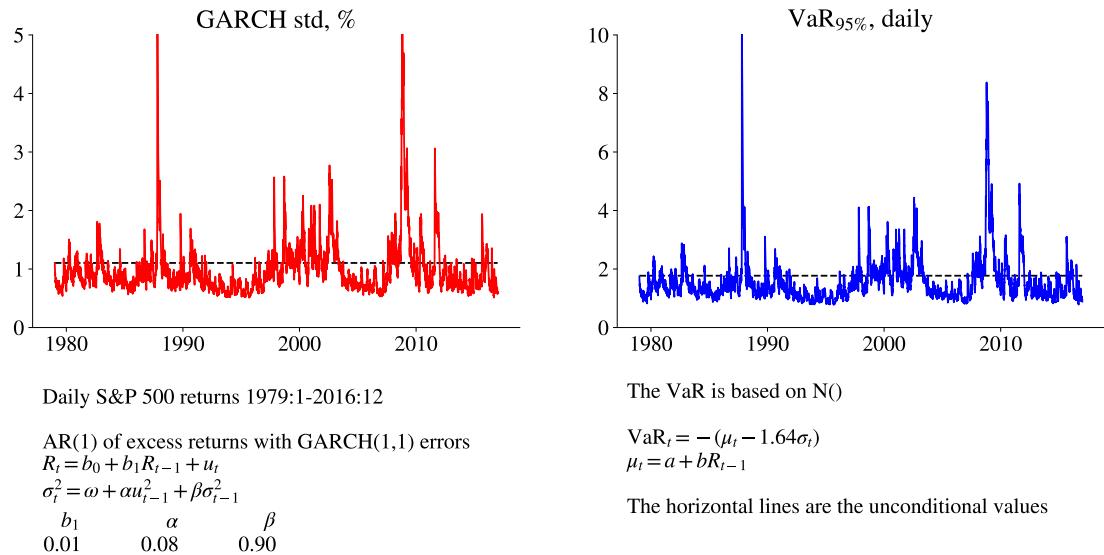


Figure 7.15: Conditional volatility and VaR

From the usual properties of GMM, we then have

$$\begin{aligned}\sqrt{T}\bar{g} &\xrightarrow{d} N(0, S_0), \text{ so} \\ \sum_{t=1}^T d_t/T &\xrightarrow{d} N(1-\alpha, S_0/T),\end{aligned}$$

where $\bar{g} = \sum_{t=1}^T g_t/T$ is the average moment condition and $S_0 = \text{Var}(\sqrt{T}\bar{g})$ is the variance. The latter can be estimated by, for instance, a Newey-West approach. Clearly, the only difference between (7.32) and the GMM approach is that the former specifies the variance as $\alpha(1-\alpha)/T$, while the latter open up the possibility to use another way of estimating the variance. In most application the results tend to be similar.

It is also important to see if there are medium- to long-run deviations from the VaR confidence level. Figure 7.18 illustrates the importance of using a dynamic VaR to capture the swings in uncertainty.

7.5 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.

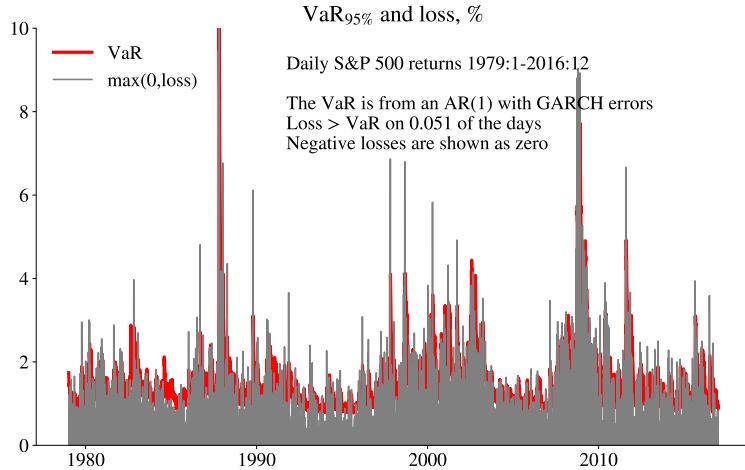


Figure 7.16: Backtesting VaR from a GARCH model, assuming normally distributed shocks

An asymmetric GARCH (Glosten, Jagannathan, and Runkle (1993)) can be constructed as

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \delta(u_{t-1} > 0) u_{t-1}^2, \text{ where} \quad (7.33)$$

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

This means that the effect of the shock u_{t-1}^2 is α if the shock was negative and $\alpha + \gamma$ if the shock was positive. With $\gamma < 0$, volatility increases more in response to a negative u_{t-1} (“bad news”) than to a positive u_{t-1} .

The EGARCH (exponential GARCH, Nelson (1991)) sets

$$\ln \sigma_t^2 = \omega + \alpha \frac{|u_{t-1}|}{\sigma_{t-1}} + \beta \ln \sigma_{t-1}^2 + \gamma \frac{u_{t-1}}{\sigma_{t-1}}. \quad (7.34)$$

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 log volatility in the same way. (Although this gives a small asymmetry for the volatility level, the effect is typically small.) The linear term in u_{t-1} modifies this to make the effect asymmetric. In particular, if $\gamma < 0$, then the log volatility increases more in response to a negative u_{t-1} (“bad news”) than to a positive u_{t-1} . This model is stationary if $|\beta| < 1$.

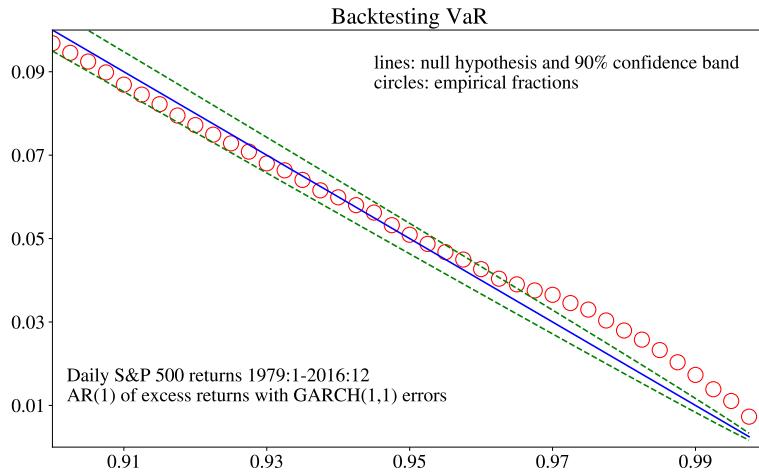


Figure 7.17: Backtesting VaR from a GARCH model, assuming normally distributed shocks

To estimate the model, we can still use the likelihood function (7.21) and do a MLE. We typically create the starting value of $u_0 = y_0 - x'_0 \hat{b}$ and $\sigma_0^2 = \text{Var}(\hat{u}_t)$, where $\hat{u}_t = y_t - x'_t \hat{b}$. See Figure 7.19 for an illustration.

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 $|u_{t-1} - b|^{3/2}$.

7.6 GARCH Models with Exogenous Variables

We could easily extend the GARCH(1,1) model by adding exogenous variables x_{t-1} (for instance, VIX)

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma x_{t-1}, \quad (7.35)$$

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

$$\ln \sigma_t^2 = \omega + \alpha \frac{|u_{t-1}|}{\sigma_{t-1}} + \beta \ln \sigma_{t-1}^2 + \gamma x_{t-1}. \quad (7.36)$$

These models can be estimated with maximum likelihood.

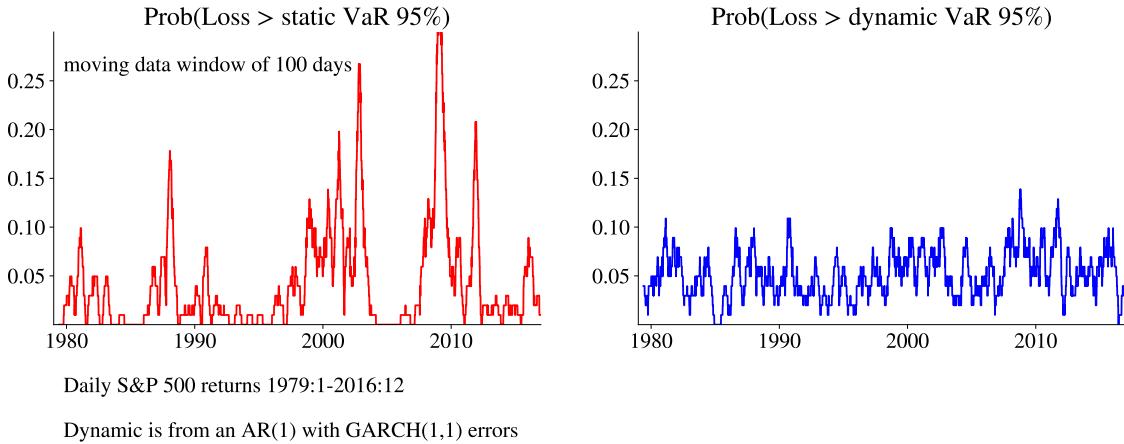


Figure 7.18: Backtesting VaR from a GARCH model on a moving data window, assuming normally distributed shocks

7.7 Stochastic Volatility Models

A *stochastic volatility model* differs from GARCH models by making the volatility truly stochastic. In contrast, in a GARCH model, the volatility in period t (σ_t) is known already in $t - 1$. This difference is important if we want to model a volatility risk premium in derivatives.

In a stochastic volatility model the log volatility may follows a ARMA process, for instance, an AR(1)

$$\begin{aligned} \ln \sigma_t^2 &= \omega + \beta \ln \sigma_{t-1}^2 + \theta \eta_t, \\ \text{with } \eta_t &\sim \text{iid } N(0, 1), \end{aligned} \tag{7.37}$$

combined with (7.14) and (7.15).

The estimation of a stochastic volatility model is complicated—and the basic reason is that it is 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 (7.15) and (7.37), this could be done as follows: square (7.15) and take logs to get

$$\ln u_t^2 = E \ln v_t^2 + \ln \sigma_t^2 + (\ln v_t^2 - E \ln v_t^2). \tag{7.38}$$

We could use this as the measurement equation in a Kalman filter (pretending that $\ln v_t^2 -$

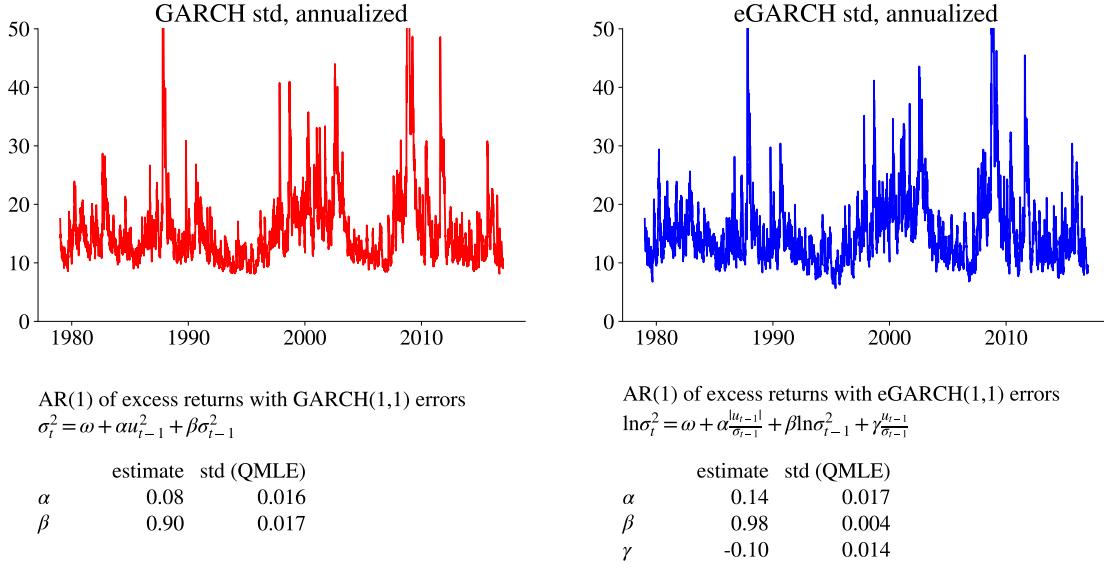


Figure 7.19: GARCH and EGARCH estimates

$E \ln v_t^2$ is normally distributed), and (7.37) 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.”

Remark 7.14 (Kalman filter) *In a Kalman filter, the state variables α_t are typically modelled as a VAR system*

$$\alpha_t = A\alpha_{t-1} + u_t,$$

where u_t are shocks. We do not observe all states variables directly, only some vector y_t , which is related to the state variables through the measurement equation

$$y_t = B\alpha_t + \epsilon_t,$$

where ϵ_t are some other shocks. The basic idea of the filter is that observing y_t gives a hint about α_t , which we can use to estimate also how the state evolves over time (the parameters in A).

If $\ln v_t^2$ is normally distributed , then this will give MLE, otherwise just a quasi-MLE. For instance, if v_t is iid $N(0, 1)$ (see Ruiz (1994)) then we have approximately $E \ln v_t^2 \approx -1.27$ and $\text{Var}(\ln v_t^2) = \pi^2/2$ (with $\pi = 3.14\dots$) so we could write the measurement

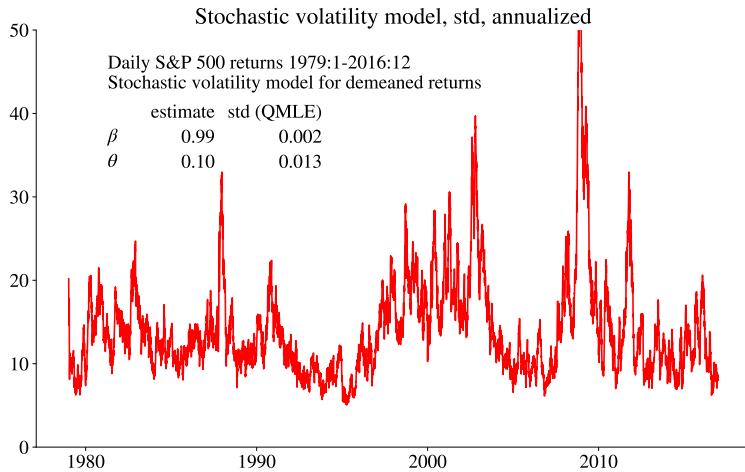


Figure 7.20: Conditional standard deviation, stochastic volatility model

equation as

$$\begin{aligned} \ln u_t^2 &= -1.27 + \ln \sigma_t^2 + w_t, \text{ with} \\ w_t &\sim N(0, \pi^2/2). \end{aligned} \quad (7.39)$$

In this case, only the state equation contains parameters that we need to estimate: ω, β, θ . See Figure 7.20 for an example.

7.8 (G)ARCH-M

It can make sense to let the conditional volatility enter the regression (“mean”) equation—for instance, as a proxy for risk which may influence the expected return.

Example 7.15 (*Mean-variance portfolio choice*) A mean variance investor solves

$$\begin{aligned} \max_{\alpha} \mathbb{E} R_p - \sigma_p^2 k / 2, \\ \text{subject to } R_p = \alpha R_m + (1 - \alpha) R_f, \end{aligned}$$

where R_m is the return on the risky asset (the market index) and R_f is the riskfree return. The solution is

$$\alpha = \frac{1}{k} \frac{\mathbb{E}(R_m - R_f)}{\sigma_m^2}.$$

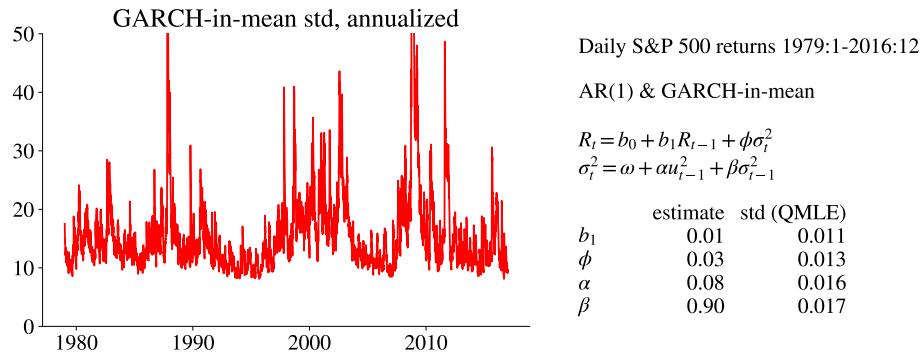


Figure 7.21: GARCH-M example

In equilibrium, this weight is one (since the net supply of bonds is zero), so we get

$$E(R_m - R_f) = k \sigma_m^2,$$

which says that the expected excess return is increasing in both the market volatility and risk aversion (k).

We modify the “mean equation” (7.14) 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

$$y_t = x'_t b + \varphi \sigma_t^2 + u_t, \quad E(u_t | x_t, \sigma_t) = 0. \quad (7.40)$$

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 (7.21) to do MLE.

It can also be noted (see Gouriéroux 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 7.21 for an example.

Remark 7.16 (Coding of (G)ARCH-M) We can use the same approach as in Remark 7.8, except that we use (7.40) instead of (7.14) to calculate the residuals (and that we obviously also need a guess of φ).

7.9 Multivariate (G)ARCH

7.9.1 Different Multivariate Models

This section gives a brief summary of some multivariate models of heteroskedasticity. Let the model (7.14) 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

$$\Sigma_t = E_{t-1} u_t u_t'. \quad (7.41)$$

The models discussed below differ with respect to the dynamics of the Σ_t matrix (which also plays a key role in the formulation of the likelihood function used for estimation).

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, $\text{vech}(\Sigma_t)$, this is still $n(n + 1)/2$ elements

$$\text{vech}(\Sigma_t) = C + A\text{vech}(u_{t-1} u'_{t-1}) + B\text{vech}(\Sigma_{t-1}). \quad (7.42)$$

This typically gives too many parameters to handle—and makes it difficult to impose sufficient restrictions to make Σ_t positive definite (compare the restrictions of positive coefficients in (7.24)).

Example 7.17 (*vech formulation, $n = 2$*) For instance, with $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = C + A \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1} u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + B \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix},$$

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 7.18 (*Diagonal model, n = 2*) With $n = 2$ we have

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{21,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} u_{1,t-1}^2 \\ u_{1,t-1}u_{2,t-1} \\ u_{2,t-1}^2 \end{bmatrix} + \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{21,t-1} \\ \sigma_{22,t-1} \end{bmatrix},$$

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

$$\Sigma_t = C + A'u_{t-1}u'_{t-1}A + B'\Sigma_{t-1}B, \quad (7.43)$$

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 $\text{vech}(\Sigma_t)$. (Recall that a quadratic form is positive definite, provided the matrices are of full rank.)

Example 7.19 (*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 7.20 (*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 (7.24) gives 7 parameters (2×3 GARCH parameters and one correlation), which is convenient.

Remark 7.21 (Imposing parameter constraints on a correlation) To impose the restriction that $-1 < \rho < 1$, iterate over $\tilde{\rho}$ and let $\rho = 1 - 2/[1 + \exp(\tilde{\rho})]$. (It is sometimes useful to notice that the inverse is $\tilde{\rho} = \ln[(1 + \rho)/(1 - \rho)]$.)

Remark 7.22 (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* (DCC) discussed in Engle (2002)) allows the correlation to change over time. 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)

$$\Sigma_t = D_t R_t D_t, \text{ with } D_t = \text{diag}(\sqrt{\sigma_{ii,t}}), \quad (7.44)$$

and R_t is the conditional correlation matrix, which is modelled below.

Remark 7.23 (*diag*(a_i) notation) *diag*(a_i) denotes the $n \times n$ matrix with elements a_1, a_2, \dots, a_n along the main diagonal and zeros elsewhere. For instance, if $n = 2$, then

$$\text{diag}(a_i) = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}.$$

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 we assume a GARCH(1,1) structure for the correlation matrix, then we have

$$Q_t = (1 - \alpha - \beta)\bar{Q} + \alpha v_{t-1} v'_{t-1} + \beta Q_{t-1}, \text{ with} \quad (7.45)$$

$$v_{i,t} = u_{i,t}/\sqrt{\sigma_{ii,t}},$$

where α and β are two *scalars* and \bar{Q} is the unconditional covariance matrix of the normalized residuals (v_t). We require that $\alpha \geq 0$ and that $\alpha + \beta < 1$.

To guarantee that the conditional correlation matrix is indeed a correlation matrix, Q_t is treated as if it were a covariance matrix and R_t is simply the implied correlation matrix. That is,

$$R_t = \text{diag}(\sqrt{q_{ii,t}})^{-1} Q_t \text{diag}(\sqrt{q_{ii,t}})^{-1}. \quad (7.46)$$

The basic idea of this model is to estimate a conditional correlation matrix as in (7.46) and then scale up with conditional variances (from univariate GARCH models) to get a conditional covariance matrix as in (7.44).

See Figure 7.22 for an illustration—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 7.24 (*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 (7.24) gives 9 parameters (2 \times 3 GARCH parameters, $(\bar{q}_{12}, \alpha, \beta)$).

To see what DCC generates, consider the correlation coefficient from a bivariate

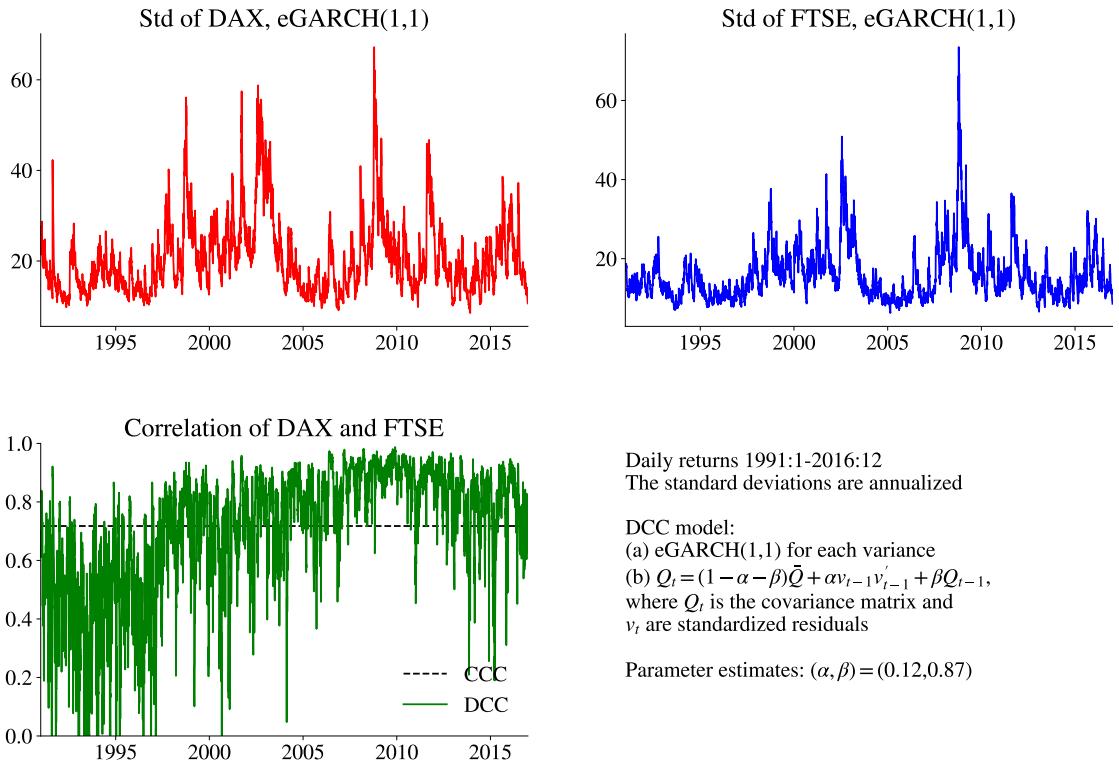


Figure 7.22: Results for multivariate eGARCH models

model

$$\rho_{12,t} = \frac{q_{12,t}}{\sqrt{q_{11,t}} \sqrt{q_{22,t}}}, \text{ where} \quad (7.47)$$

$$q_{12,t} = (1 - \alpha - \beta)\bar{q}_{12} + \alpha v_{1,t-1}v_{2,t-1} + \beta q_{12,t-1}$$

$$q_{11,t} = (1 - \alpha - \beta) + \alpha v_{1,t-1}v_{1,t-1} + \beta q_{11,t-1}$$

$$q_{22,t} = (1 - \alpha - \beta) + \alpha v_{2,t-1}v_{2,t-1} + \beta q_{22,t-1}.$$

This is a complicated expression, but 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.

7.9.2 Estimation of a Multivariate Model

In principle, it is straightforward to specify the likelihood function of the model and then maximize it with respect to the model parameters. For instance, if u_t is iid $N(0, \Sigma_t)$, then the log likelihood function is

$$\ln \mathcal{L} = -\frac{Tn}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^T \ln |\Sigma_t| - \frac{1}{2} \sum_{t=1}^T u_t' \Sigma_t^{-1} u_t. \quad (7.48)$$

In practice, the optimization problem can be difficult since there are typically many parameters. At least, good starting values are required.

Remark 7.25 (*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.

Remark 7.26 (*Estimation of the dynamic correlation model*) The DCC model can be estimated 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 (7.48 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.

7.10 LAD and Quantile Regressions*

Quantile regressions are useful for estimating models where the heteroskedasticity is related to the regressors.

7.10.1 LAD

Reference: Amemiya (1985) 4.6, Greene (2012) 7.3

The least absolute deviations (LAD) estimator is a special case of quantile regressors—and a good way to introduce the general concept. LAD minimizes the sum of absolute residuals (rather than the squared residuals)

$$\hat{b}_{LAD} = \arg \min_b \sum_{t=1}^T |y_t - x_t' b| \quad (7.49)$$

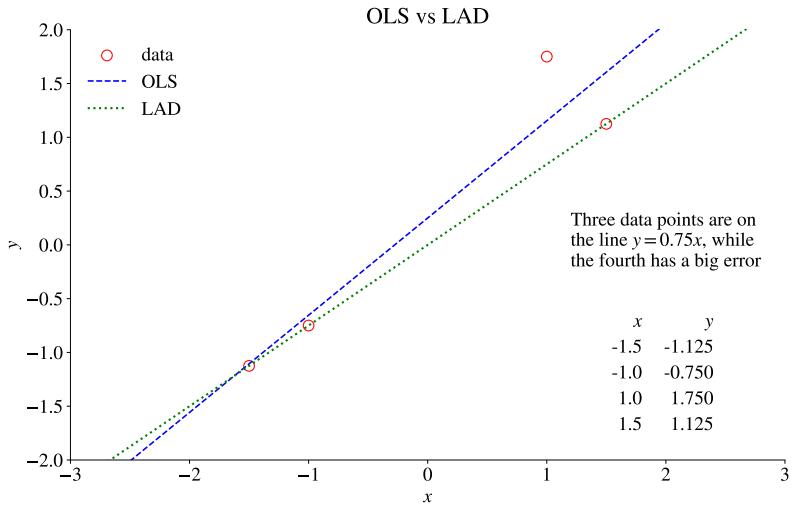


Figure 7.23: Data and regression line from OLS and LAD

The optimization is a non-linear problem, but a simple iteration works nicely (see below). The estimator is typically less sensitive to outliers than OLS. (There are also other ways to estimate robust regression coefficients.) This is illustrated in Figure 7.23.

See Figure 7.24 for an empirical example.

If we assume that the median of the true residual, u_t , is zero, then (under strict assumptions, discussed below) we have

$$\sqrt{T}(\hat{b}_{LAD} - b_0) \xrightarrow{d} N\left[0, f(0)^{-2} \Sigma_{xx}^{-1}/4\right], \text{ where} \quad (7.50)$$

$$\Sigma_{xx} = \text{plim} \sum_{t=1}^T x_t x_t' / T,$$

where $f(0)$ is the value of the pdf of the residual at zero. Unless we know this density function (or else we would probably have used MLE instead of LAD), we need to estimate it—for instance with a kernel density method. However, to arrive at the result in (7.50) we must assume that the residual is independent of the regressors. (This is discussed in some detail below, see quantile regressions).

Example 7.27 ($N(0, \sigma^2)$) When $u_t \sim N(0, \sigma^2)$, then $f(0) = 1/\sqrt{2\pi\sigma^2}$, so the covariance matrix in (7.50) becomes $\pi\sigma^2 \Sigma_{xx}^{-1}/2$. This is $\pi/2$ times larger than when using LS.

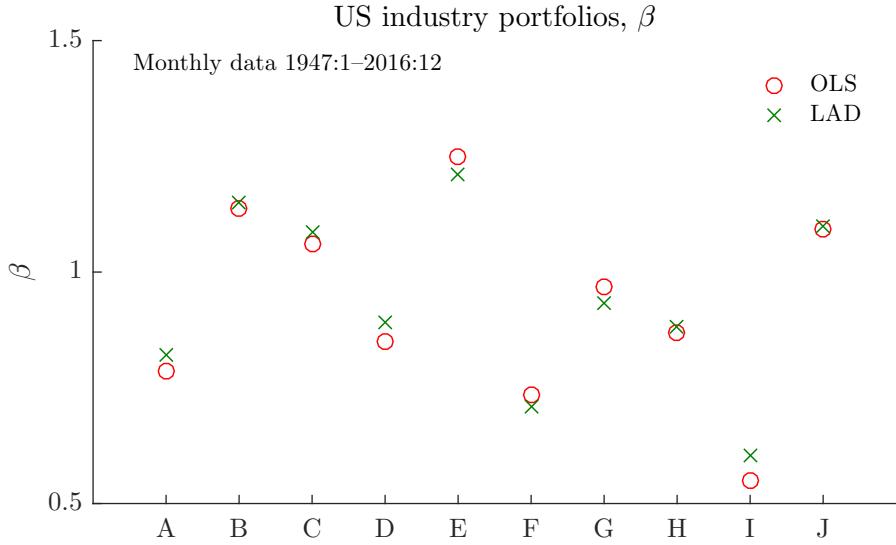


Figure 7.24: Betas of US industry portfolios

Remark 7.28 (*Algorithm for LAD*) The LAD estimator can be written

$$\hat{b}_{LAD} = \arg \min_b \sum_{t=1}^T w_t \hat{u}_t(b)^2, \quad w_t = 1/|\hat{u}_t(b)|, \text{ with}$$

$$\hat{u}_t(\hat{b}) = y_t - x'_t \hat{b}$$

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

7.10.2 Reinterpreting the LAD

Consider a linear regression

$$y_t = x'_t b + u_t. \quad (7.51)$$

In the OLS context we typically assume $E u_t = 0$ and $\text{Cov}(x_t, u_t) = 0$. The latter is the same as $E(u_t | x_t) = 0$ which means that

$$E(y_t | x_t) = x'_t b. \quad (7.52)$$

We can interpret the LAD estimator as an alternative way of getting good estimates of b , especially when the error distribution has fat tails. In fact, when the errors have a Laplace distribution, $f(u) = \exp(-|u|/\sigma)/2\sigma$, then LAD is the MLE.

Remark 7.29 ($E(u|x) = 0$ or $(E u_t = 0, \text{Cov}(x_t, u_t) = 0)^*$) For any random variables u and x ,

$$\text{Cov}(x, u) = \text{Cov}[x, E(u|x)].$$

The condition $E(u|x) = 0$ therefore implies $\text{Cov}(x, u) = 0$. It also implies $E u = 0$ since $E u = E_x[E(u|x)] = E_x[0] = 0$.

Remark 7.30 (Mean and median as solutions to minimization problems) If u is a random variable, then the mean, μ , is the solution to $\min_{\mu} E(u - \mu)^2$, while the median, m , is the solution to $\min_m E|u - m|$. (There are some restrictions on u for this to be true, but we disregard that here.)

The previous remark shows that the LAD estimator (7.49) amounts to finding the b coefficients (in a linear model) so that

$$\text{Median}(u_t|x_t) = 0, \text{ which implies} \quad (7.53)$$

$$\text{Median}(y_t|x_t) = x_t'b. \quad (7.54)$$

This is the alternative interpretation of the LAD: it tries to set the median of the residuals, at a given x_t vector, equal to zero. In contrast, OLS tries to set the mean of the residuals, at a given x_t vector, to zero.

7.10.3 Quantile Regressions

A quantile regression is a generalization of the LAD. Instead of focusing on the 0.5th quantile (the median), as is done in (7.53), it rather states that the q th quantile (conditional on x_t) of the residual is zero

$$Q(u_t|x_t; q) = 0, \text{ which implies} \quad (7.55)$$

$$Q(y_t|x_t; q) = x_t'b^{(q)}. \quad (7.56)$$

Here $Q(u_t|x_t; q)$ denotes the q th quantile of u_t at a particular value of x_t and we also index the coefficients $b^{(q)}$ to remember that this refers to the q th quantile. Clearly, the LAD is the special case when $q = 0.5$.

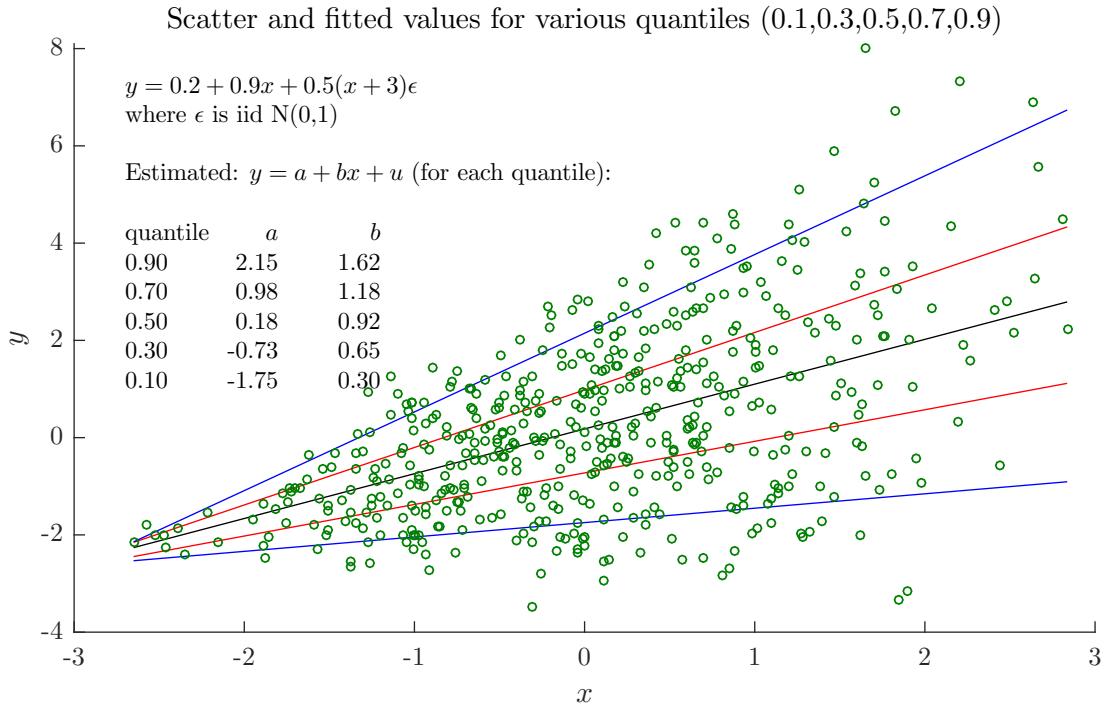


Figure 7.25: Example of quantile regression

We could estimate (see below for how) such coefficients for various quantile. When x_t just contains a constant and one more regressor, then it is easy to illustrate. See Figure 7.25 for an example where the slopes differ across the quantiles and Figure 7.26 where they do not. In particular, in Figure 7.25 the data follows a *location and scale model*

$$y_t = x_t' \beta + u_t \text{ where } u_t = x_t' \gamma \varepsilon_t, \text{ and } \varepsilon_t \text{ is iid.} \quad (7.57)$$

This is basically a linear model ($y_t = x_t' \beta$), but where the residuals (u_t) are heteroskedastic. In particular, the volatility of u_t is increasing in $|x_t' \gamma|$. This highlights the key feature of quantile regressions: they are well suited for showing how both the typical (median) and tails (for instance, the 0.1th and 0.9th quantiles) are related to the regressors. Notice, however, that we are always referring to *conditional quantiles*, that is, to quantiles of y_t at a particular value of x_t . We are *not* referring to unconditional quantiles of y_t . This means that the slopes for a high quantile (0.9, say) do not necessarily describe the relation between y_t and x_t at generally (unconditionally) high y_t (or x_t) values—see Figure 7.25. Rather, the slopes describes the relation between y_t and x_t at high ε_t values. This

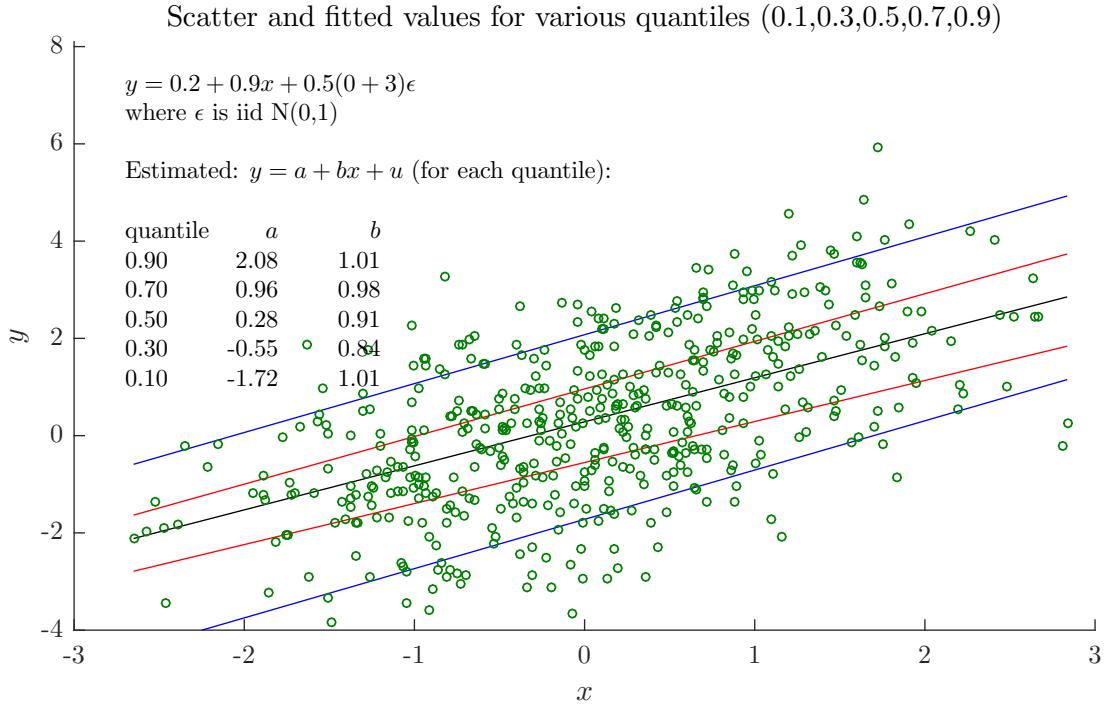


Figure 7.26: Example of quantile regression

is perhaps best seen by using the location and scale model in (7.57) which implies

$$Q(y_t|x_t; q) = x_t' \beta + Q(u_t|x_t; q) \quad (7.58)$$

$$= x_t' \beta + x_t' \gamma Q(\varepsilon_t; q) \quad (7.59)$$

$$= x_t' [\beta + \gamma Q(\varepsilon_t; q)]. \quad (7.60)$$

(In the second line, $Q(\varepsilon_t; q)$ need not be conditioned on x_t since ε_t is independent of x_t .)

Comparing with (7.56) shows that

$$b^{(q)} = \beta + \gamma Q(\varepsilon_t; q). \quad (7.61)$$

For instance, if $\gamma > 0$, then $b^{(q)}$ is increasing with q since $Q(\varepsilon_t; q)$ is. For instance, if $\varepsilon_t \sim N(0, 1)$, then $Q(\varepsilon_t; 0.05) = -1.64$ and $Q(\varepsilon_t; 0.95) = 1.64$. This is the case illustrated in Figure 7.25—where the higher slopes at high quantiles basically capture heteroskedasticity. In contrast, Figure 7.26 shows the case where the γ coefficient on the non-constant regressors are all zero: the $b^{(q)}$ coefficients (except for the constants) are the same across quantiles.

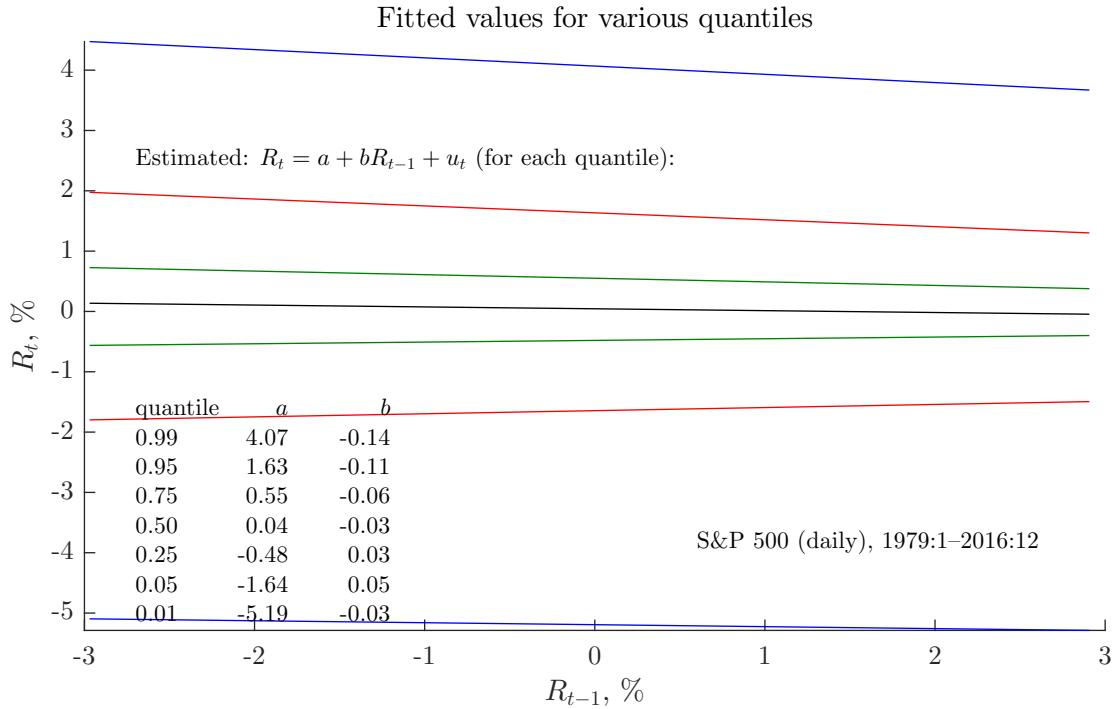


Figure 7.27: Quantile regression of AR(1) of daily returns

Figure 7.27 illustrates these points by showing the predicted quantiles of a return as a function of the lagged return. The empirical evidence suggests that the typical (median) effect of a lagged return on today's return is almost zero (there is a weak pattern of negative return to be followed by positive returns and vice versa). More pronounced is the smaller dispersion of returns after positive returns—and this is where the real payoff of the quantile regressions is.

The estimated coefficients for the q th quantile, $b^{(q)}$, solves the following problem

$$\min_{b^{(q)}} \sum_{t:u_t \geq 0} q|u_t| + \sum_{t:u_t < 0} (1-q)|u_t|, \text{ where} \quad (7.62)$$

$$u_t = y_t - x_t' b^{(q)}.$$

This is a highly non-linear problem (and the objective function does not have continuous derivatives), which can be solved by either a linear programming method or a derivative-free minimization algorithm. As a special case, $q = 0.5$ gives the LAD where (7.62) becomes

$$\min_{b^{(0.5)}} 0.5 \sum_{t=1}^T |u_t|, \text{ where } u_t = y_t - x_t' b^{(q)}, \quad (7.63)$$

which is clearly the same as (7.49).

Remark 7.31 (*Alternative way of writing (7.62)*) Suppose $u_1 \geq 0$ and $u_2 < 0$, then the sum in (7.62) can be written $qu_1 + (q-1)u_2$. This suggests that if we define a dummy d_t to be 1 if $u_t < 0$ and zero otherwise, then we can write the sum as $(q-d_1)u_1 + (q-d_2)u_2$. In general, the minimisation problem can then be written

$$\min_{b^{(q)}} \sum_{t=1}^T (q - d_t) u_t.$$

The term $(q - d_t)u_t$ is sometimes written as a function $\rho_q(u_t)$, so the sum becomes $\sum_{t=1}^T \rho_q(u_t)$. As a function of u_t , the $\rho_q(u_t)$ function has a (skewed) v-shape around $u_t = 0$. For $u_t < 0$ the function is $(q-1)u_t$ so it is linear with a (negative) slope of $q-1$, and for u_t it is also linear but with a slope of q .

The asymptotic distribution of the estimates is typically

$$\begin{aligned} \sqrt{T}(\hat{b}^{(q)} - b_0^{(q)}) &\xrightarrow{d} N[0, q(1-q)C^{-1}\Sigma_{xx}C^{-1}], \text{ where} \\ \Sigma_{xx} &= \text{plim } \sum_{t=1}^T x_t x_t' / T \text{ and} \\ C &= \text{plim } \sum_{t=1}^T f(0|x_t) x_t x_t' / T, \end{aligned} \quad (7.64)$$

where $f(0|x_t)$ is the value of the pdf of the residual, conditional on the regressor value, at a zero residual. If x_t is independent of the regressor, then $f(0|x_t) = f(0)$ where the latter is the unconditional density of the residual. In this case, the covariance matrix can be written $, q(1-q)f(0)^{-2}\Sigma_{xx}^{-1}$, which gives the result in (7.50) once we set $q = 0.5$.

One way of obtaining a consistent estimate of C is via a kernel density estimate

$$\begin{aligned} C &= \sum_{t=1}^T w_t x_t x_t' / T, \text{ with} \\ w_t &= \frac{1}{h\sqrt{2\pi}} \exp\left[-(\hat{u}_t/h)^2/2\right], \end{aligned} \quad (7.65)$$

and where \hat{u}_t is the fitted residual.

7.11 “A Closed-Form GARCH Option Valuation Model” by Heston and Nandi

References: Heston and Nandi (2000) (HN); Duan (1995)

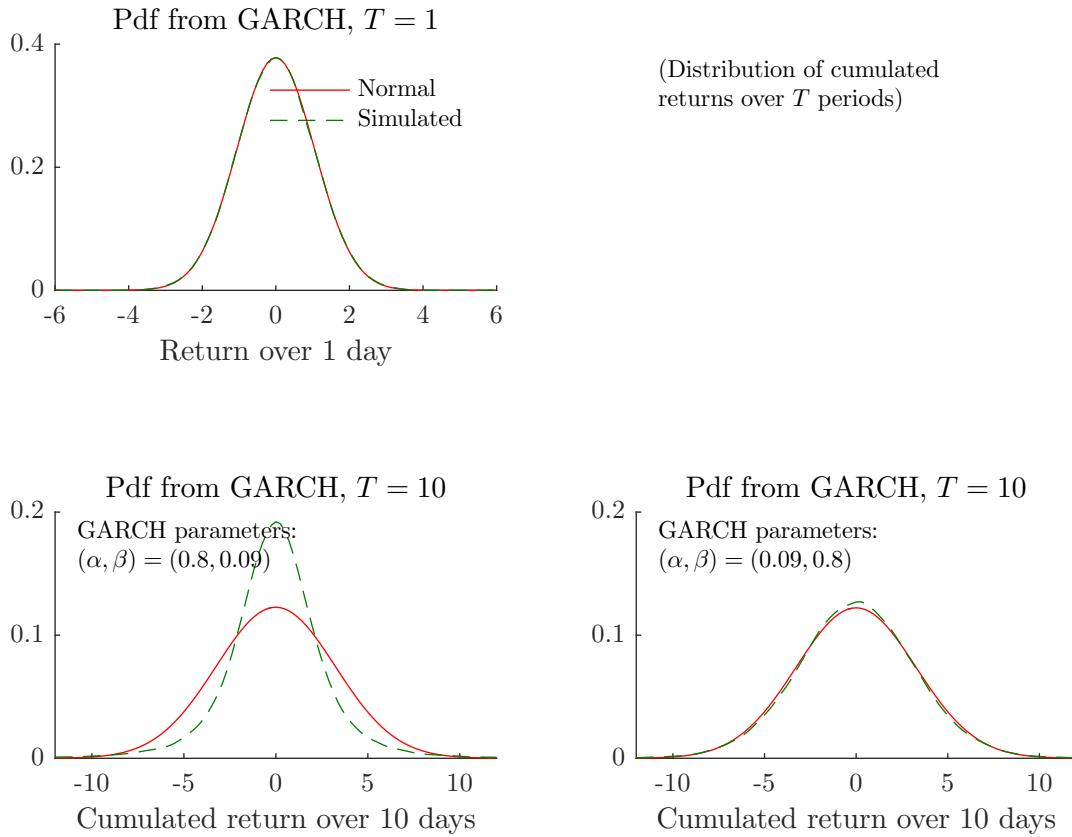


Figure 7.28: Comparison of normal and simulated distribution of m -period returns

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 compared to a Black-Scholes formula.

7.11.1 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 7.28 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 ($\beta_1 \sigma_{t-1}^2$).

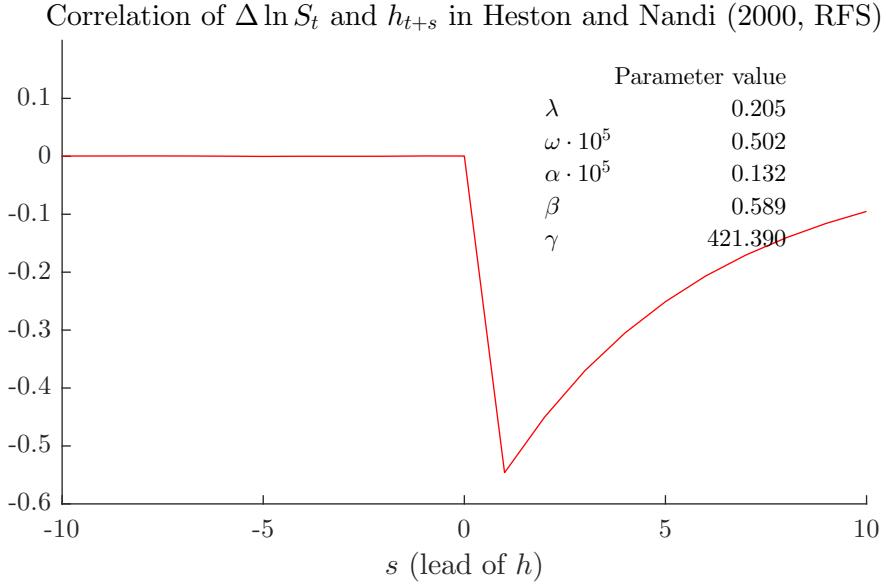


Figure 7.29: Simulated correlations of $\Delta \ln S_t$ and h_{t+s}

7.11.2 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

$$\ln S_t - \ln S_{t-\Delta} - r = \lambda h_t + \sqrt{h_t} z_t, \text{ where } z_t \text{ is iid } N(0, 1) \quad (7.66)$$

$$h_t = \omega + \alpha_1(z_{t-\Delta} - \gamma_1 \sqrt{h_{t-\Delta}})^2 + \beta_1 h_{t-\Delta}. \quad (7.67)$$

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 $\gamma_1 \sqrt{h_{t-\Delta}}$ 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 7.29.

The process (7.66)–(7.67) 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 options on this asset, since we cannot use a dynamic hedging approach to establish a no-

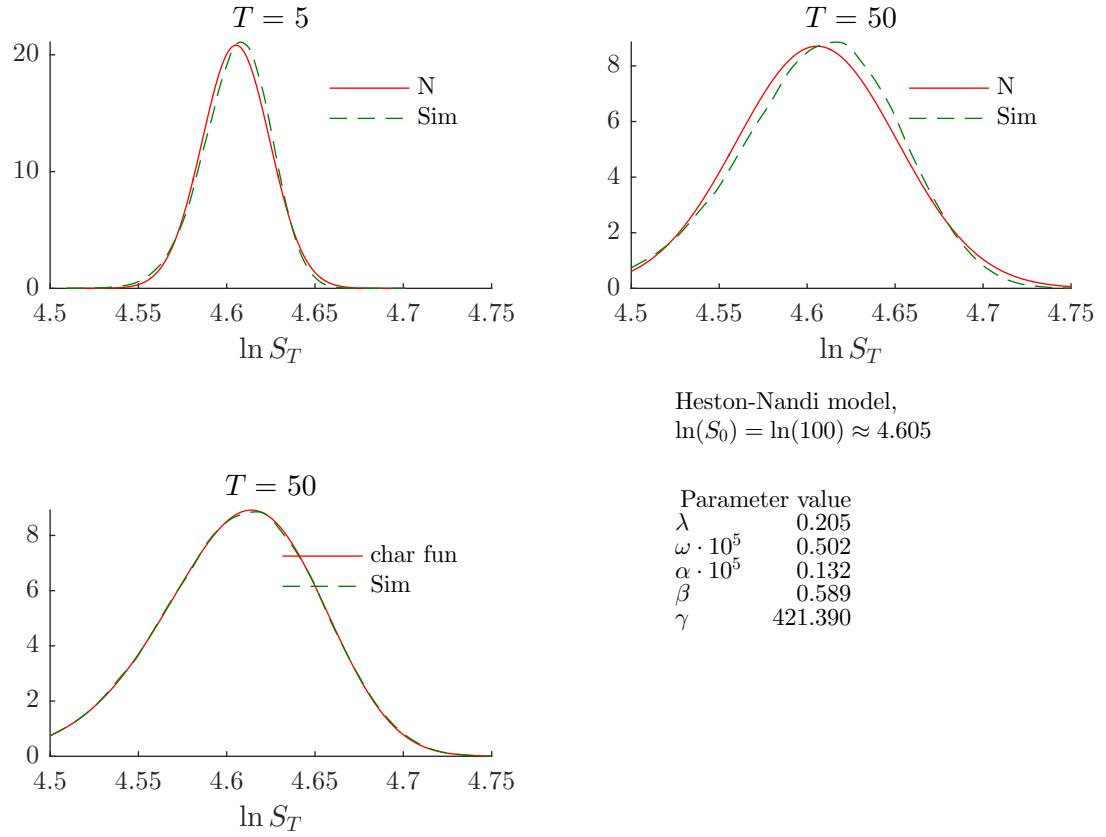


Figure 7.30: Distribution (physical) of $\ln S_T$ in the Heston-Nandi model

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

$$C_{t-\Delta} = E_{t-\Delta} \{ M_t \max [S_t - K, 0] \}. \quad (7.68)$$

Alternatively, we can write

$$C_{t-\Delta} = e^{-r\Delta} E_{t-\Delta}^* \{ \max [S_t - K, 0] \}, \quad (7.69)$$

where $E_{t-\Delta}^*$ 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 7.5. These estimates suggests that λ has the wrong sign (high volatility predicts low future returns) and the persistence of volatility is much higher than in HN (β is much higher).

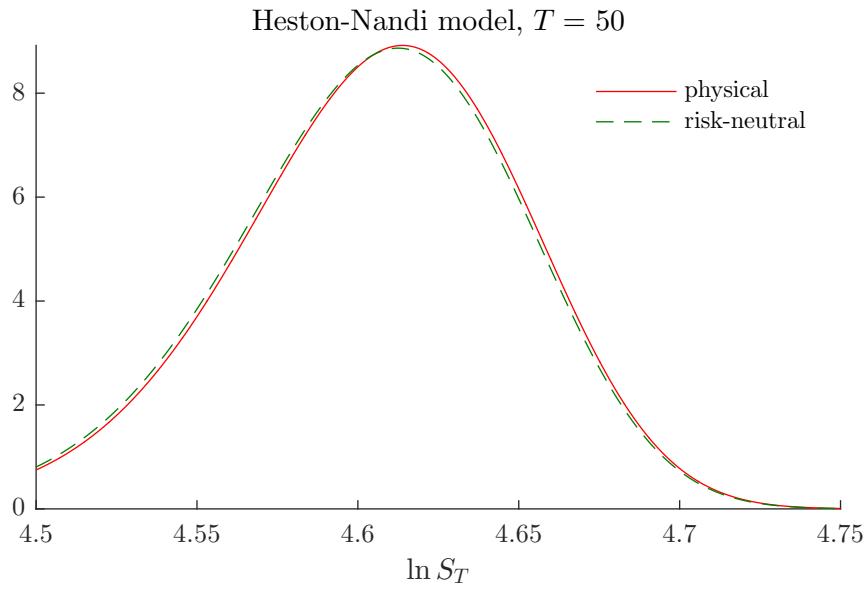


Figure 7.31: Physical and riskneutral distribution of $\ln S_T$ in the Heston-Nandi model

λ	-2.5
ω	1.22e-006
α	0.00259
β	0.903
γ	6.06

Table 7.5: Estimate of the Heston-Nandi model on daily S&P500 excess returns, in %.
Sample: 1990:1-2011:5

7.11.3 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 motivate 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 (7.66)–(7.67), 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

$$C_t(S_t, r, K, T) = e^{-r\Delta} E_t^* \max [S_T - K, 0] \quad (7.70)$$

$$= S_t P_1 - e^{-r\Delta} K P_2, \quad (7.71)$$

where P_1 and P_2 are two risk neutral probabilities (implied by the risk neutral version of (7.66)–(7.67), 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 , $f(\phi) = E_t^* \exp(i\phi \ln S_T)$, where $i^2 = -1$, and then inverting to get the probabilities.

Remark 7.32 (*Characteristic function and the pdf*) *The characteristic function of a random variable x is*

$$\begin{aligned} f(\phi) &= E \exp(i\phi x) \\ &= \int_x \exp(i\phi x) \text{pdf}(x) dx, \end{aligned}$$

where $\text{pdf}(x)$ 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

$$\text{pdf}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-i\phi x) f(\phi) d\phi.$$

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 7.33 (*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 $\phi.$

Remark 7.34 (*Characteristic function in the iid case*) In the special case when α_1, γ_1 and β_1 are all zero, then process (7.66)–(7.67) 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

$$f(\phi) = \exp(i\phi \ln S_0 + A_0 + B_0 \omega)$$

$$= \exp[i\phi [\ln S_0 + T(r + \omega\lambda)] - \phi^2 T\omega/2],$$

which is the characteristic function of a normally distributed variable with mean $\ln S_0 + T(r + \omega\lambda)$ and variance $T\omega.$

7.11.4 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 (7.66)–(7.67) 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 (7.71) depend (nonlinearly) on the parameters of the risk neutral version of (7.66)–(7.67). 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.

7.12 “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.

7.12.1 Basic Model

It is assumed that the individual stock markets can be described by CAPM

$$R_{it}^e = \beta_i R_{mt}^e + \varepsilon_{it}, \quad (7.72)$$

where R_{mt}^e 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 therefore have a GARCH-M (“-in-Mean”) process

$$R_{mt}^e = \lambda \sigma_{mt}^2 + \varepsilon_{mt}, \quad \text{E}_{t-1} \varepsilon_{mt} = 0 \text{ and } \text{Var}_{t-1}(\varepsilon_{mt}) = \sigma_{mt}^2, \quad (7.73)$$

$$\sigma_{mt}^2 = \zeta + \gamma \varepsilon_{m,t-1}^2 + \delta \sigma_{m,t-1}^2. \quad (7.74)$$

(Warning: BL uses a different timing/subscript convention for the GARCH model.)

7.12.2 The Price-Dividend Ratio

A gross return

$$R_{i,t+1} = \frac{D_{i,t+1} + P_{i,t+1}}{P_{it}}, \quad (7.75)$$

can be approximated in terms of logs (lower case letters)

$$r_{i,t+1} \approx \rho_i \underbrace{(p_{i,t+1} - d_{i,t+1})}_{z_{i,t+1}} - \underbrace{(p_{it} - d_{it})}_{z_{it}} + \underbrace{(d_{i,t+1} - d_{it})}_{g_{i,t+1}}, \quad (7.76)$$

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)

$$p_{it} - d_{it} = z_{it} \approx \sum_{s=0}^{\infty} \rho_i^s \text{E}_t (g_{i,t+s+1} - r_{i,t+s+1}). \quad (7.77)$$

To calculate the right hand side of (7.77), 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 (7.72)–(7.74).

Since all these three processes are modelled as univariate first-order time-series pro-

cesses, the solution is

$$p_{it} - d_{it} = z_{it} = A_{i,0} + A_{i,1}g_{it} + A_{i,2}\sigma_{m,t+1}^2 + A_{i,3}r_{ft}. \quad (7.78)$$

(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 $p_{it} - d_{it} = z_{it}$ defined in (7.78). This is the “overidentified” part of the model.

7.12.3 A Benchmark Case with No Predictability

As a benchmark for comparison, consider the case when the right hand side in (7.77) 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.

7.12.4 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 7.32.

Remark 7.35 (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

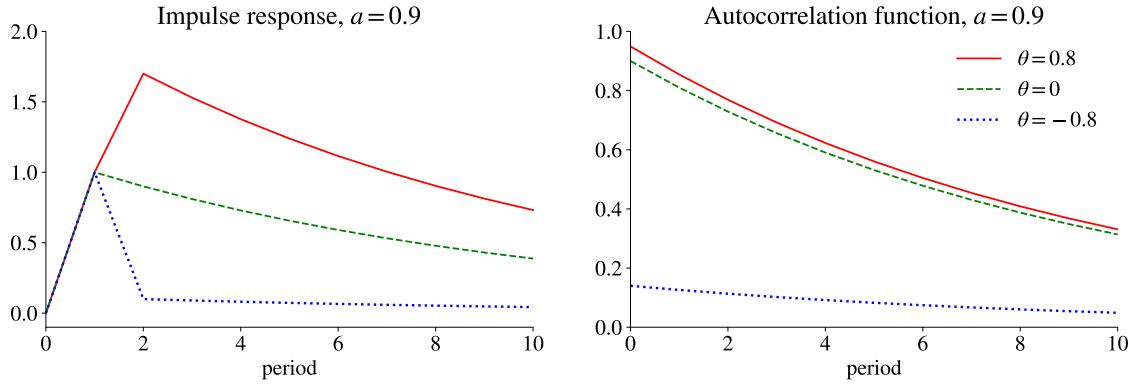
$$\rho_1 = \frac{(1 + a\theta)(a + \theta)}{1 + \theta^2 + 2a\theta}, \text{ and } \rho_s = a\rho_{s-1} \text{ for } s = 2, 3, \dots$$

and the conditional expectations are

$$\mathbb{E}_t y_{t+s} = a^{s-1}(ay_t + \theta\varepsilon_t), \quad s = 1, 2, \dots$$

7.12.5 Results

1. The hypothesis that the CAPM regressions have zero intercepts (for all five country indices) cannot be rejected.
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



ARMA(1,1): $y_t = ay_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$

Figure 7.32: Impulse response and autocorrelation functions of ARMA(1,1)

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).

7.13 Appendix: Using an FFT to Calculate the PDF from the Characteristic Function

7.13.1 Characteristic Function

The characteristic function $h(x)$ of a random variable x is

$$\begin{aligned} h(\phi) &= E \exp(i\phi x) \\ &= \int_{-\infty}^{\infty} \exp(i\phi x) f(x) dx, \end{aligned} \quad (7.79)$$

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

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-i\phi x) h(\phi) d\phi. \quad (7.80)$$

In practice, we typically use a fast (discrete) Fourier transform to perform this calculation, since there are very quick computer algorithms for doing that.

7.13.2 Inverting the Characteristic Function

Approximate the characteristic function (7.79) as the integral over $[x_{\min}, x_{\max}]$ (assuming the pdf is zero outside)

$$h(\phi) = \int_{x_{\min}}^{x_{\max}} e^{i\phi x} f(x) dx. \quad (7.81)$$

Now, approximate the integral by a Riemann sum

$$h(\phi) \approx \sum_{k=1}^N e^{i\phi x_k} f(x_k) \Delta x. \quad (7.82)$$

Split up $[x_{\min}, x_{\max}]$ into N intervals of equal size, so the step (and interval width) is

$$\Delta x = \frac{x_{\max} - x_{\min}}{N}. \quad (7.83)$$

The mid point of the k th interval is

$$x_k = x_{\min} + (k - 1/2)\Delta x, \quad (7.84)$$

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 7.36 With $(x_{\min}, x_{\max}) = (1, 7)$ and $N = 3$, then $\Delta x = (7 - 1)/3 = 2$. The x_j values are

$$\begin{bmatrix} k & x_k = x_{\min} + (k - 1/2)\Delta x \\ 1 & 1 + 1/2 \times 2 = 2 \\ 2 & 1 + 3/2 \times 2 = 4 \\ 3 & 1 + 5/2 \times 2 = 6. \end{bmatrix}$$

This gives the Riemann sum

$$h_j \approx \sum_{k=1}^N e^{i\phi[x_{\min} + (k - 1/2)\Delta x]} f_k \Delta x, \quad (7.85)$$

where $h_j = h(\phi_j)$ and $f_k = f(x_k)$.

We want

$$\phi_j = b + \frac{2\pi}{N} \frac{j - 1}{\Delta x}, \quad (7.86)$$

so we can control the central location of ϕ . Use that in the Riemann sum

$$h_j \approx \sum_{k=1}^N e^{i[x_{\min} + (k - 1/2)\Delta x] \frac{2\pi}{N} \frac{j - 1}{\Delta x}} e^{i[x_{\min} + (k - 1/2)\Delta x] b} f_k \Delta x, \quad (7.87)$$

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

$$\underbrace{e^{-i(x_{\min} + 1/2\Delta x)\frac{2\pi}{N}\frac{j-1}{\Delta x}} \frac{1}{N} h_j}_{q_j} \approx \frac{1}{N} \sum_{k=1}^N e^{\frac{2\pi i}{N}(j-1)(k-1)} \underbrace{e^{i[x_{\min} + (k-1/2)\Delta x]b} f_k \Delta x}_{Q_k}, \quad (7.88)$$

which has the same form as the ifft. We should therefore be able to calculate Q_k by applying the fft on q_j . We can then recover the density function as

$$f_k = e^{-i[x_{\min} + (k-1/2)\Delta x]b} Q_k / \Delta x. \quad (7.89)$$

7.14 Appendix: Some Proofs

7.14.1 ARCH Models

Proof. (of (7.17)–(7.18)) Notice that $E_t \sigma_{t+2}^2 = \omega + \alpha E_t v_{t+1}^2 E_t \sigma_{t+1}^2$ since v_t is independent of σ_t . Moreover, $E_t v_{t+1}^2 = 1$ and $E_t \sigma_{t+1}^2 = \sigma_{t+1}^2$ (known in t). Combine to get $E_t \sigma_{t+2}^2 = \omega + \alpha \sigma_{t+1}^2$. Similarly, $E_t \sigma_{t+3}^2 = \omega + \alpha E_t \sigma_{t+2}^2$. Substitute for $E_t \sigma_{t+2}^2$ to get $E_t \sigma_{t+3}^2 = \omega + \alpha(\omega + \alpha \sigma_{t+1}^2)$, which can be written as (7.17). Further periods follow the same pattern.

To prove (7.18), notice that $\text{Var}_t(u_{t+s}) = E_t v_{t+s}^2 \sigma_{t+s}^2 = E_t v_{t+s}^2 E_t \sigma_{t+s}^2$ since v_{t+s} and σ_{t+s} are independent. In addition, $E_t v_{t+s}^2 = 1$, which proves (7.18). ■

Proof. (of (7.20)) Since v_t and σ_t are independent, we have $E(u_t^2) = E(v_t^2 \sigma_t^2) = E \sigma_t^2$ and $E(u_t^4) = E(v_t^4 \sigma_t^4) = E(\sigma_t^4) E(v_t^4) = E(\sigma_t^4) 3$, where the last equality follows from $E(v_t^4) = 3$ for a standard normal variable. To find $E(\sigma_t^4)$, square (7.16) and take expectations (and use $E \sigma_t^2 = \omega/(1-\alpha)$)

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

Multiplying by 3 and dividing by $(E u_t^2)^2 = \omega^2/(1-\alpha)^2$ gives (7.20). ■

7.14.2 GARCH Models

Proof. (of (7.25)–(7.27)) Notice that $E_t \sigma_{t+2}^2 = \omega + \alpha E_t v_{t+1}^2 E_t \sigma_{t+1}^2 + \beta \sigma_{t+1}^2$ since v_t is independent of σ_t . Moreover, $E_t v_{t+1}^2 = 1$ and $E_t \sigma_{t+1}^2 = \sigma_{t+1}^2$ (known in t). Combine

to get $E_t \sigma_{t+2}^2 = \omega + (\alpha + \beta)\sigma_{t+1}^2$. Similarly, $E_t \sigma_{t+3}^2 = \omega + (\alpha + \beta)E_t \sigma_{t+2}^2$. Substitute for $E_t \sigma_{t+2}^2$ to get $E_t \sigma_{t+3}^2 = \omega + (\alpha + \beta)[\omega + (\alpha + \beta)\sigma_{t+1}^2]$, which can be written as (7.25). Further periods follow the same pattern.

To prove (7.27), notice that the first line can be written $= K\bar{\sigma}^2 + \sum_{s=1}^K (\alpha + \beta)^{s-1} (\sigma_{t+1}^2 - \bar{\sigma}^2)$. Then, use (7.25) and notice that $\sum_{s=1}^K (\alpha + \beta)^{s-1} = [1 - (\alpha + \beta)^K] / [1 - (\alpha + \beta)]$. ■

Proof. (of (7.28)) Since v_t and σ_t are independent, we have $E(u_t^2) = E(v_t^2\sigma_t^2) = E\sigma_t^2$ and $E(u_t^4) = E(v_t^4\sigma_t^4) = E(\sigma_t^4)E(v_t^4) = E(\sigma_t^4)3$, where the last equality follows from $E(v_t^4) = 3$ for a standard normal variable. We also have $E(u_t^2\sigma_t^2) = E\sigma_t^4$

$$\begin{aligned} E\sigma_t^4 &= E(\omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2)^2 \\ &= \omega^2 + \alpha^2 E u_{t-1}^4 + \beta^2 E \sigma_{t-1}^4 + 2\omega\alpha E u_{t-1}^2 + 2\omega\beta E \sigma_{t-1}^2 + 2\alpha\beta E(u_{t-1}^2\sigma_{t-1}^2) \\ &= \omega^2 + \alpha^2 E(\sigma_t^4)3 + \beta^2 E \sigma_t^4 + 2\omega\alpha E \sigma_t^2 + 2\omega\beta E \sigma_t^2 + 2\alpha\beta E \sigma_t^4 \\ &= \frac{\omega^2 + 2\omega(\alpha + \beta)E\sigma_t^2}{1 - 2\alpha^2 - (\alpha + \beta)^2}. \end{aligned}$$

Use $E\sigma_t^2 = \omega/(1-\alpha-\beta)$, multiply by 3 and divide by $(E u_t^2)^2 = \omega^2/(1-\alpha-\beta)^2$ gives (7.28). ■

Proof. (of (7.29)) Substitute for σ_{t-1}^2 in (7.24), 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 (7.29). ■

Chapter 8

Factor Models

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

8.1 CAPM Tests: Overview

Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 5

Let R_{it}^e be the excess return on asset i in excess over the riskfree asset, and let f_t be the excess return on the market portfolio (f for factor). CAPM with a riskfree return says that $\alpha_i = 0$ in

$$R_{it}^e = \alpha_i + \beta_i f_t + \varepsilon_{it}, \text{ where} \quad (8.1)$$
$$\mathbb{E} \varepsilon_{it} = 0 \text{ and } \text{Cov}(f_t, \varepsilon_{it}) = 0.$$

The basic test of CAPM is to estimate (8.1) 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 the factor (f_t) is the correct benchmark, then it is a test of whether asset i is “correctly” priced (this is the approach in mutual fund evaluations). Alternatively, if we assume that asset i is correctly priced, then it is a test of the mean-variance efficiency of the factor (compare the Roll critique).

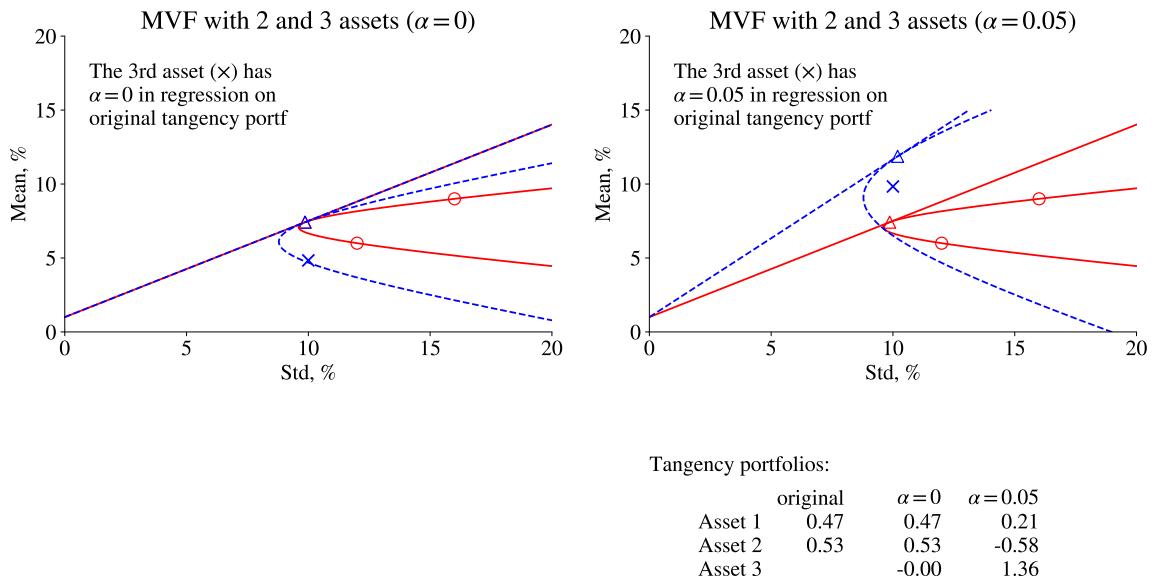


Figure 8.1: MV frontiers with 2 and 3 assets

8.2 Testing CAPM: Traditional LS Approach

8.2.1 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 (8.1) and form a t-test of the null hypothesis that the intercept is zero.

The variance of the estimated intercept in the CAPM regression (8.1) is

$$\text{Var}(\hat{\alpha}_i) = (1 + SR^2)\sigma_i^2/T, \quad (8.2)$$

where σ_i^2 is the variance of the residual in (8.1) and SR^2 is the squared Sharpe ratio of the market portfolio (recall: f_t is the excess return on market portfolio). The result is well known, but a simple proof is found in Appendix 8.10. Equation (8.2) shows 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. Clearly, (8.2) can be used to construct a t-test.

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

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} \xrightarrow{d} \chi_1^2 \text{ under } H_0: \alpha_0 = 0. \quad (8.3)$$

It is quite straightforward to use the properties of minimum-variance frontiers (see Gibbons, Ross, and Shanken (1989), MacKinlay (1995) and the simple proof in Appendix 8.10) to show that the test statistic in (8.3) can be written

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} = \frac{(\widehat{SR}_c)^2 - SR^2}{(1 + SR^2)/T}, \quad (8.4)$$

where SR 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 asset i and the market portfolio (so the market portfolio is mean-variance efficient also when we take the test asset into account) then the test statistic, $\hat{\alpha}_i^2 / \text{Var}(\hat{\alpha}_i)$, is zero—and CAPM is not rejected. The economic importance of a non-zero intercept (α) is thus that the tangency portfolio changes if the test asset is added to the investment opportunity set. See Figure 8.1 for an illustration.

8.2.2 CAPM with Several Assets: Traditional LS Approach

Suppose we have n test assets. Stack the expressions (8.1) for $i = 1, \dots, n$ as

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} f_t + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}, \text{ where} \quad (8.5)$$

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

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

$$\sqrt{T}\hat{\alpha} \xrightarrow{d} N\left[\mathbf{0}_{n \times 1}, \Sigma(1 + SR^2)\right], \text{ where} \quad (8.6)$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \dots & \hat{\sigma}_{nn} \end{bmatrix} \text{ with } \sigma_{ij} = \text{Cov}(\varepsilon_{it}, \varepsilon_{jt})$$

and where $SR^2 = (\text{E } f)^2 / \text{Var}(f)$.

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 8.11. To test the null hypothesis that all intercepts are zero, we then use the test statistic

$$T\hat{\alpha}'(1 + SR^2)^{-1}\Sigma^{-1}\hat{\alpha} \sim \chi_n^2. \quad (8.7)$$

8.2.3 CAPM with Several Assets: Bonferroni Test

Remark 8.1 (*The Bonferroni inequality*) Suppose we perform $i = 1 \dots n$ different tests, each at the significance level p_i . The Bonferroni inequality then says that if the null hypotheses are all true, then

$$\Pr(\text{not rejecting in any of the } n \text{ tests}) \geq 1 - \sum_{i=1}^n p_i.$$

It follows that rejecting in at least one of the n tests has a probability of less than or equal to $\sum_{i=1}^n p_i$. For instance, with $p_i = 0.05/n$, there is 5% chance of rejecting in at least one test: $\Pr(\text{rejecting in at least one of the } n \text{ tests}) \leq 0.05$.

As an alternative to the joint test, we could instead study each of the n assets separately. Clearly, if we can safely reject the null hypothesis for at least one asset, then the joint hypothesis is also rejected. However, this cannot be implemented with traditional critical values since the chance of at least one false rejection increases with the number of test assets.

To control this “family-wise error rate,” a Bonferroni correction is applied. To do this, let t_i be the t -stat for asset i ($t_i = \hat{\alpha}_i / \text{Std}(\hat{\alpha}_i)$). As usual, we would reject the hypothesis that $\alpha_i = 0$ on the 5% level if $|t_i| > 1.96$.

Redo this for each asset—and reject the joint hypothesis on the family-wise significance level of 5% if at least one of the individual test statistics exceeds the $0.05/n$ critical value. For instance, with 10 test assets, we compare $|t_i|$ with 2.81 instead of 1.96 (since

2.81 is the 99.75th percentile of a $N(0, 1)$ distribution, whereas the 97.5 percentile is 1.96). To use another significance level ρ , use ρ/n instead of $0.05/n$.

It can be noticed that since we focus on the highest individual test statistic, the Bonferroni and the Holm-Bonferroni (Holm, 1979) methods give the same result. This would be different if we wanted to see how many of the alphas that differ from 0. In that case the Holm-Bonferroni method is more powerful.

8.3 Testing CAPM: GMM

8.3.1 CAPM with Several Assets: GMM and a Wald Test

To test n assets at the same time when the errors are non-iid we can use of the GMM framework.

Write the n regressions in (8.5) on vector form as

$$R_t^e = \alpha + \beta f_t + \varepsilon_t, \text{ where} \quad (8.8)$$

$$\mathbb{E} \varepsilon_t = \mathbf{0}_{n \times 1} \text{ and } \text{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{1 \times n},$$

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 ,

$$\mathbb{E} g_t(\alpha, \beta) = \mathbf{0}_{2n \times 1}, \text{ where} \quad (8.9)$$

$$g_t(\alpha, \beta) = \begin{bmatrix} \varepsilon_t \\ f_t \varepsilon_t \end{bmatrix} = \begin{bmatrix} R_t^e - \alpha - \beta f_t \\ f_t (R_t^e - \alpha - \beta f_t) \end{bmatrix}. \quad (8.10)$$

There are as many parameters as moment conditions, so the GMM estimator picks values of α and β such that the sample analogues of (8.9) are satisfied exactly, 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 8.2.

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

$$\hat{\alpha}' \text{Var}(\hat{\alpha})^{-1} \hat{\alpha} \xrightarrow{d} \chi_n^2. \quad (8.11)$$

Remark 8.2 (*Easy coding of the GMM Problem (8.9)*) Estimate (8.8) by LS (equation by equation). Then, plug in the fitted residuals in (8.10) to generate time series of the

moments (will be important for the tests).

Remark 8.3 (Distribution of GMM) Let the parameter vector in the moment condition have the true value b_0 . Define

$$S_0 = \text{Cov} \left[\sqrt{T} \bar{g}(b_0) \right] \text{ and } D_0 = \text{plim} \frac{\partial \bar{g}(b_0)}{\partial b'}.$$

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

$$\sqrt{T}(\hat{b} - b_0) \xrightarrow{d} N(\mathbf{0}_{k \times 1}, V), \text{ where } V = (D_0 S_0^{-1} D_0')^{-1}.$$

When D_0 is invertible (as it would be in an exactly identified model), then we can also write $V = D_0^{-1} S_0 (D_0^{-1})'$.

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 be written as

$$\begin{aligned} \text{plim} \frac{\partial \bar{g}_t(\alpha, \beta)}{\partial [\alpha, \beta]} &= D_0 = -E \begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} \otimes I_n \\ &= -E \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix}' \right) \otimes I_n, \end{aligned} \quad (8.12)$$

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

$$D_0^{-1} = - \left[E \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix}' \right) \right]^{-1} \otimes I_n, \quad (8.13)$$

since $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (if conformable).

Remark 8.4 (*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 8.5 (*Two test assets*) With assets 1 and 2, the parameter vector is $b = [\alpha_1, \alpha_2, \beta_1, \beta_2]'$. Write out the sample analogues of (8.9) 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

$$\begin{aligned} \frac{\partial \bar{g}(\alpha, \beta)}{\partial [\alpha_1, \alpha_2, \beta_1, \beta_2]'} &= \begin{bmatrix} \partial \bar{g}_1 / \partial \alpha_1 & \partial \bar{g}_1 / \partial \alpha_2 & \partial \bar{g}_1 / \partial \beta_1 & \partial \bar{g}_1 / \partial \beta_2 \\ \partial \bar{g}_2 / \partial \alpha_1 & \partial \bar{g}_2 / \partial \alpha_2 & \partial \bar{g}_2 / \partial \beta_1 & \partial \bar{g}_2 / \partial \beta_2 \\ \partial \bar{g}_3 / \partial \alpha_1 & \partial \bar{g}_3 / \partial \alpha_2 & \partial \bar{g}_3 / \partial \beta_1 & \partial \bar{g}_3 / \partial \beta_2 \\ \partial \bar{g}_4 / \partial \alpha_1 & \partial \bar{g}_4 / \partial \alpha_2 & \partial \bar{g}_4 / \partial \beta_1 & \partial \bar{g}_4 / \partial \beta_2 \end{bmatrix} \\ &= -\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & 0 & f_t & 0 \\ 0 & 1 & 0 & f_t \\ f_t & 0 & f_t^2 & 0 \\ 0 & f_t & 0 & f_t^2 \end{bmatrix} = -\frac{1}{T} \sum_{t=1}^T \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix}' \right) \otimes I_2. \end{aligned}$$

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

$$S_0 = \text{Cov} \left(\frac{\sqrt{T}}{T} \sum_{t=1}^T g_t(\alpha, \beta) \right). \quad (8.14)$$

The Newey-West estimator is often a good estimator of S_0 .

From Remark 8.3, we can write the covariance matrix of the $2n \times 1$ vector of parameters (n parameters in α and another n in β) as

$$\text{Cov} \left(\sqrt{T} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \right) = D_0^{-1} S_0 (D_0^{-1})'. \quad (8.15)$$

8.3.2 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 (8.9). The moment conditions are then

$$\mathbb{E} g(\beta) = \mathbb{E} \begin{bmatrix} R_t^e - \beta f_t \\ f_t(R_t^e - \beta f_t) \end{bmatrix} = \mathbf{0}_{2n \times 1}. \quad (8.16)$$

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

$$\min_b \bar{g}(b)' W \bar{g}(b), \quad (8.17)$$

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

$$A_{n \times 2n} \mathbb{E} g(\beta) = \mathbf{0}_{n \times 1}. \quad (8.18)$$

The model is then tested by testing if all $2n$ moment conditions in (8.16) 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.

For instance, to effectively use only the last n moment conditions in the estimation, we specify

$$A \mathbb{E} g(\beta) = \begin{bmatrix} 0_{n \times n} & I_n \end{bmatrix} \mathbb{E} \begin{bmatrix} R_t^e - \beta f_t \\ f_t(R_t^e - \beta f_t) \end{bmatrix} = \mathbf{0}_{n \times 1}. \quad (8.19)$$

This clearly gives the classical LS estimator without an intercept

$$\hat{\beta} = \frac{\sum_{t=1}^T f_t R_t^e / T}{\sum_{t=1}^T f_t^2 / T}. \quad (8.20)$$

Example 8.6 (*Combining moment conditions, CAPM on two assets*) With two assets we can combine the four moment conditions into only two by

$$A \mathbb{E} g_t(\beta_1, \beta_2) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbb{E} \begin{bmatrix} R_{1t}^e - \beta_1 f_t \\ R_{2t}^e - \beta_2 f_t \\ f_t(R_{1t}^e - \beta_1 f_t) \\ f_t(R_{2t}^e - \beta_2 f_t) \end{bmatrix} = \mathbf{0}_{2 \times 1}.$$

Remark 8.7 (*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

$$T \bar{g}(\hat{\beta})' S_0^{-1} \bar{g}(\hat{\beta}) \xrightarrow{d} \chi_{q-k}^2,$$

where q is the number of moment conditions and k is the number of parameters.

Remark 8.8 (*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 from those in Remarks 8.3 and 8.7.

8.3.3 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 worsen the 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).

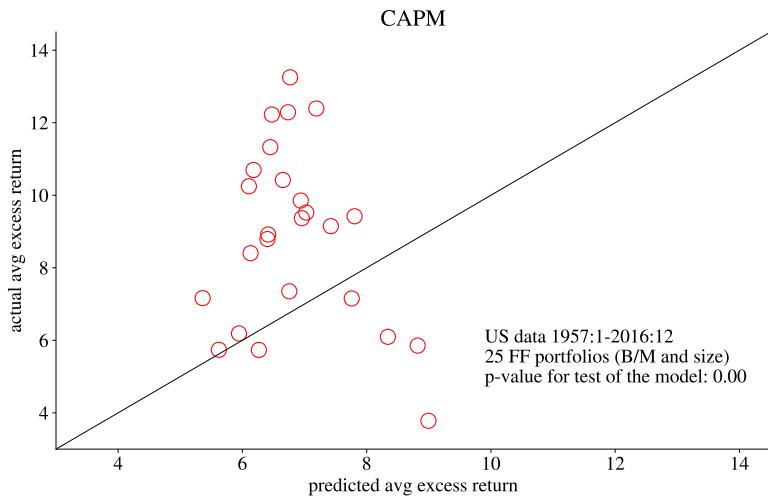


Figure 8.2: CAPM, FF portfolios

8.3.4 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. Also, 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 be particularly informative.

8.3.5 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,

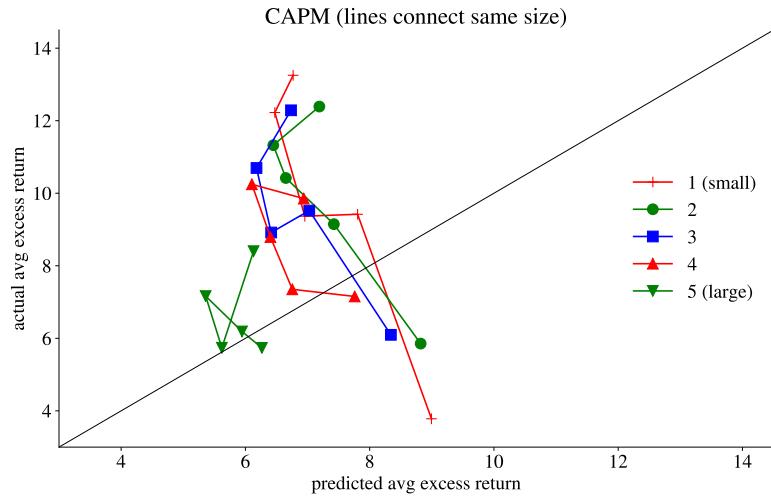


Figure 8.3: CAPM, FF portfolios

where portfolio $i j$ belongs to the i th size quantile *and* the j th BE/ME quantile (so this is a *double-sort*). This is illustrated in Table 8.1.

		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 8.1: Numbering of the FF 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 $E R_{it}^e = \beta_i E R_{mt}^e$). However, there is little relation between $E R_{it}^e$ and β_i (see Figure 8.2). This lack of relation 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 $E R_{it}^e$ and β_i —in stark contrast to CAPM (see Figure 8.3). Second, *within a BE/ME quantile*, there is a positive relation (across size quantiles) between $E R_{it}^e$ and β_i —as predicted by CAPM (see Figure 8.4).

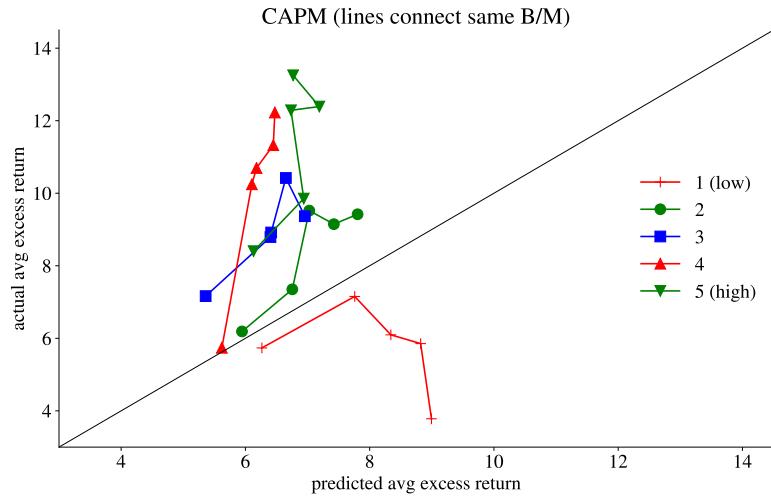


Figure 8.4: CAPM, FF portfolios

8.4 Testing Multi-Factor Models (Factors are Excess Returns)

Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 6.2.1

8.4.1 A Multi-Factor Model

When the K factors, f_t , are excess returns, the null hypothesis typically says that $\alpha_i = 0$ in

$$R_{it}^e = \alpha_i + \beta_i' f_t + \varepsilon_{it}, \text{ where} \quad (8.21)$$

$$\mathbb{E} \varepsilon_{it} = 0 \text{ and } \text{Cov}(f_t, \varepsilon_{it}) = \mathbf{0}_{K \times 1},$$

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

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_{11} & \dots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \dots & \beta_{nK} \end{bmatrix} \begin{bmatrix} f_{1t} \\ \vdots \\ f_{Kt} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}$$

or in vector form

$$R_t^e = \alpha + \beta f_t + \varepsilon_t, \text{ where} \quad (8.22)$$

$$\mathbb{E} \varepsilon_t = \mathbf{0}_{n \times 1} \text{ and } \text{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{K \times n},$$

where α is $n \times 1$ and β is $n \times K$. Notice that β_{ij} shows how the i th asset depends on the j th factor.

This is, of course, very similar to the CAPM (one-factor) model—and both the LS and GMM approaches discussed before are straightforward to extend.

8.4.2 Multi-Factor Model: Traditional LS (SURE)

The results from the LS approach of testing CAPM generalizes directly (see Appendix 8.11 for details). In particular, (8.7) still holds—but where the residuals are from the multi-factor regressions (8.21) and where the Sharpe ratio of the tangency portfolio (based on the factors) depends on the means and covariance matrix of all factors

$$T \hat{\alpha}'(1 + SR^2)^{-1} \Sigma^{-1} \hat{\alpha} \sim \chi_n^2, \text{ where} \quad (8.23)$$

$$SR^2 = \mathbb{E} f' \text{Cov}(f)^{-1} \mathbb{E} f.$$

8.4.3 Multi-Factor Model: GMM

The moment conditions are

$$\mathbb{E} g_t(\alpha, \beta) = \mathbb{E} \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes \varepsilon_t \right) = \mathbb{E} \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \right) = \mathbf{0}_{n(1+K) \times 1}. \quad (8.24)$$

Note that this expression looks similar to (8.9)—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 the sample analogues of (8.24), which is the same as LS (equation by equation). The model can be tested by testing if all alphas are zero, as in (8.11).

Instead, when we restrict $\alpha = \mathbf{0}_{n \times 1}$ (overidentified system), then we either specify a weighting matrix W and solve

$$\min_{\beta} \bar{g}(\beta)' W \bar{g}(\beta), \quad (8.25)$$

or we specify a matrix A to combine the moment conditions and solve

$$A_{nK \times n(1+K)} \bar{g}(\beta) = \mathbf{0}_{nK \times 1}. \quad (8.26)$$

Example 8.9 (*Moment condition with two assets and two factors*) The moment conditions for $n = 2$ and $K = 2$ are

$$\mathbb{E} g_t(\alpha, \beta) = \mathbb{E} \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_{11}f_{1t} - \beta_{12}f_{2t} \\ R_{2t}^e - \alpha_2 - \beta_{21}f_{1t} - \beta_{22}f_{2t} \\ f_{1t}(R_{1t}^e - \alpha_1 - \beta_{11}f_{1t} - \beta_{12}f_{2t}) \\ f_{1t}(R_{2t}^e - \alpha_2 - \beta_{21}f_{1t} - \beta_{22}f_{2t}) \\ f_{2t}(R_{1t}^e - \alpha_1 - \beta_{11}f_{1t} - \beta_{12}f_{2t}) \\ f_{2t}(R_{2t}^e - \alpha_2 - \beta_{21}f_{1t} - \beta_{22}f_{2t}) \end{bmatrix} = \mathbf{0}_{6 \times 1}.$$

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 (8.12)–(8.13). Notice that in this Jacobian we differentiate the moment conditions (8.24) with respect to $\text{vec}(\alpha, \beta)$, 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 (8.14).

8.4.4 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).

Remark 8.10 (*The Fama-French factors*) The SMB and HML are created by a double sort. First classify firms according to size: small or big, using the median as a cutoff.

Second, classify firms according the book/market value: low (growth stocks, using 30th percentile as cutoff), neutral or high (value stocks, using 70th percentile as cutoff). Create six value weighted portfolios from the intersection of those groups

	<u>Low book/market</u>	<u>Medium book/market</u>	<u>High book/market</u>
Small:	Small Growth (SG)	Small Neutral (SN)	Small Value (SV)
Big:	Big Growth (BG)	Big Neutral (BN)	Big Value (BV)

The SMB is the average of the small portfolios minus the average of the big portfolios: $SMB = 1/3(SG + SN + SV) - 1/3(BG + BN + BV)$. Rearranging gives $SMB = 1/3(SG - BG) + 1/3(SN - BN) + 1/3(SV - BV)$, which shows that it represents the return on small stocks (for a given book/market) minus the return on big stocks (for same book/market). The HML is the average of the value stocks minus the growth stocks, $HML = 1/2(SV + BV) - 1/2(SG + BG)$, which can be rearranged as $HML = 1/2(SV - SG) + 1/2(BV - BG)$, which shows that it represents the return on value stocks (for a given size) minus the return on growth stocks (for the same size).

The Fama-French 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 8.5–8.8.

Is it a trivial finding that the 25 FF portfolios are better explained once we use the HML and SMB factors? No, as argued by Fama and French (1996) it just shows that there is a common (possibly unknown) pricing factor. To see that in a simplified setting, suppose excess returns are generated by some one-factor model $R_{it}^e = \beta_i f_t + \varepsilon_{it}$, although we may not know what the factor is. In addition, assume that the portfolio (HML or SMB, say) we create is just an equally weighted average across all n assets like

$$x_t = \beta_x f_t + \sum_{i=1}^n \varepsilon_{it} / n, \text{ where } \beta_x = \sum_{i=1}^n \beta_i / n. \quad (8.27)$$

(With an appropriate interpretation of the signs of the betas, this could actually be a long-short portfolio.) Regressing R_{it}^e on this portfolio gives a slope coefficient $\gamma_i = \text{Cov}(R_{it}^e, x_t) / \text{Var}(x_t)$. If we assume that all residuals are uncorrelated with the factor and with each other, then the numerator of γ_i can be simplified as

$$\text{Cov}(R_{it}^e, x_t) = \beta_i \beta_x \text{Var}(f_t) + \text{Var}(\varepsilon_{it}) / n. \quad (8.28)$$

The last term is due to the fact that ε_{it} shows up both in R_{it}^e and x_t , but its importance decreases as the number of assets in the portfolio (n) increases. This shows that if the

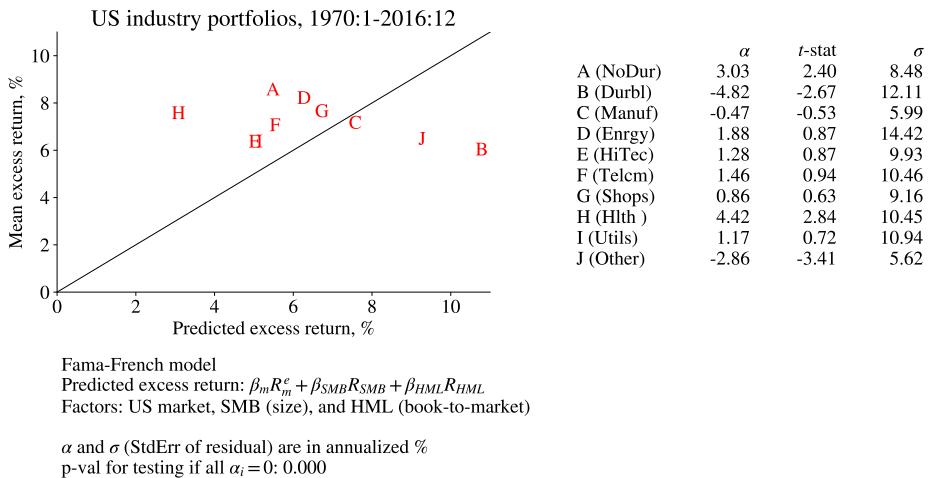


Figure 8.5: Three-factor model, US industry portfolios

cross-section (n) is large, then γ_i depends mostly on the first term. Clearly, the first term is non-zero if all three ingredients are non-zero. This means that both asset i and the portfolio are exposed to a (time-varying) factor, although we may not know what that factor represents. However, the pattern of γ_i across assets may give us a clue. (There are clearly other methods to investigate if there are common factors, for instance, principal component analysis.)

Remark 8.11 (*Factor structure after having controlled for the market movements**) *If the purpose is to investigate if there is a remaining factor structure after having controlled for the market movements, we can do the following. First, create “abnormal returns” as $R_{it}^e - \hat{b}_i R_{mt}^e$, where \hat{b}_i is the coefficient obtained from regressing R_{it}^e on R_{mt}^e (and a constant). Then, replace R_{it}^e (also in the definition of x_t) in (8.27)–(8.28) with this abnormal return. By the properties of OLS, this gives the same as running multiple regressions using R_{mt}^e and x_t as regressors (this is the Frisch-Waugh theorem).*

8.4.5 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* approach. First define M discrete investor groups (for instance, age 18–30, 31–40, etc) and calculate their

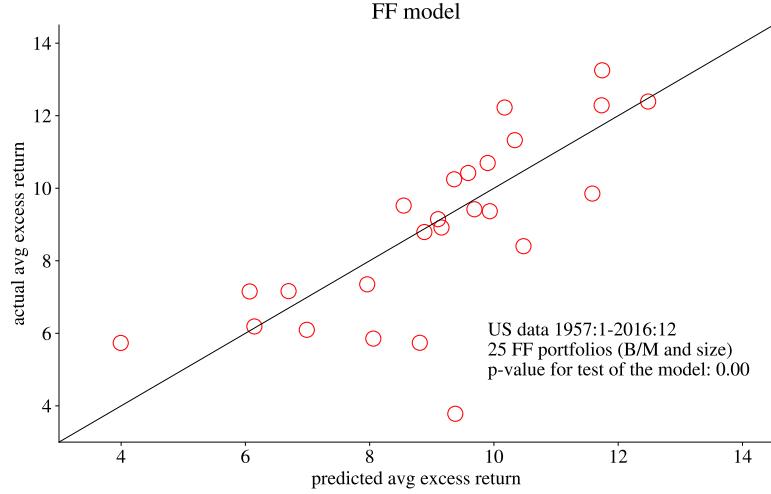


Figure 8.6: FF, FF portfolios

respective average excess returns (\bar{R}_{jt}^e for group j)

$$\bar{R}_{jt}^e = \frac{1}{N_j} \sum_{i \in \text{Group } j} R_{it}^e, \quad (8.29)$$

where N_j is the number of individuals in group j .

Then, we run a factor model

$$\bar{R}_{jt}^e = \alpha_j + \beta'_j f_t + v_{jt}, \text{ for } j = 1, 2, \dots, M \quad (8.30)$$

where f_t typically includes various return factors (for instance, excess returns on equity and bonds). By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the “alpha”) is higher for the M th group than for the first group.

Example 8.12 (*Calendar time approach with two investor groups*) *With two investor groups, estimate the following SURE system*

$$\begin{aligned} \bar{R}_{1t}^e &= \alpha_1 + \beta'_1 f_t + v_{1t}, \\ \bar{R}_{2t}^e &= \alpha_2 + \beta'_2 f_t + v_{2t}. \end{aligned}$$

The calendar time approach is straightforward and the cross-sectional correlations are fairly easy to handle (in the SURE approach). However, it forces us to define discrete

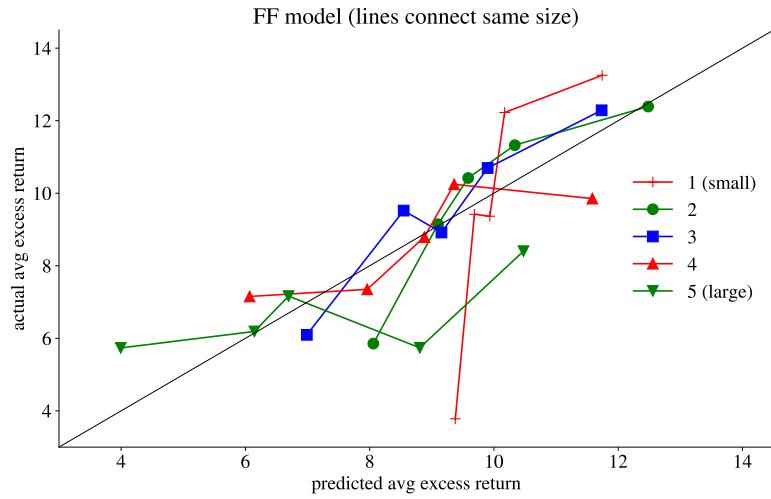


Figure 8.7: FF, FF portfolios

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* approach is to first estimate the factor model for each investor

$$R_{it}^e = \alpha_i + \beta_i' f_t + \varepsilon_{it}, \text{ for } i = 1, 2, \dots, N \quad (8.31)$$

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

$$\hat{\beta}_{pi} = z_i' c_p + w_{pi}. \quad (8.32)$$

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

8.5 Testing Multi-Factor Models (General Factors)

Reference: Cochrane (2005) 12.2; Campbell, Lo, and MacKinlay (1997) 6.2.3 and 6.3

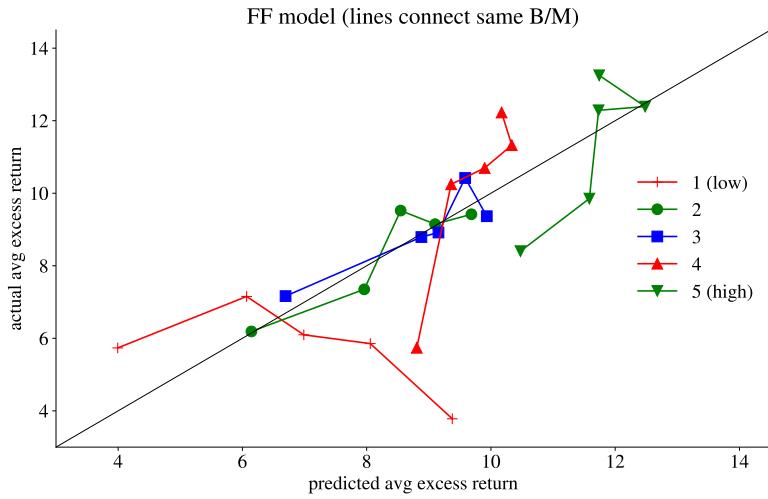


Figure 8.8: FF, FF portfolios

8.5.1 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)

$$\mathbb{E} R_{it}^e = \beta_i' \lambda, \text{ where } \lambda \text{ is } K \times 1, \text{ for } i = 1, \dots, n, \quad (8.33)$$

where the β_i are the loading of asset i on the factors, as estimated from (8.21).

Stacking the test assets gives

$$\mathbb{E} \begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \dots & \beta_{nK} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_K \end{bmatrix}, \text{ or} \\ \mathbb{E} R_t^e = \beta \lambda, \quad (8.34)$$

where β is $n \times K$.

When the factors are excess returns, then the factor risk premia must equal the expected excess returns of those factors, $\lambda = \mathbb{E} f_t$. (To see this, let the factor also be one of the test assets. It will then get a beta equal to unity on itself.) In any case, 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.

Remark 8.13 (*Factor mimicking portfolios*) *It is more difficult to estimate and test a*

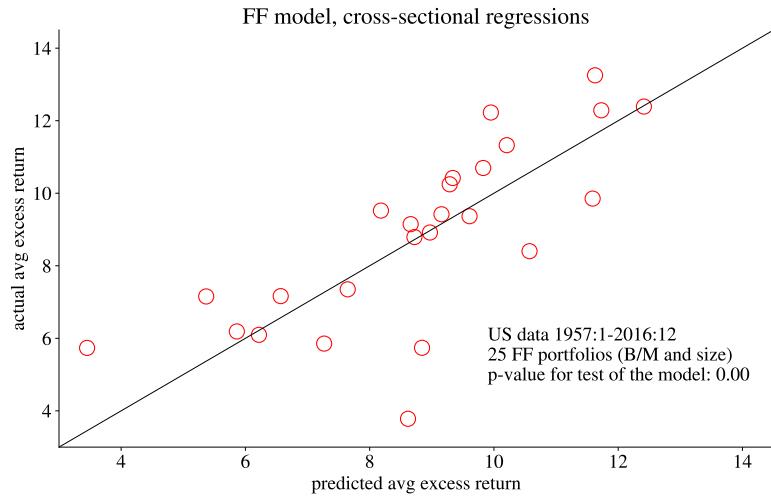


Figure 8.9: FF, FF portfolios

model with general factors than a model with excess return factors. A common approach to get around the difficulties is to replace any general factor with the linear combination of excess returns that best mimics the general factor. This linear combination can be constructed by either forming a regression of the general factor on a vector of excess returns, or by creating an arbitrage portfolio that is long assets that are highly correlated with the general factor and short assets that are less or even negatively correlated with the factor.

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 (8.21) (equation by equation); second, use $\hat{\beta}_i$ as regressors in a regression equation of the type (8.33) with a residual added

$$\bar{R}_i^e = \hat{\beta}'_i \lambda + u_i, \quad (8.35)$$

where $\bar{R}_i^e = \sum_{t=1}^T R_{it}^e / T$ is the (time-series) average of R_{it}^e .

It is then tested if $u_i = 0$ for all assets $i = 1, \dots, n$. This approach is often called a *cross-sectional* regression while the previous tests are time series regressions. Clearly, this approach relies on the assumption that the betas are indeed non-zero (and preferably not too similar across the test assets).

An issue with 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 down-

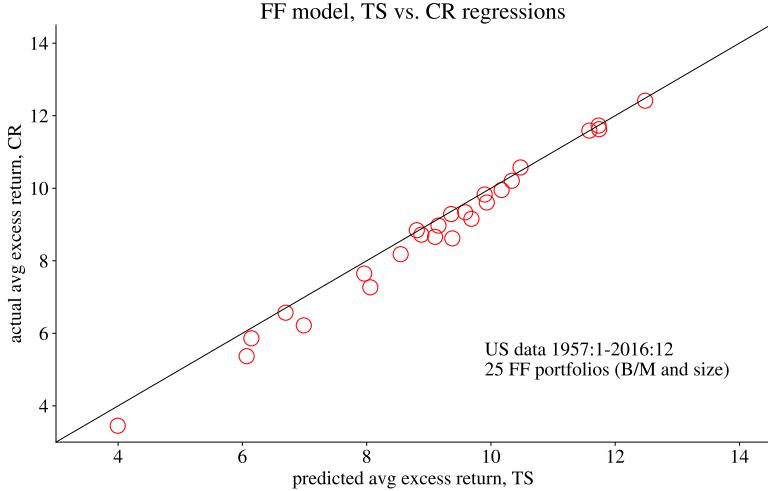


Figure 8.10: FF, FF portfolios

wards bias of the estimates of λ and an upward bias of the mean of the fitted residuals; and (ii) invalidate the standard expression of the test of λ .

A way to handle these problems is to combine the moment conditions for the time series regressions (8.24) (to estimate β) with (8.34) (to estimate λ) to get a joint system

$$E g_t(\alpha, \beta, \lambda) = E \left[\begin{bmatrix} 1 \\ f_t \\ R_t^e - \beta \lambda \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \right] = \mathbf{0}_{n(1+K+1) \times 1}. \quad (8.36)$$

We can then test the overidentifying restrictions of the model. There are $n(1 + K + 1)$ 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. From the GMM estimation using (8.36) 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. Notice that this is, in general, a non-linear estimation problem, since the parameters in β multiply the parameters in λ .

See Figures 8.9 for an empirical example based on the FF model, and Figure 8.10 for a comparison with the results from the time series approach.

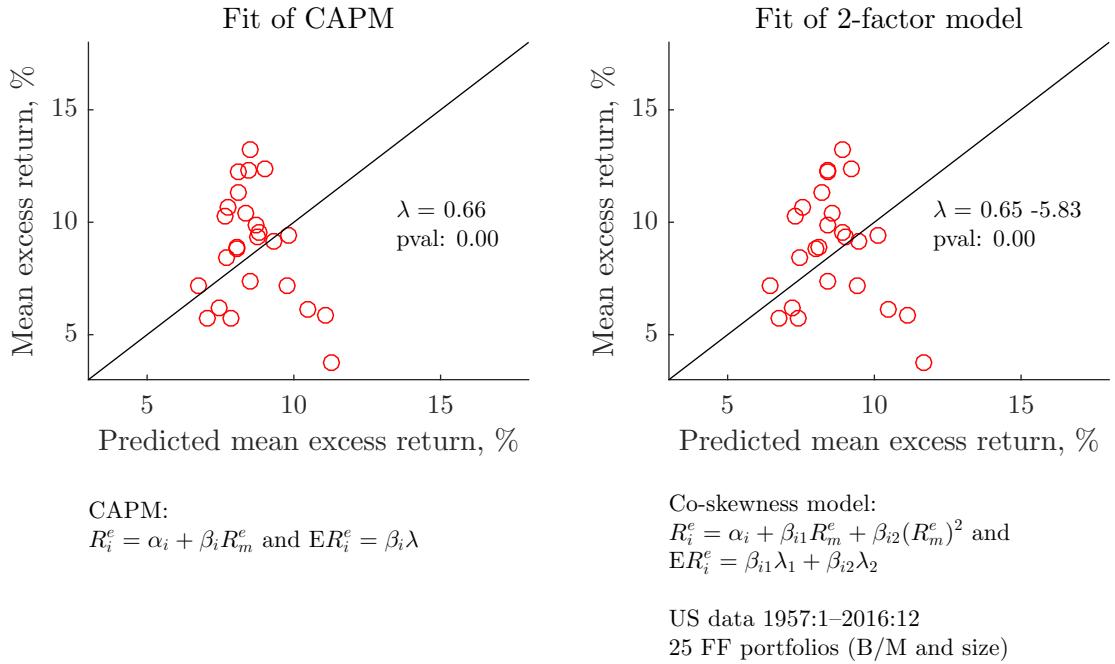


Figure 8.11: CAPM and quadratic model (co-skewness model)

See Figures 8.11–8.13 for an empirical example of a quadratic (co-skewness) model.

One approach to estimate the model is to specify a weighting matrix W and then solve a minimization problem like (8.25). 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 8.3. For other choices of the weighting matrix, the expression for the covariance matrix is more complicated.

It is straightforward to show that the Jacobian of these moment conditions (with respect to $\text{vec}(\alpha, \beta, \lambda)$) is

$$D_0 = - \begin{bmatrix} \frac{1}{T} \sum_{t=1}^T \left(\begin{bmatrix} 1 \\ f_t \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ f_t \\ \lambda' \end{bmatrix}' \right) \otimes I_n & \mathbf{0}_{n(1+K) \times K} \\ \mathbf{0}_{n \times K} & \beta_{n \times K} \end{bmatrix} \quad (8.37)$$

where the upper left block is similar to the expression for the case with excess return factors (8.12), while the other blocks are new.

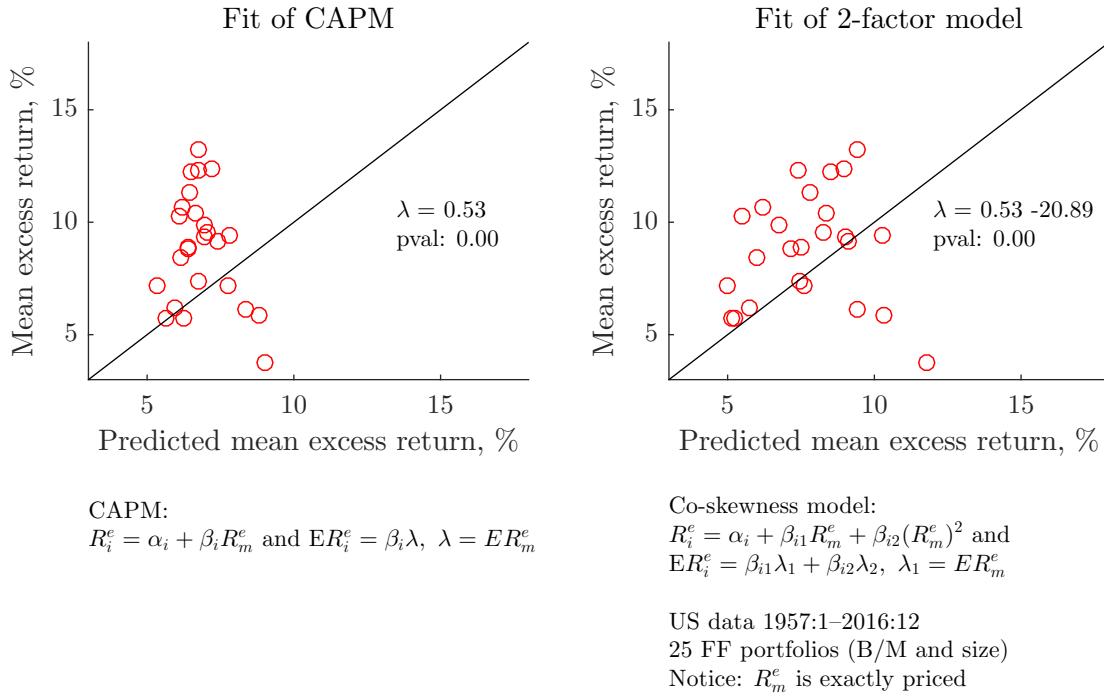


Figure 8.12: CAPM and quadratic model (co-skewness model) when the market excess is exactly priced

Example 8.14 (*Two assets and one factor*) we have the moment conditions

$$E g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = E \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ R_{1t}^e - \beta_1 \lambda \\ R_{2t}^e - \beta_2 \lambda \end{bmatrix} = \mathbf{0}_{6 \times 1}.$$

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

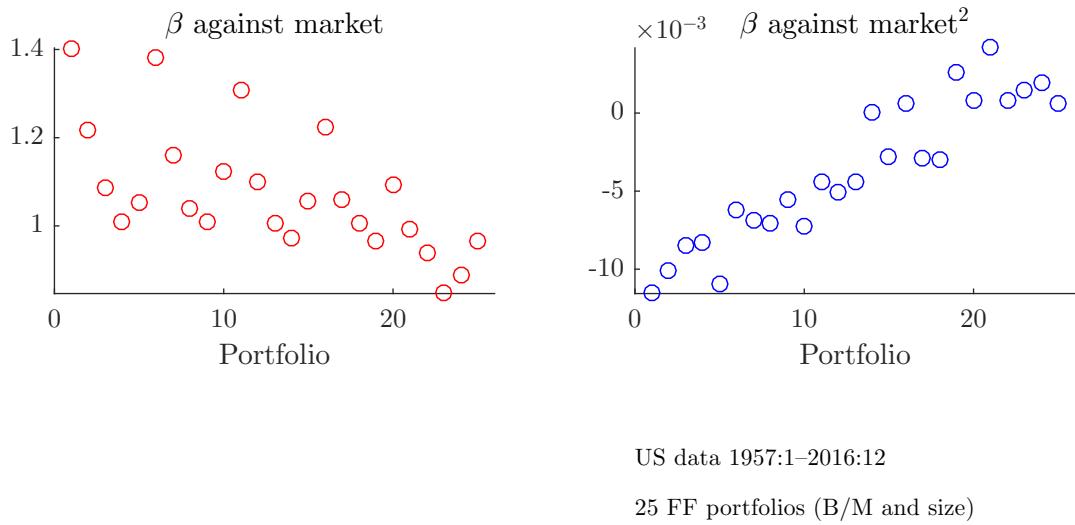


Figure 8.13: CAPM and quadratic model

than factors. In this case, the Jacobian is

$$\begin{aligned} \frac{\partial \bar{g}}{\partial [\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda]'} &= -\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & 0 & f_t & 0 & 0 \\ 0 & 1 & 0 & f_t & 0 \\ f_t & 0 & f_t^2 & 0 & 0 \\ 0 & f_t & 0 & f_t^2 & 0 \\ 0 & 0 & \lambda & 0 & \beta_1 \\ 0 & 0 & 0 & \lambda & \beta_2 \end{bmatrix} \\ &= - \begin{bmatrix} \frac{1}{T} \sum_{t=1}^T \left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix}' \right) \otimes I_2 & \mathbf{0}_{4 \times 1} \\ [0, \lambda] \otimes I_2 & \beta \end{bmatrix}. \end{aligned}$$

8.5.2 Traditional Cross-Sectional Regressions as a Special Case

Instead of estimating the overidentified model (8.36) (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 $A E g_t = \mathbf{0}$. This does not generate any overidentifying restrictions, but it still allows us to test hypotheses about the 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

$$A \mathbf{E} g_t = \mathbf{0}_{[n(1+K)+K] \times 1} \quad (8.38)$$

$$\begin{bmatrix} I_{n(1+K)} & \mathbf{0}_{n(1+K) \times n} \\ \mathbf{0}_{K \times n(1+K)} & \theta_{K \times n} \end{bmatrix} \mathbf{E} \begin{bmatrix} 1 \\ f_t \\ R_t^e - \beta\lambda \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) = \mathbf{0} \quad (8.39)$$

$$\mathbf{E} \begin{bmatrix} 1 \\ f_t \\ \theta(R_t^e - \beta\lambda) \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) = \mathbf{0} \quad (8.40)$$

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 8.15 (*Calculation of the estimates based on (8.39)*) 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

$$\theta \mathbf{E}(R_t^e - \beta\lambda) = \mathbf{0}_{K \times 1} \text{ or } \lambda = (\theta\beta)^{-1} \theta \mathbf{E} R_t^e,$$

which is just like a cross-sectional instrumental variables regression of $\mathbf{E} R_t^e = \beta\lambda$ (with β being the regressors, θ the instruments, and $\mathbf{E} R_t^e$ the dependent variable).

With $\theta = \beta'$, we get the traditional cross-sectional approach (8.33). 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 estimation of the factor model. Then, using $\theta = \beta' \Sigma^{-1}$ gives a traditional GLS cross-sectional approach.

See Table 10.2 shows estimates of the factor risk premia from several methods based on the 25 FF portfolios.

To test the asset pricing implications, we test if the moment conditions $\mathbf{E} g_t = \mathbf{0}$ in (8.38) 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 8.3).

Example 8.16 (*LS cross-sectional regression, two assets and one factor*) With the mo-

	Data	CR	FMB1	FMB2
Rm	6.34 (1.95)	5.98 (2.05)	5.98 (1.97)	-8.51 (3.55)
SMB	2.46 (1.33)	2.30 (1.41)	2.30 (1.37)	1.98 (1.37)
HML	4.38 (1.22)	4.97 (1.36)	4.97 (1.26)	4.53 (1.26)

Table 8.2: Different estimates of factor risk premia, annualized %. Numbers in (parentheses) are standard deviations. The 25 FF portfolios 1957:1-2016:12. Data are the mean excess returns of the factors; CR are estimates of the factor risk premia from a cross-sectional regression; FMB1 are from Fama-MacBeth without intercept in the cross-sectional regression; FMB2 are from Fama-MacBeth with intercept in the cross-sectional regression. In both FMB regressions, the betas are estimated from the full sample.

ment conditions in Example (8.14) and the weighting vector $\theta = [\beta_1, \beta_2]$ (8.40) is

$$A E g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = E \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ \beta_1(R_{1t}^e - \beta_1 \lambda) + \beta_2(R_{2t}^e - \beta_2 \lambda) \end{bmatrix} = \mathbf{0}_{5 \times 1},$$

which has as many parameters as moment conditions. The test of the asset pricing model is then to test if

$$E g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = E \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ R_{1t}^e - \beta_1 \lambda \\ R_{2t}^e - \beta_2 \lambda \end{bmatrix} = \mathbf{0}_{6 \times 1},$$

are satisfied at the estimated parameters.

Example 8.17 (Structure of $\theta E(R_t^e - \beta \lambda)$) If there are 2 factors and three test assets,

then $0_{2 \times 1} = \theta E(R_t^e - \beta\lambda)$ is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \left(\begin{bmatrix} E R_{1t}^e \\ E R_{2t}^e \\ E R_{3t}^e \end{bmatrix} - \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \right).$$

8.5.3 What If the Factors Are Excess Returns?*

It would (perhaps) be natural if the tests discussed in this section coincided with those in Section 8.4 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 8.4 impose $\lambda = E f_t$. This can be done in (8.39)–(8.40) by doing two things. First, define a new set of test assets by stacking the original test assets and the excess return factors

$$\tilde{R}_t^e = \begin{bmatrix} R_t^e \\ f_t \end{bmatrix}, \quad (8.41)$$

which is an $(n + K) \times 1$ vector. Second, define the $K \times (n + K)$ matrix θ as

$$\tilde{\theta} = \begin{bmatrix} \mathbf{0}_{K \times n} & I_K \end{bmatrix}. \quad (8.42)$$

(Clearly, the betas of f_t (as test assets) must equal I_K and their residuals must be zero. This means that the GLS approach to (8.40), $\theta = \beta' \Sigma^{-1}$, is conceptually the same as (8.42), since all weight is on the betas of f_t . However, (8.42) numerically more robust.) Together, this gives

$$\lambda = E f_t. \quad (8.43)$$

It is also straightforward to show that this gives precisely the same test statistics as the Wald test on the multifactor model (8.21).

Proof. (of (8.43)) 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 $\theta E(R_t^e - \beta\lambda) = \mathbf{0}$ is then

$$\begin{bmatrix} \mathbf{0}_{K \times n} & I_K \end{bmatrix} E \begin{bmatrix} R_t^e \\ f_t \end{bmatrix} - \begin{bmatrix} \mathbf{0}_{K \times n} & I_K \end{bmatrix} \begin{bmatrix} \beta_{n \times K} \\ I_K \end{bmatrix} \lambda = \mathbf{0}, \text{ or} \\ E f_t = \lambda.$$

■

Remark 8.18 (*Two assets, one excess return factor*) By including the factors among the test assets and using the weighting vector $\theta = [0, 0, 1]$ gives

$$A \mathbb{E} g_t(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \lambda) = \mathbb{E} \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t - \alpha_3 - \beta_3 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ f_t(f_t - \alpha_3 - \beta_3 f_t) \\ 0(R_{1t}^e - \beta_1 \lambda) + 0(R_{2t}^e - \beta_2 \lambda) + 1(f_t - \beta_3 \lambda) \end{bmatrix} = \mathbf{0}_{7 \times 1}.$$

Since $\alpha_3 = 0$ and $\beta_3 = 1$, this gives the estimate $\lambda = \mathbb{E} f_t$. 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

$$\mathbb{E} g_t(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \lambda) = \mathbb{E} \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t - \alpha_3 - \beta_3 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ f_t(f_t - \alpha_3 - \beta_3 f_t) \\ R_{1t}^e - \beta_1 \lambda \\ R_{2t}^e - \beta_2 \lambda \\ f_t - \beta_3 \lambda \end{bmatrix} = \mathbf{0}_{9 \times 1}.$$

In fact, this gives the same test statistic as when testing if α_1 and α_2 are zero in (8.11).

Remark 8.19 (*What is an excess return?**) Short answer: the return of a zero cost portfolio. More detailed answer: consider a portfolio with the (net) return

$$R_p = v_1 R_1 + v_2 R_2 + v_3 R_3 + (1 - v_1 - v_2 - v_3) R_4,$$

where v_i is the portfolio weight on asset i which has the net return R_i . The balance $(1 - v_1 - v_2 - v_3)$ is made up of asset 4 with the net return R_4 (which may be a riskfree asset). Rearrange as

$$R_p - R_4 = v_1 (R_1 - R_4) + v_2 (R_2 - R_4) + v_3 (R_3 - R_4).$$

Clearly, $R_p - R_4$ is an excess return—and it is a linear combination of other excess returns (even if v_1 , v_2 and/or v_3 happen to be negative and they do not sum to unity). If $v_3 = -v_2$, then we can rearrange further to get

$$R_p - R_4 = v_1(R_1 - R_4) + v_2(R_2 - R_3).$$

This is still an excess return, although the “excess” on the right hand side is over different returns. When we use excess returns as factors, then we typically require the portfolio weights (see above) to be constant over time.

When Some (but Not All) of the Factors Are Excess Returns*

Partition the vector of factors as

$$f_t = \begin{bmatrix} Z_t \\ F_t \end{bmatrix}, \quad (8.44)$$

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 (8.39)–(8.40) by doing two things. First, define a new set of test assets by stacking the original test assets and the excess return factors

$$\tilde{R}_t^e = \begin{bmatrix} R_t^e \\ Z_t \end{bmatrix}, \quad (8.45)$$

which is an $(n + v) \times 1$ vector. Second, define the $K \times (n + K)$ matrix θ

$$\tilde{\theta} = \begin{bmatrix} \mathbf{0}_{v \times n} & I_v \\ \vartheta_{w \times n} & \mathbf{0}_{w \times v} \end{bmatrix}, \quad (8.46)$$

where ϑ is some $w \times n$ matrix. Together, this ensures that

$$\tilde{\lambda} = \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} = \begin{bmatrix} \mathbb{E} Z_t \\ (\vartheta \beta^F)^{-1} \vartheta (\mathbb{E} R_t^e - \beta^Z \lambda_Z) \end{bmatrix}, \quad (8.47)$$

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

$$\beta_{n \times K} = \begin{bmatrix} \beta_{n \times v}^Z & \beta_{n \times w}^F \end{bmatrix}. \quad (8.48)$$

One possible choice of ϑ is $\vartheta = \beta^F$, since then λ_F are the same as when running a cross-sectional regression of the expected “abnormal return” ($E R_t^e - \beta^Z \lambda_Z$) on the betas (β^F).

Proof. (of (8.47)) The betas of the \tilde{R}_t^e vector are

$$\tilde{\beta} = \begin{bmatrix} \beta_{n \times v}^Z & \beta_{n \times w}^F \\ I_v & 0_{v \times w} \end{bmatrix}.$$

The expression corresponding to $\theta E(R_t^e - \beta \lambda) = \mathbf{0}$ is then

$$\begin{aligned} \tilde{\theta} E \tilde{R}_t^e &= \tilde{\theta} \tilde{\beta} \tilde{\lambda} \\ \begin{bmatrix} \mathbf{0}_{v \times n} & I_v \\ \vartheta_{w \times n} & \mathbf{0}_{w \times v} \end{bmatrix} \begin{bmatrix} E R_t^e \\ E Z_t \end{bmatrix} &= \begin{bmatrix} \mathbf{0}_{v \times n} & I_v \\ \vartheta_{w \times n} & \mathbf{0}_{w \times v} \end{bmatrix} \begin{bmatrix} \beta_{n \times v}^Z & \beta_{n \times w}^F \\ I_v & 0_{v \times w} \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} \\ \begin{bmatrix} E Z_t \\ \vartheta_{w \times n} E R_t^e \end{bmatrix} &= \begin{bmatrix} I_v & \mathbf{0}_{v \times w} \\ \vartheta_{w \times n} \beta_{n \times v}^Z & \vartheta_{w \times n} \beta_{n \times w}^F \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix}. \end{aligned}$$

The first v equations give

$$\lambda_Z = E Z_t.$$

The remaining w equations give

$$\begin{aligned} \vartheta E R_t^e &= \vartheta \beta^Z \lambda_Z + \vartheta \beta^F \lambda_F, \text{ so} \\ \lambda_F &= (\vartheta \beta^F)^{-1} \vartheta (E R_t^e - \beta^Z \lambda_Z). \end{aligned}$$

■

Example 8.20 (*Structure of θ to identify λ for excess return factors*) Continue Example 8.17 (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 0.) If we set $(\theta_{11}, \theta_{12}, \theta_{13}) = (0, 0, 1)$, then the moment conditions in Example 8.17 can be written

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \left(\begin{bmatrix} E R_{1t}^e \\ E R_{2t}^e \\ E Z_t \end{bmatrix} - \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} \right).$$

The first line reads

$$0 = \mathbb{E} Z_t - \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix}, \text{ so } \lambda_Z = \mathbb{E} Z_t.$$

8.5.4 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 (2001b) estimate models where consumption growth is the factor—with mixed results.

8.6 Linear SDF Models

This section discusses how we can estimate and test the asset pricing equation

$$\mathbb{E} m_t R_t^e = 0. \quad (8.49)$$

Assume that the SDF is linear in the factors

$$m_t = \bar{m} + b'(f_t - \mathbb{E} f_t), \quad (8.50)$$

where the $K \times 1$ vector f_t contains the factors and where $\bar{m} \neq 0$. Combining with (8.49) gives the moment conditions

$$g_t(b) = m_t R_t^e = \bar{m} R_t^e + b'(f_t - \bar{f}_t) R_t^e, \quad (8.51)$$

since m_t is a scalar. There are K parameters (in b) and n moment conditions (the number of assets). The mean of the SDF cannot be estimated from excess returns (it could if we used returns), but it is straightforward to show that the choice of \bar{m} (as long as not zero) does not matter for the test based on excess returns.

Remark 8.21 (*The SDF model and the mean SDF*) Take expectations of the moment conditions (8.51) and set equal to zero to get

$$b' \text{Cov}(f_t, R_t^e) = -\bar{m} \mathbb{E} R_t^e.$$

This would be satisfied by $(\bar{m}, b) = (0, \mathbf{0})$, which makes no sense. Instead, for any $\bar{m} \neq 0$,

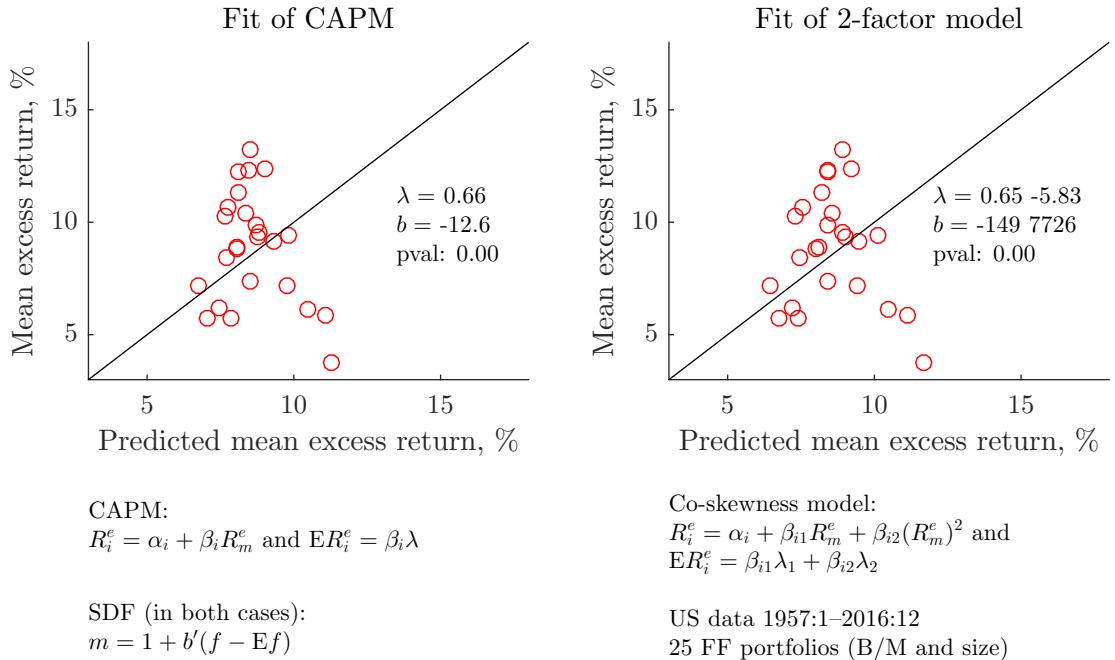


Figure 8.14: CAPM and quadratic model (co-skewness)

we could have

$$E R_t^e = \frac{-1}{\bar{m}} b' \text{Cov}(f_t, R_t^e),$$

which allows us to test if there is a $K \times 1$ vector b that prices all n assets, given how the covariance matrix of the returns and factors looks like.

To estimate this model with a weighting matrix W , we minimize the loss function

$$J = \bar{g}(b)' W \bar{g}(b). \quad (8.52)$$

Alternatively, the moment conditions are combined into K effective conditions as

$$A_{K \times n} \bar{g}(b) = \mathbf{0}_{K \times 1}. \quad (8.53)$$

To test the asset pricing implications, we test if the moment conditions $E g_t = \mathbf{0}$ 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.

8.6.1 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):

$$m_t = \bar{m} + (R_{1t}^e - \bar{R}_{1t}^e)' b, \quad (8.54)$$

where \bar{m} is a constant and b is chosen to make m_t “price” R_{1t}^e in the sample, that is, so

$$\sum_{t=1}^T \mathbb{E} R_{1t}^e m_t / T = \mathbf{0}. \quad (8.55)$$

Consider the test assets with excess returns R_{2t}^e , and “SDF-based performance”

$$\bar{g}_{2t} = \frac{1}{T} \sum_{t=1}^T R_{2t}^e m_t. \quad (8.56)$$

Compare with the linear factor portfolio model

$$R_{2t}^e = \alpha + \beta R_{1t}^e + \varepsilon_t, \quad (8.57)$$

(where $\mathbb{E} \varepsilon_t = \mathbf{0}$ and $\text{Cov}(R_{1t}^e, \varepsilon_t) = \mathbf{0}$) to see that the SDF-performance (“pricing error”) is proportional to a traditional alpha

$$\bar{g}_{2t}/\bar{m} = \hat{\alpha}. \quad (8.58)$$

In both cases we are thus testing if α is zero or not.

Proof. (of (8.58)) (Here written in terms of population moments, to simplify the notation.) It follows directly that $b = -\text{Var}(R_{1t}^e)^{-1} (\mathbb{E} R_{1t}^e \bar{m})$. Using this and the expression

for m_t in (8.56) gives

$$\mathbb{E} g_{2t} = \mathbb{E} R_{2t}^e \bar{m} - \text{Cov}(R_{2t}^e, R_{1t}^e) \text{Var}(R_{1t}^e)^{-1} \mathbb{E} R_{1t}^e \bar{m}.$$

We now rewrite this equation in terms of the parameters in the factor portfolio model (8.57). The latter implies $\mathbb{E} R_{2t}^e = \alpha + \beta \mathbb{E} R_{1t}^e$, and the least squares estimator of the slope coefficients is $\hat{\beta} = \text{Cov}(R_{2t}^e, R_{1t}^e) \text{Var}(R_{1t}^e)^{-1}$. Using these two facts in the equation above—and replacing population moments with sample moments, gives (8.58). ■

8.7 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 functions of them (this is sometimes called “scaled factors” to indicate that we scale the original factors with instruments). For instance, if R_{mt}^e 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

$$f_{1t} = R_{mt}^e \text{ and } f_{2t} = z_{t-1} R_{mt}^e. \quad (8.59)$$

Since the second factor is not an excess return, the test is done as in (8.36).

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 (8.36), $R_{it}^e = \alpha + \beta f_t + \varepsilon_{it}$,

$$\begin{aligned} R_{it}^e &= \alpha + \beta_1 R_{mt}^e + \beta_2 z_{t-1} R_{mt}^e + \varepsilon_{it} \\ &= \alpha + (\beta_1 + \beta_2 z_{t-1}) R_{mt}^e + \varepsilon_{it}. \end{aligned} \quad (8.60)$$

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 $z_{t-1} R_{mt}^e$ is not a traditional excess return.

See Figure 8.15–8.16 for an empirical illustration.

Remark 8.22 (Figures 8.15–8.16, equally weighted 25 FF portfolios) Figure 8.15 shows

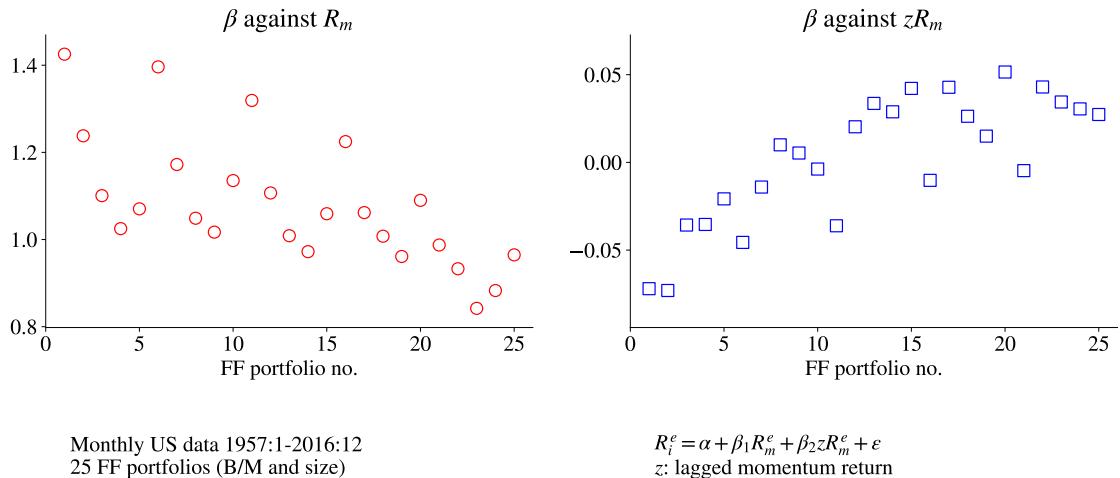


Figure 8.15: Conditional betas of the 25 FF portfolios

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 8.16.

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 (8.36).

Remark 8.23 (*Dynamic Portfolios**) The returns on our factors, f_t , could be the excess return on dynamic portfolios, $R_{1t}^e = s_{t-1} \otimes R_{0t}^e$, where s_{t-1} are some information variables (not payoffs as before), for instance, lagged returns or market volatility, and R_{0t}^e are some basic benchmarks (S&P500 and bond, perhaps). The reason is that if R_{0t}^e are excess returns, so are $R_{1t}^e = s_{t-1} \otimes R_{0t}^e$. Therefore, the typical cross-sectional test (of $E R^e = \beta' \lambda$) coincides with the test of the alpha—and also of zero SDF pricing errors. Notice also that the returns of our test assets, R_{it}^e , could be the excess return on dynamic strategies in terms of some basic test assets (mutual funds, say), $R_{2t}^e = z_{t-1} \otimes R_{pt}^e$, where z_{t-1} are information variables and R_{pt}^e are basic test assets. In this case, we are testing the performance of these dynamic strategies.

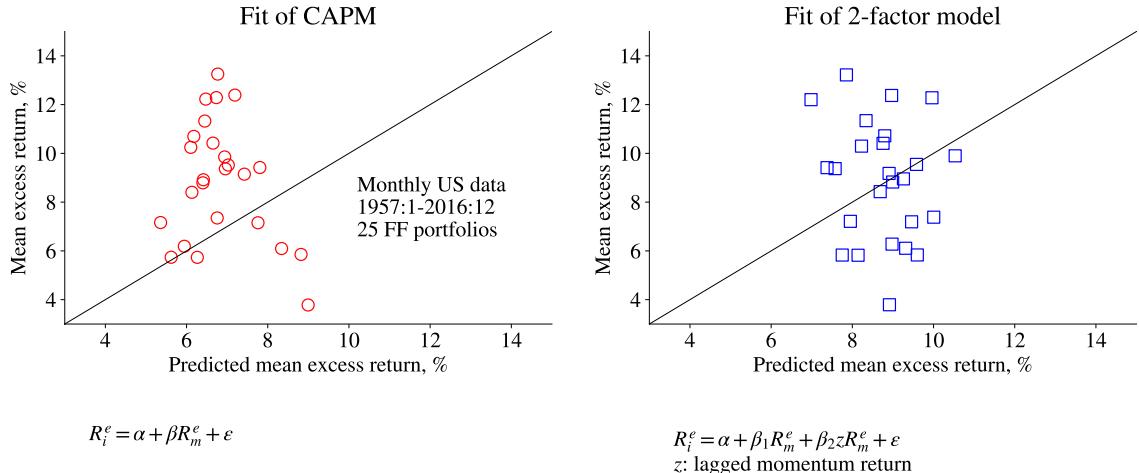


Figure 8.16: Unconditional and conditional CAPM tests of the 25 FF portfolios

8.8 Conditional Models with “Regimes”

Reference: Christiansen, Ranaldo, and Söderlind (2011)

It is also possible to estimate non-linear factor models. The model could be piecewise linear or include higher order terms. For instance, Treynor and Mazuy (1966) extend the CAPM regression by including a squared term (of the market excess return) to capture market timing.

Alternatively, the conditional model (8.60) could be changed so that the time-varying 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

$$G(z) = \frac{1}{1 + \exp[-\gamma(z - c)]}, \quad (8.61)$$

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 8.17 for an illustration. A logistic smooth transition regression

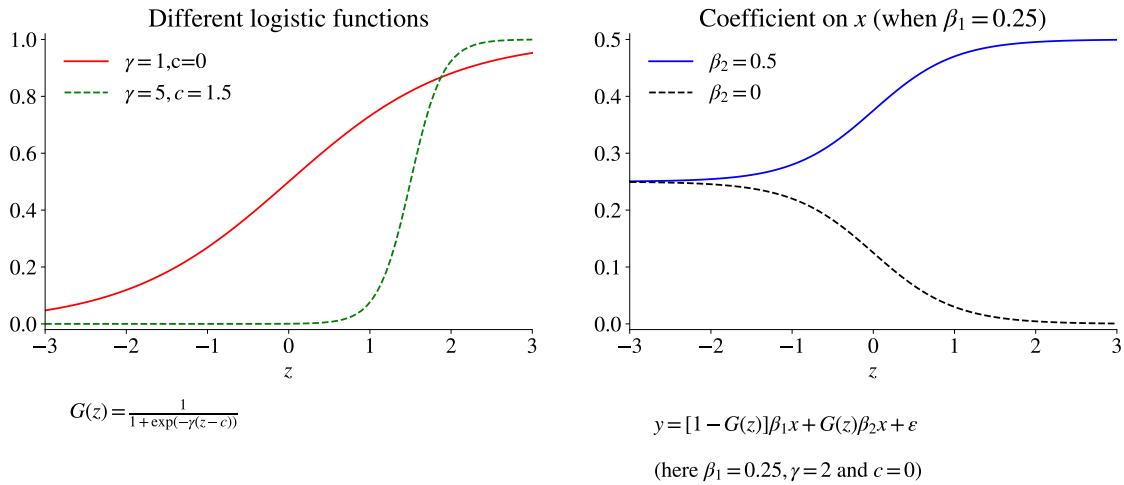


Figure 8.17: Logistic function and the effective slope coefficient in a Logistic smooth transition regression

is

$$\begin{aligned}
 R_{it}^e &= \beta(z_{t-1})' x_t + \varepsilon_t \\
 &= \{[1 - G(z_{t-1})] \beta'_1 + G(z_{t-1}) \beta'_2\} x_t + \varepsilon_t \\
 &= [1 - G(z_{t-1})] \beta'_1 x_t + G(z_{t-1}) \beta'_2 x_t + \varepsilon_t.
 \end{aligned} \tag{8.62}$$

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 8.17 for an illustration.

Remark 8.24 (*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” $([1 - G(z_{t-1})] x_t, G(z_{t-1}) x_t)$.*

The most common application of this model is by letting $x_t = R_{i,t-s}^e$. 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 8.18–8.19.

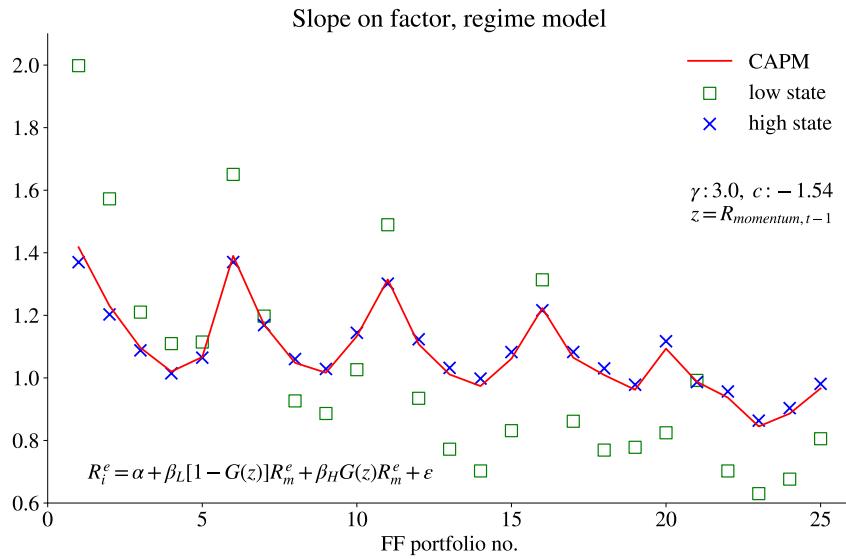


Figure 8.18: Betas on the market in the low and high regimes, 25 FF portfolios

8.9 Fama-MacBeth*

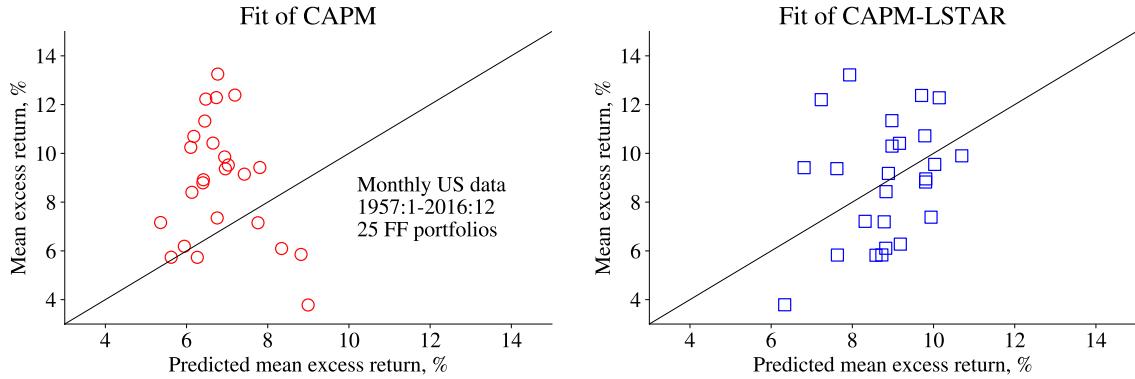
Reference: Cochrane (2005) 12.3; Campbell, Lo, and MacKinlay (1997) 5.8; Fama and MacBeth (1973)

The Fama and MacBeth (1973) approach (called FMB below) is a bit different from the regression approaches discussed so far—although it seems most related to what we discussed in Section 8.5. The method has three steps, described below.

- First, estimate the betas β_i ($i = 1, \dots, n$) from (8.1) (this is a time-series regression). This is often done on the whole sample—assuming the betas are constant. Sometimes, the betas are estimated separately for different sub samples (so we could let $\hat{\beta}_i$ carry a time subscript in the equations below).
- Second, run a cross sectional regression for every t . That is, for period t , estimate λ_t from the cross section (across the assets $i = 1, \dots, n$) regression

$$R_{it}^e = \gamma_t + \lambda'_t \hat{\beta}_i + \varepsilon_{it}, \quad (8.63)$$

where $\hat{\beta}_i$ are the regressors. Note the difference to the traditional cross-sectional approach discussed in (8.8), where the second stage regression regressed \bar{R}_{it}^e (the time-series average of R_{it}^e) on $\hat{\beta}_i$, while the Fama-French approach runs one re-



$$R_i^e = \alpha + \beta R_m^e + \epsilon$$

$$R_i^e = \alpha + \beta_L [1 - G(z)] R_m^e + \beta_H G(z) R_m^e + \epsilon$$

z : lagged momentum return

Figure 8.19: Test of 1 and 2-factor models, 25 FF portfolios

gression for every time period. The intercept γ_t (which capture time-fixed effects) is often dropped from the regression.

- Third, estimate the time averages

$$\hat{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_{it} \text{ for } i = 1, \dots, n, \text{ (for every asset)} \quad (8.64)$$

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^T \hat{\lambda}_t. \quad (8.65)$$

Since $\hat{\lambda}_t$ measures the cross-sectional effect, $\hat{\lambda}$ is just the average of the cross-sectional effect. How R_{it}^e for a given i varies across time as β_{it} does does not affect the estimate.

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.

Remark 8.25 (*Fama-MacBeth with constant betas*) If the betas are (restricted to be) con-

stant across time, then the estimate $\hat{\lambda}$ from (8.65) without intercept (exclude the γ_t term) is the same as from the traditional cross-sectional regression (8.35). To see that, consider the simplifying case of only one factor (so $\hat{\beta}_i$ is a scalar). Then, the FMB from the second step regression (8.63) without an intercept gives $\hat{\lambda}_t = (\sum_{i=1}^n \hat{\beta}_i R_{it}^e) / (\sum_{i=1}^n \hat{\beta}_i^2)$. Notice that the denominator is the same across time, so we can calculate the time-average (8.65) as

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^T \frac{\sum_{i=1}^n \hat{\beta}_i R_{it}^e}{\sum_{i=1}^n \hat{\beta}_i^2} = \frac{\sum_{i=1}^n \hat{\beta}_i \bar{R}_{it}^e}{\sum_{i=1}^n \hat{\beta}_i^2},$$

which is the same as from the CR approach.

We clearly want portfolios which have different betas, or else the second step regression (8.63) 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 (8.64) that ε_i is the time average of the residual for asset i , ε_{it}), by forming a t-test $\hat{\varepsilon}_i / \text{Std}(\hat{\varepsilon}_i)$. Fama and MacBeth (1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\varepsilon}_{it}$. In particular, they suggest that the variance of $\hat{\varepsilon}_{it}$ (not $\hat{\varepsilon}_i$) can be estimated by the (average) squared variation around its mean

$$\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T} \sum_{t=1}^T (\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2. \quad (8.66)$$

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,

$$\text{Var}(\hat{\varepsilon}_i) = \frac{1}{T} \text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2. \quad (8.67)$$

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

$$\text{Var}(\hat{\lambda}) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\lambda}_t - \hat{\lambda})^2. \quad (8.68)$$

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.

The approach can also be extended to include other variables in the cross-sectional regressions, so (8.63) would become

$$R_{it}^e = \gamma_t + \lambda'_t \begin{bmatrix} \hat{\beta}_i \\ z_{it} \end{bmatrix} + \varepsilon_{it}, \quad (8.69)$$

where z_{it} could be a vector of asset (i) specific characteristics in period t (for instance, the leverage). Testing the λ coefficients of z_{it} is done in the same way as before. It can be noticed that when z_{it} is time-varying, then the FMB approach is not the same as OLS on pooled data. In fact, FMB is focused on the average cross-sectional effect, not on the time-series effect. (Actually, regressions where all fixed effects have been taken out by demeaning are the same in FMB and pooled OLS.)

8.10 Appendix: Details of CAPM Regression

Proof. (of (8.2)) 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

$$\sqrt{T}(\hat{b}_{LS} - b_0) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Sigma_{xx}^{-1}), \text{ where } \sigma^2 = E u_t^2 \text{ and } \Sigma_{xx} = E \Sigma_{t=1}^T x_t x'_t / T.$$

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

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

(In the last line we use $\text{Var}(f_t) = E f_t^2 - (E f_t)^2$.) The upper left cell is (8.2). ■

Proof. (of (8.4)) From the CAPM regression (8.1) 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 $\mu^{e'} \Sigma^{-1} \mu^e$, 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 (8.2) and to get (8.4). ■

8.11 Appendix: Details of SURE Systems

Proof. (of (8.6)) Write each of the regression equations in (8.5) 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 $\hat{\theta}_1, \dots, \hat{\theta}_n$. As in (8.2), 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 SURE coefficients, used in proof of (8.6)*) 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

$$\begin{aligned}\hat{\beta} &= \beta + \frac{1}{\sum_{t=1}^T x_t x_t} (x_1 u_1 + x_2 u_2 + \dots x_T u_T) \\ \hat{\gamma} &= \gamma + \frac{1}{\sum_{t=1}^T x_t x_t} (x_1 v_1 + x_2 v_2 + \dots x_T v_T).\end{aligned}$$

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

$$\begin{aligned}\text{Cov}(\hat{\beta}, \hat{\gamma}) &= \left(\frac{1}{\sum_{t=1}^T x_t x_t} \right)^2 [x_1^2 \text{Cov}(u_1, v_1) + x_2^2 \text{Cov}(u_2, v_2) + \dots x_T^2 \text{Cov}(u_T, v_T)] \\ &= \left(\frac{1}{\sum_{t=1}^T x_t x_t} \right)^2 \left(\sum_{t=1}^T x_t x_t \right) \sigma_{uv}, \text{ where } \sigma_{uv} = \text{Cov}(u_t, v_t), \\ &= \frac{1}{\sum_{t=1}^T x_t x_t} \sigma_{uv}.\end{aligned}$$

Divide and multiply by T to get the result in the proof of (8.6). (We get the same results if we relax the assumption that x_t are fixed numbers, and instead derive the asymptotic distribution.) ■

Remark 8.26 (*General results on SURE distribution, same regressors*) Let the regression equations be

$$y_{it} = x_t' \theta_i + \varepsilon_{it}, i = 1, \dots, n,$$

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 $x_{2t} \varepsilon_t$

$$\mathbb{E} g_t = \mathbb{E}(x_t \otimes \varepsilon_t),$$

then Jacobian (with respect to the coefficients of x_{1t} , then the coefficients 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 $S_0 = \sum_{s=-\infty}^{\infty} \mathbb{E} g_t g_t'$. 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

$$\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \bar{g}_3 \\ \bar{g}_4 \end{bmatrix} = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} y_{1t} - \alpha_1 - \beta_1 f_t \\ y_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(y_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(y_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix},$$

and

$$\begin{aligned} \frac{\partial \bar{g}}{\partial [\alpha_1, \alpha_2, \beta_1, \beta_2]'} &= \begin{bmatrix} \partial \bar{g}_1 / \partial \alpha_1 & \partial \bar{g}_1 / \partial \alpha_2 & \partial \bar{g}_1 / \partial \beta_1 & \partial \bar{g}_1 / \partial \beta_2 \\ \partial \bar{g}_2 / \partial \alpha_1 & \partial \bar{g}_2 / \partial \alpha_2 & \partial \bar{g}_2 / \partial \beta_1 & \partial \bar{g}_2 / \partial \beta_2 \\ \partial \bar{g}_3 / \partial \alpha_1 & \partial \bar{g}_3 / \partial \alpha_2 & \partial \bar{g}_3 / \partial \beta_1 & \partial \bar{g}_3 / \partial \beta_2 \\ \partial \bar{g}_4 / \partial \alpha_1 & \partial \bar{g}_4 / \partial \alpha_2 & \partial \bar{g}_4 / \partial \beta_1 & \partial \bar{g}_4 / \partial \beta_2 \end{bmatrix} \\ &= -\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & 0 & f_t & 0 \\ 0 & 1 & 0 & f_t \\ f_t & 0 & f_t^2 & 0 \\ 0 & f_t & 0 & f_t^2 \end{bmatrix} = \left(-\frac{1}{T} \sum_{t=1}^T x_t x_t' \right) \otimes I_2. \end{aligned}$$

Remark 8.27 (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

$$E g_t = E(\varepsilon_t \otimes x_t),$$

then the Jacobian (wrt the coefficients in regression 1, then the coefficients 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 8.26 gives

$$\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \bar{g}_3 \\ \bar{g}_4 \end{bmatrix} = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} y_{1t} - \alpha_1 - \beta_1 f_t \\ f_t(y_{1t} - \alpha_1 - \beta_1 f_t) \\ y_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(y_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix},$$

and

$$\begin{aligned}
\frac{\partial \bar{g}}{\partial [\alpha_1, \beta_1, \alpha_2, \beta_2]'} &= \begin{bmatrix} \partial \bar{g}_1 / \partial \alpha_1 & \partial \bar{g}_1 / \partial \beta_1 & \partial \bar{g}_1 / \partial \alpha_2 & \partial \bar{g}_1 / \partial \beta_2 \\ \partial \bar{g}_2 / \partial \alpha_1 & \partial \bar{g}_2 / \partial \beta_1 & \partial \bar{g}_2 / \partial \alpha_2 & \partial \bar{g}_2 / \partial \beta_2 \\ \partial \bar{g}_3 / \partial \alpha_1 & \partial \bar{g}_3 / \partial \beta_1 & \partial \bar{g}_3 / \partial \alpha_2 & \partial \bar{g}_3 / \partial \beta_2 \\ \partial \bar{g}_4 / \partial \alpha_1 & \partial \bar{g}_4 / \partial \beta_1 & \partial \bar{g}_4 / \partial \alpha_2 & \partial \bar{g}_4 / \partial \beta_2 \end{bmatrix} \\
&= -\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & f_t & 0 & 0 \\ f_t & f_t^2 & 0 & 0 \\ 0 & 0 & 1 & f_t \\ 0 & 0 & f_t & f_t^2 \end{bmatrix} = I_2 \otimes \left(-\frac{1}{T} \sum_{t=1}^T x_t x_t' \right).
\end{aligned}$$

Chapter 9

Consumption-Based Asset Pricing

Reference: Bossaert (2002); Campbell (2003); Cochrane (2005)

9.1 Consumption-Based Asset Pricing

9.1.1 The Basic Asset Pricing Equation

The basic asset pricing equation says

$$E_{t-1} R_t m_t = 1. \quad (9.1)$$

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.

In a consumption-based model, (9.1) is the Euler equation for optimal saving in $t - 1$ where m_t is the ratio of marginal utilities in t and $t - 1$, $m_t = \delta u'(C_t)/u'(C_{t-1})$. where δ is the time discounting (“impatience”). 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 (so they would not affect marginal utility). With constant relative risk aversion (CRRA) γ , the stochastic discount factor is

$$m_t = \delta(C_t/C_{t-1})^{-\gamma}, \text{ so} \quad (9.2)$$

$$\ln m_t = \ln \delta - \gamma \Delta c_t, \text{ where } \Delta c_t = \ln C_t/C_{t-1}. \quad (9.3)$$

The second line is only there to introduce the convenient notation Δc_t for the consumption

growth rate. Notice that this defines a real (not nominal) SDF—which should be related to real (not nominal) excess return. In practice, real and nominal excess returns are very similar (and for log returns they are the same).

The next few sections study if the pricing model consisting of (9.1) and (9.2) 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. Second, complete markets are not assumed. The key assumption is rather that the basic asset pricing equation (9.1) 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 risk premium as

$$E_{t-1} R_t^e = \text{Cov}_{t-1}(R_t^e, \Delta c_t)\gamma. \quad (9.4)$$

The intuition for this expression 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 (9.1).)

We can relax the assumption that the excess return is normally distributed: (9.4) 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 9.1.1 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 9.1 (*Stein's lemma*) If x and y have a bivariate normal distribution and $h(y)$ is a differentiable function such that $E[|h'(y)|] < \infty$, then $\text{Cov}[x, h(y)] = \text{Cov}(x, y) E[h'(y)]$.

Proof. (of (9.4)) For an excess return R^e , (9.1) says $E R^e m = 0$, so

$$E R^e = -\text{Cov}(R^e, m)/E m.$$

Stein's lemma gives $\text{Cov}[R^e, \exp(\ln m)] = \text{Cov}(R^e, \ln m) E m$. (In terms of Stein's

lemma, $x = R^e$, $y = \ln m$ and $h() = \exp()$. Finally, notice that $\text{Cov}(R^e, \ln m) = -\gamma \text{Cov}(R^e, \Delta c)$. ■

To discuss the historical average excess returns, it is convenient to work with the unconditional version of the pricing expression (9.4)

$$\mathbb{E} R_t^e = \text{Cov}(R_t^e, \Delta c_t)\gamma, \quad (9.5)$$

which is obtained by taking unconditional expectations of both sides of (9.4).

Remark 9.2 (*From covariances to betas*) Divide and multiply (9.5) by $\text{Var}(\Delta c_t)$ to get a beta expression

$$\mathbb{E} R_t^e = \beta\lambda, \text{ where } \beta = \frac{\text{Cov}(R_t^e, \Delta c_t)}{\text{Var}(\Delta c_t)} \text{ and } \lambda = \gamma \text{Var}(\Delta c_t).$$

Clearly, β is the coefficient obtained from regressing R_t^e on Δc_t , and λ is risk factor premium (same for all assets).

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 9.1* 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.

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 9.2* gives an illustration.

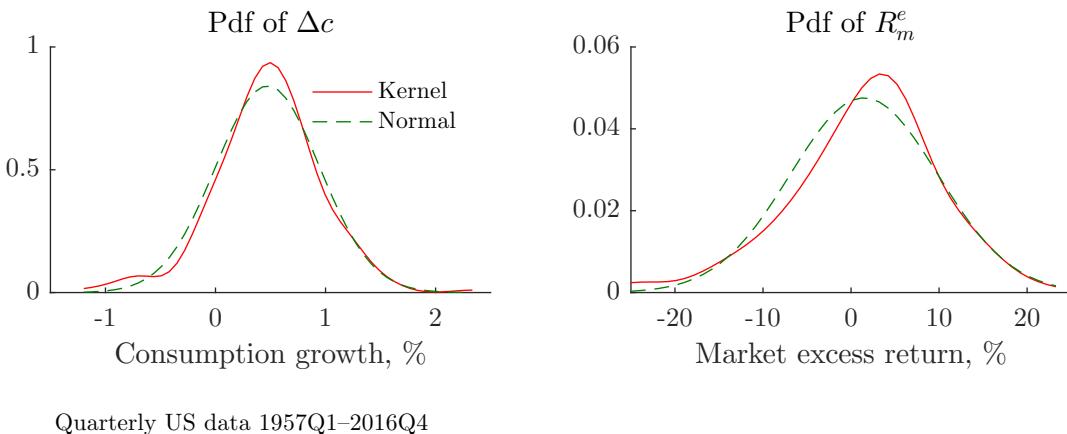


Figure 9.1: **Density functions of consumption growth and equity market excess returns.** The kernel density function of a variable x is estimated by using a $N(0, 1)$ kernel with a bandwidth of $1.06 \text{ Std}(x)T^{-1/5}$. The normal distribution is calculated from the estimated mean and variance of the same variable.

Proposition 9.3 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 $E|h'(y)| < \infty$. Then $\text{Cov}[x, h(y)] = E h'(y) \text{Cov}(x, y)$. (See Söderlind (2009) for a proof.)

9.2 Asset Pricing Puzzles

9.2.1 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, by focusing on the unconditional pricing expression in (9.5).

Table 9.1 shows the key statistics for quarterly US real returns and consumption growth.

	Mean	Std	Autocorr	Corr with Δc
Δc	1.89	0.95	0.45	1.00
R_m^e	6.73	16.74	0.07	0.16
Riskfree	0.85	2.78	0.50	0.20

Table 9.1: US quarterly data, 1957Q1–2016Q4 , (annualized, in %, in real terms)

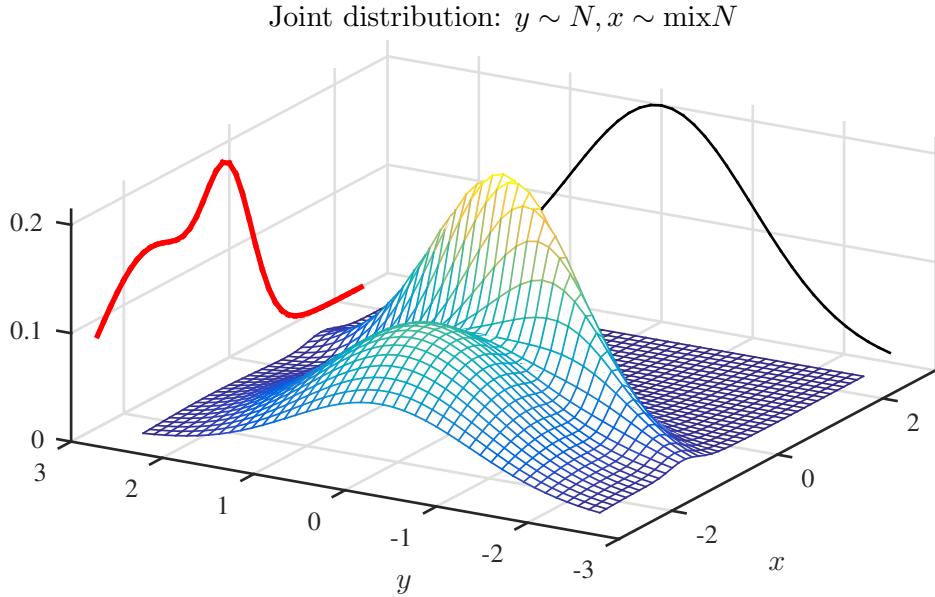


Figure 9.2: **Example of a bivariate mixed-normal distribution where y has a normal marginal distribution.** The marginal distributions are drawn at the back.

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 6.5% (annualized), and that returns are only weakly correlated with consumption growth. These figures will be important in the following sections.

Table 9.1 shows that we can write (9.5) as

$$\mathbb{E} R_t^e = \text{Corr}(R_t^e, \Delta c_t) \times \text{Std}(R_t^e) \times \text{Std}(\Delta c_t) \gamma \quad (9.6)$$

$$0.065 \approx 0.17 \times 0.17 \times 0.01 \gamma. \quad (9.7)$$

which requires a value of $\gamma \approx 225$ 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. Indeed, even if the correlation was one, (9.7) would require $\gamma \approx 38$. This is the *equity premium puzzle* stressed by Mehra and Prescott (1985) (although they approach the issue from another angle).

9.2.2 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.

Some 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, notice that (9.4) says that the conditional expected excess return should equal the conditional covariance times the risk aversion.

The upper left subfigure of Figure 9.3 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 9.3 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.

9.2.3 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

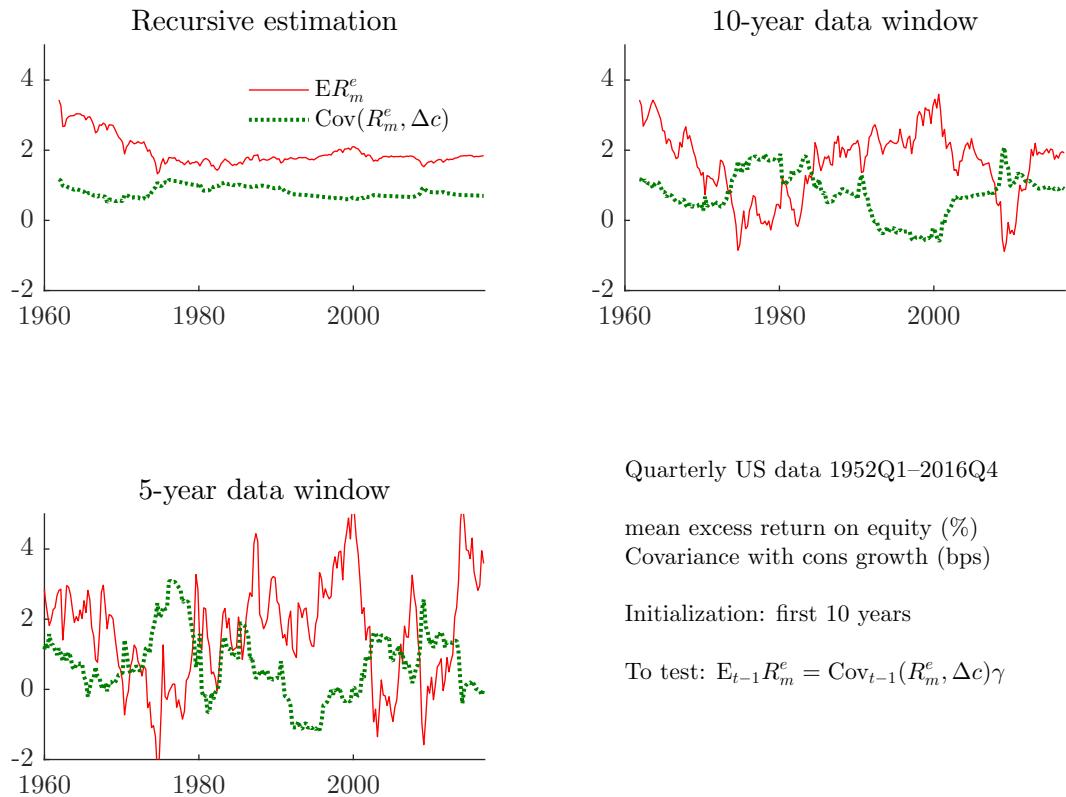


Figure 9.3: The equity premium puzzle for different samples

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 (9.1) we have $E_{t-1}(R_f t)$ $E_{t-1}(m_t) = 1$. 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 9.1 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

¹As 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 $Cov_{t-1}(x_t, y_t) = 0$, but $Cov(x_t, y_t) = \sigma^2(z_{t-1})$.

log expected riskfree rate is

$$r_{ft} \approx -\ln \delta + \gamma E_{t-1} \Delta c_t, \quad (9.8)$$

where r_{ft} is the net real interest rate. To relate this equation to historical data, we take unconditional expectations to get

$$E r_{ft} \approx -\ln \delta + \gamma E \Delta c_t. \quad (9.9)$$

Proof. (of (9.8)) For a riskfree gross return R_{ft} that is known in t , (9.1) with the SDF (9.2) says $R_{ft} E_{t-1} [\delta(C_t/C_{t-1})^{-\gamma}] = 1$. Take logs and replace $\ln R_{ft}$ by r_{ft} . Recall that if $x \sim N(\mu, \sigma^2)$ and $y = \exp(x)$ then $E y = \exp(\mu + \sigma^2/2)$. When Δc_t is conditionally normally distributed, the log of $E_{t-1} [\delta(C_t/C_{t-1})^{-\gamma}]$ equals $\ln \delta - \gamma E_{t-1} \Delta c_t + \gamma^2 \text{Var}_{t-1}(\Delta c_t)/2$. The variance term is typically very small, so we disregard it. ■

According to (9.9) there are two ways to reconcile a positive consumption growth rate with a low real interest rate (around 1% in Table 9.1): investors may prefer to consume later rather than sooner ($\delta > 1$) 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 (9.9) is to hold. For instance, with $\gamma = 25$ (which is a very conservative guess of what we need to fit the equity premium) equation (9.9) says

$$0.01 \approx -\ln \delta + 25 \times 0.02 \quad (9.10)$$

which requires $\delta \approx 1.6$. 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).

9.3 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 (9.5), which I repeat here for ease of reading

$$\mathbb{E} R_t^e = \text{Cov}(R_t^e, \Delta c_t) \gamma.$$

This can be tested with a GMM framework or a 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 9.4 (*GMM estimation of (9.5)*) Let there be N assets. The original moment conditions are

$$\bar{g}(\theta) = \begin{bmatrix} \frac{1}{T} \sum_{t=1}^T (\Delta c_t - \mu_{\Delta c}) = 0 \\ \frac{1}{T} \sum_{t=1}^T (R_{it}^e - \mu_i) = 0 \text{ for } i = 1, 2, \dots, N \\ \frac{1}{T} \sum_{t=1}^T [(\Delta c_t - \mu_c)(R_{it}^e - \mu_i) - \sigma_{ci}] = 0 \text{ for } i = 1, 2, \dots, N \\ \frac{1}{T} \sum_{t=1}^T (R_{it}^e - \alpha - \sigma_{ci}\kappa) = 0 \text{ for } i = 1, 2, \dots, N, \end{bmatrix}$$

where $\mu_{\Delta c}$ is the mean of Δc_t , μ_i the mean of R_{it}^e , σ_{ci} the covariance of Δc_t and R_{it}^e . 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

$$A\bar{g}(\theta) = \mathbf{0}_{(2N+3) \times 1}, \text{ where}$$

$$A_{(2N+3) \times (1+3N)} = \begin{bmatrix} 1 & \mathbf{0}_{1 \times N} & \mathbf{0}_{1 \times N} & \mathbf{0}_{1 \times N} \\ \mathbf{0}_{N \times 1} & I_N & \mathbf{0}_{N \times N} & \mathbf{0}_{N \times N} \\ \mathbf{0}_{N \times 1} & \mathbf{0}_{N \times N} & I_N & \mathbf{0}_{N \times N} \\ 0 & \mathbf{0}_{1 \times N} & \mathbf{0}_{1 \times N} & \sigma'_{ic} \\ 0 & \mathbf{0}_{1 \times N} & \mathbf{0}_{1 \times N} & \mathbf{1}_{1 \times N} \end{bmatrix},$$

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 $\mathbb{E} R_{it}^e$ on a constant and σ_{ci} . The test that the pricing errors are all zero is a Wald test that $\bar{g}(\theta)$ 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 (9.5) hold. The 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 with consumption growth.

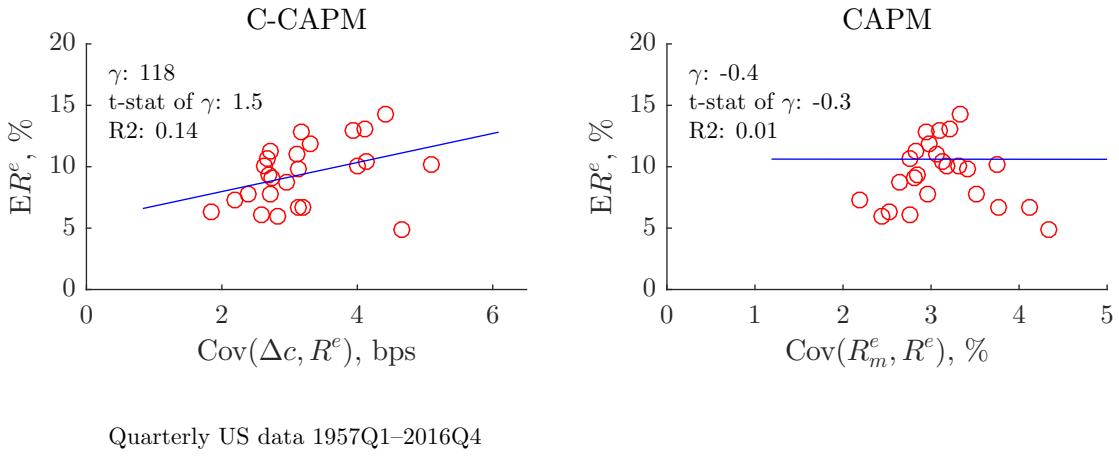


Figure 9.4: Test of C-CAPM and CAPM on 25 FF portfolios

Figure 9.4 shows the results of both C-CAPM and the standard CAPM—for the 25 Fama and French (1993) portfolios. It is clear that both models work badly, but CAPM actually worse.

Figure 9.5 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).

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.

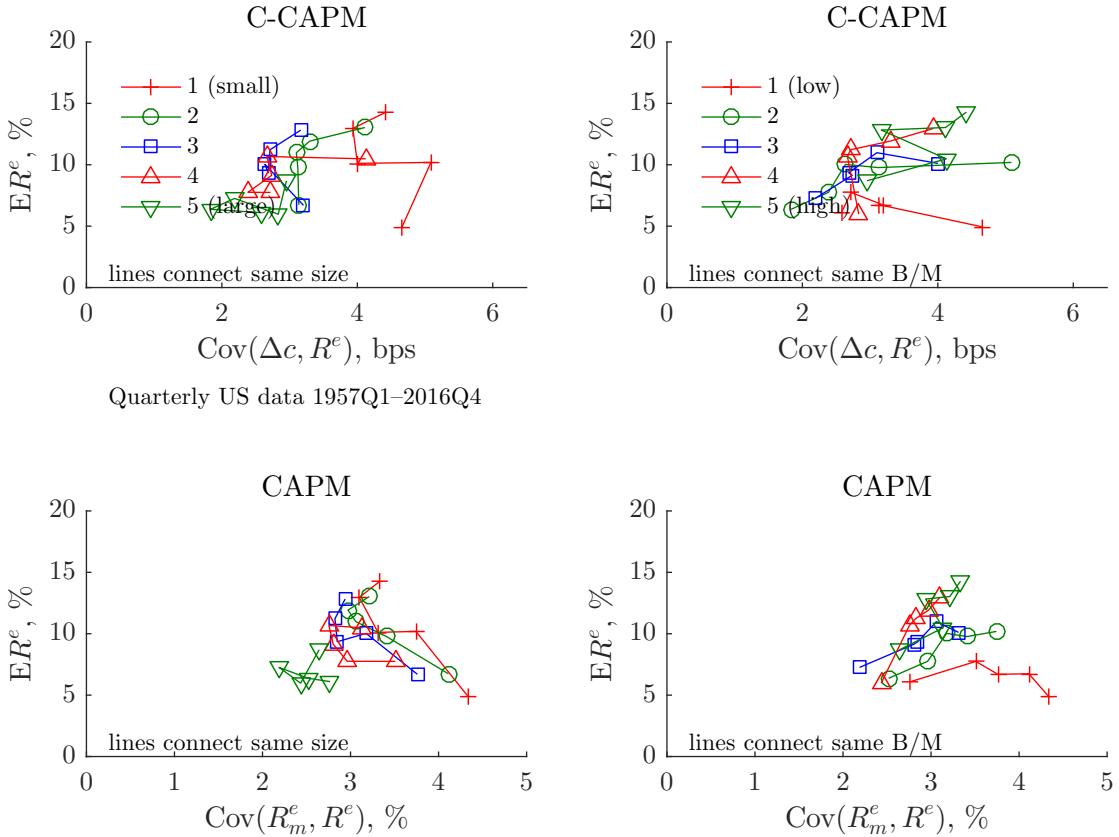


Figure 9.5: Diagnosing C-CAPM and CAPM, 25 FF portfolios

9.4 The Cross-Section of Returns: Conditional Models

The basic asset pricing model is about conditional moment and it can be summarized as in (9.4) which is given here again

$$E_{t-1} R_t^e = \text{Cov}_{t-1}(R_t^e, \Delta c_t) \gamma. \quad (\text{EPP3c again})$$

Expressing this in terms of unconditional moments as in (9.5) shows only part of the story.

However, it can be shown (see, for instance, 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 (9.4) holds, but with a time varying effective risk aversion coefficient (so γ should carry a time subscript).

9.4.1 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/total 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 an AR(1) process). Therefore, *cay* is modelled as

$$cay_t = c_t - \omega a_t - (1 - \omega)y_t, \quad (9.11)$$

where c_t is log consumption, a_t log financial wealth and y_t is log income. Log total wealth (financial wealth plus human capital) is approximately $\omega a_t + (1 - \omega)y_t$, where ω is the fraction of financial wealth in total wealth and $1 - \omega$ is the fraction of human capital. Information about ω would allow us to estimate what current total wealth is. Presumably, *cay* is a stationary variable, even if consumption, financial wealth and income are not. Equation (9.11) is therefore a cointegrating relation, so ω can be estimated by OLS (the point estimate is around 0.3).

Remark 9.5 (*The cay variable*) First, total wealth is $W = A + H$, where A is financial wealth and H is human capital. If the income process is a first-order autoregression, then human capital is a linear function of current income, $H = \kappa Y$. Let lower case letters denote logarithms. Log wealth can then be written $\ln(e^a + \kappa e^y)$. A first-order Taylor approximation is

$$\Delta w_t \approx \omega \Delta a_t + (1 - \omega) \Delta y_t,$$

where ω is the ratio of financial wealth to total wealth at the point around which we approximate. Second, collect all the constants and write the log consumption/wealth ratio as (9.11) plus a constant.

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

$$m_t = a_t + b_t \Delta c_t, \text{ where} \quad (9.12)$$

$$a_t = \gamma_0 + \gamma_1 cay_{t-1} \text{ and } b_t = \eta_0 + \eta_1 cay_{t-1}. \quad (9.13)$$

This is a *conditional C-CAPM*. It is clearly the same as specifying a linear factor model

$$R_{it}^e = \alpha + \beta_{i1} cay_{t-1} + \beta_{i2} \Delta c_t + \beta_{i3} (\Delta c_t \times cay_{t-1}) + \varepsilon_{it}, \quad (9.14)$$

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

$$\mathbb{E} R_t^e = \beta \lambda, \quad (9.15)$$

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 (9.14)–(9.15) 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 provide better fits of the cross-sectional returns than the unconditional models (including the Fama-French model)—and that the C-CAPM is actually a pretty good model.

9.4.2 Approach 2 of Testing the Conditional CCAPM: An Explicit Volatility Model

Reference: Duffee (2005)

Duffee (2005) estimates the conditional model (9.4) 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 (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

$$R_t^e = \alpha'_R Y_{R,t-1} + e_{R,t} \quad (9.16)$$

$$\Delta c_t = \alpha'_c Y_{c,t-1} + e_{c,t}. \quad (9.17)$$

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

$$R_t^e = b_0 + (b_1 + b_2 p_{t-1}) e_{R,t} e_{c,t} + w_t, \quad (9.18)$$

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 (9.4). 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 (9.16)–(9.18) 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

$$\begin{aligned} \widehat{\mathbb{E}_{t-1}} R_t^e &= \alpha'_R Y_{R,t-1} \text{ and} \\ \widehat{\text{Cov}_{t-1}}(e_{R,t}, e_{c,t}) &= \alpha'_v Z_{t-1} \end{aligned} \quad (9.19)$$

and then relating this proxies as

$$\widehat{\mathbb{E}_{t-1}} R_t^e = b_0 + (b_1 + b_2 p_{t-1}) \widehat{\text{Cov}_{t-1}}(e_{R,t}, e_{c,t}) + u_t. \quad (9.20)$$

The point of using a (GMM) system is that this allows handling the estimation uncertainty 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 (9.4)). This seems related to the results in Figure 9.3. 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).

9.5 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

$$\mathbb{E}_{t-1} R_t^e (C_t/C_{t-1})^{-\gamma} = 0 \quad (9.21)$$

Similarly, an n -period *real* bond price ($P_{n,t}$) satisfies

$$\mathbb{E}_t \delta^n (C_{t+n}/C_t)^{-\gamma} = P_{nt}, \text{ so} \quad (9.22)$$

$$C_t^{-\gamma} = \mathbb{E}_t \delta^n C_{t+n}^{-\gamma} / P_{n,t}. \quad (9.23)$$

Use in (9.21) to get

$$\begin{aligned} \mathbb{E}_{t-1} R_t^e m_{n,t} &= 0, \text{ where} \\ m_{n,t} &= (1/P_{n,t})(C_{t+n}/C_{t-1})^{-\gamma}. \end{aligned} \quad (9.24)$$

This expression relates the one-period excess return to an n -period SDF—which involves the real interest rate ($1/P_{n,t}$) and ratio of marginal utilities n periods apart.

If we 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

$$\begin{aligned} \mathbb{E}_{t-1} R_t^e &= -\text{Cov}_{t-1}(R_t^e, \ln m_{n,t}) \\ &= \text{Cov}_{t-1}[R_t^e, \gamma \ln(C_{t+n}/C_{t-1})] - \text{Cov}_{t-1}[R_t^e, y_{n,t}]. \end{aligned} \quad (9.25)$$

This first term is very similar to the traditional expression (9.2), 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 (9.2). Clearly, the moments in (9.25) 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. See Figure 9.6 for an illustration.

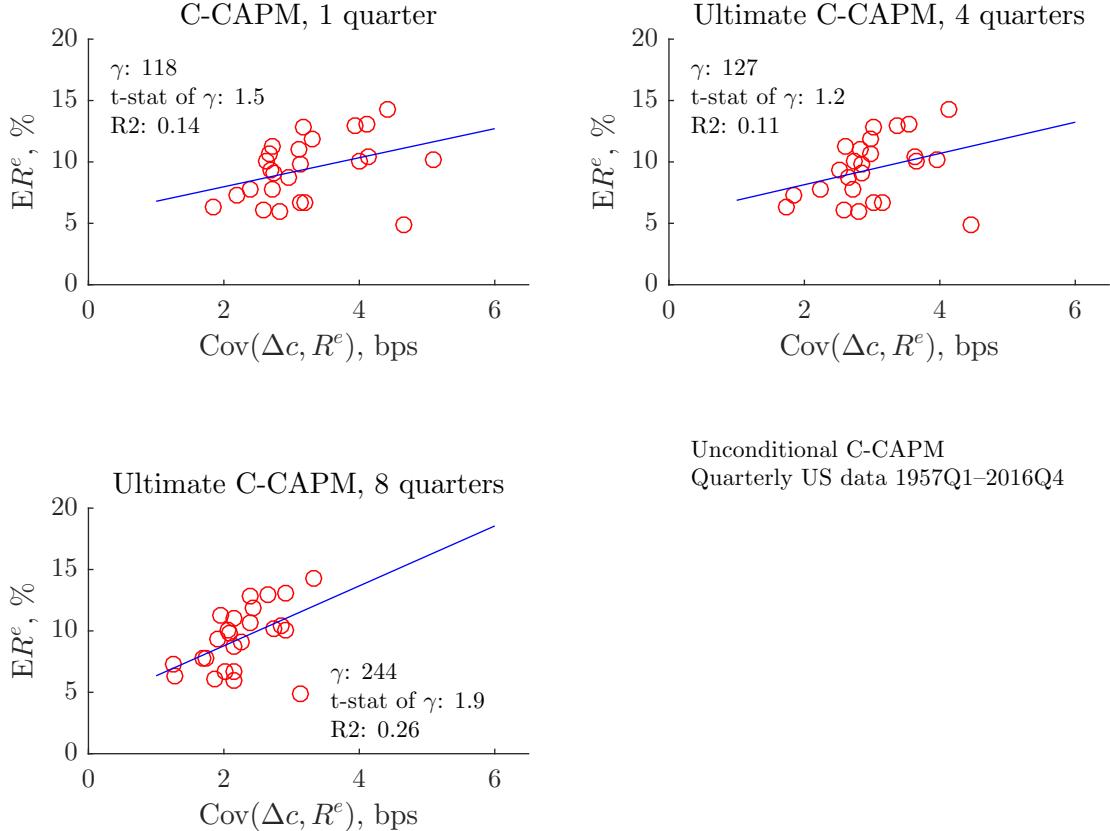


Figure 9.6: **C-CAPM and ultimate consumption, 25 FF portfolio.**

Proof. (of (9.22)–(9.24)) To prove (9.22), let $m_{t+1} = \delta(C_{t+1}/C_t)^{-\gamma}$ denote the SDF and P_{nt} the price of an n -period bond. Clearly, $P_{2t} = E_t m_{t+1} P_{1,t+1}$, so $P_{2t} = E_t m_{t+1} E_{t+1}(m_{t+2} P_{0,t+2})$. Use the law of iterated expectations (LIE) and $P_{0,t+2} = 1$ to get $P_{2t} = E_t m_{t+2} m_{t+1}$. The extension from 2 to n is straightforward, which gives (9.22). To prove (9.24), use (9.23) in (9.21), apply LIE and simplify. ■

9.6 Long Run Risk

Reference: Bansal and Yaron (2004)

The long run risk models use a combination of recursive (non-expected) utility from (Epstein and Zin (1989), Epstein and Zin (1991)) with predictable consumption growth to generate higher risk premia than in standard consumption-based models. In an extension they also allow for time-varying uncertainty to capture changes in the risk premia.

9.6.1 Epstein-Zin Utility

The basic idea of the recursive utility function in Epstein and Zin (1989) is to form a certainty equivalent of future utility as $Z_t = [E_t(U_{t+1}^{1-\gamma})]^{1/(1-\gamma)}$ 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 and δ is the time preference rate.

Epstein and Zin (1991) show that if all wealth is marketable, so the budget restriction can be written $W_t = R_{at}(W_{t-1} - C_{t-1})$ where R_{at} is the return on the wealth portfolio (typically not observed), then the Euler equation for an asset with return R_t is

$$E_{t-1} m_t R_t = 1, \text{ where} \quad (9.26)$$

$$m_t = \delta^\theta (C_t/C_{t-1})^{-\theta/\psi} R_{at}^{\theta-1}, \text{ and} \quad (9.27)$$

$$\theta = (1 - \gamma)/(1 - 1/\psi). \quad (9.28)$$

When $\theta = 1$, then m_t becomes a traditional CRRA model. The values of θ are illustrated in Figure 9.7.

To see that the intertemporal substitution is governed by the parameter ψ , close down all systematic risk by assuming that the market return is a riskfree rate (set $R_{at} = R_t = R_{ft}$ in $E_{t-1} M_t R_t = 1$) and that log consumption has a normal distribution. We then get that the log riskfree rate (r_{ft}) is

$$r_{ft} = -\ln \delta + \frac{1}{\psi} E_{t-1}(\Delta c_t) - \frac{\theta}{\psi^2} \text{Var}_{t-1}(\Delta c_t)/2. \quad (9.29)$$

This is the same relation between the real interest rate and consumption growth as in the CRRA case—except that the elasticity of intertemporal substitution is no longer forced to be the inverse of the risk aversion. This could help us to solve the riskfree rate puzzle.

Proof. (of (9.29)) Set $R_{at} = R_t = R_{ft}$ in ((9.26)–(9.28)) to get $E_{t-1} \delta^\theta (C_t/C_{t-1})^{-\theta/\psi} R_{ft}^\theta = 1$. Rewrite as $\delta^\theta \exp(\theta r_{ft}) E_{t-1} \exp(-\theta/\psi \times \Delta c_t) = 1$. If Δc is normally distributed, then the expectation equals $\exp[-\theta/\psi \times E_{t-1} \Delta c_t + \theta^2/\psi^2 \times \text{Var}_{t-1}(\Delta c_t)/2]$. Take logs and simplify. ■

To illustrate that (9.26)–(9.28) *could* have the same implications for risky assets as the CRRA model, assume that the consumption-wealth ratio is constant (for instance, because the market return is iid). This effectively closes down all interaction between the risk and intertemporal substitution (see Svensson (1989) and Campbell (1993)). It then

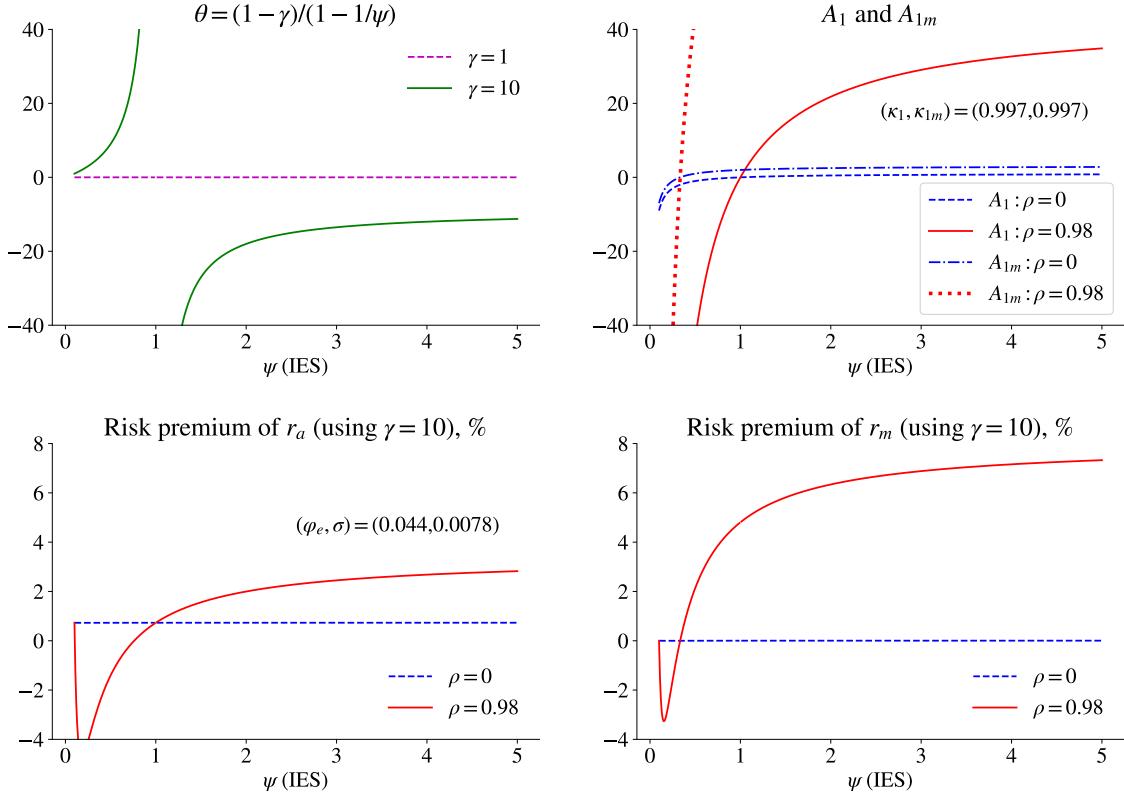


Figure 9.7: Illustration of the Bansal-Yaron (2004) model

follows from the budget constraint that consumption growth and the market return are proportional so (9.26)–(9.28) give the expected excess return as

$$E_{t-1}[(C_t/C_{t-1})^{-\gamma} R_t^e] = 0, \quad (9.30)$$

which is the same as in the CRRA model. In this case, the Epstein-Zin model inherits the problem with explaining the cross-section of returns. This is also approximately true when the consumption-wealth ratio is not too variable. Of course, if there are large predictable movements in the wealth return, then it is no longer true.

Proof. (of (9.30)) Let $C_t/W_t = 1/\alpha_t$, and substitute for wealth in the budget restriction to get $C_t\alpha_t = R_{at}(\alpha_{t-1}C_{t-1} - C_{t-1})$, or $C_t/C_{t-1} = R_{at}(\alpha_{t-1} - 1)/\alpha_t$ which in turn gives $(C_t/C_{t-1})\alpha_t/(\alpha_{t-1} - 1) = R_{at}$. Using this in (9.26) gives $E_{t-1}[(C_t/C_{t-1})^{-\gamma}\alpha_t^{\theta-1}R_{it}^e] = 0$. [Campbell \(1993\)](#) shows that there are no innovations in α_t if $\psi = 1$. Moreover, α is a constant if the wealth returns are iid or if $\psi = 1$ and all innovations are homoskedastic. In either case, $\alpha_t^{\theta-1}$ can be eliminated from the expression. ■

9.6.2 Long-Run Predictability

Bansal and Yaron (2004) (henceforth BY) use a model where consumption growth contains a predictable long-run component (x_t) which follows an AR(1)

$$x_t = \rho x_{t-1} + \varphi_e \sigma e_t. \quad (9.31)$$

Consumption growth (BY uses g_t to denote $\ln \Delta c_t$) follows the exogenous process

$$\Delta c_t = \mu + x_{t-1} + \sigma \eta_t. \quad (9.32)$$

All shocks are assumed to be iid normally distributed and uncorrelated with each other.

Remark 9.6 (ARMA(1,1) model*) Equations (9.31)-(9.32) imply a non-standard time series model for Δc_t . In contrast, if $e_t = \eta_t$ (which would lead to other expressions for the risk premia than the ones presented below), then Δc_t would be an ARMA(1,1) as in Bansal and Lundblad (2002). To see that, notice that (9.32) gives $x_{t-1} = \Delta c_t - \mu - \sigma \eta_t$. Use this in (9.31) to get $\Delta c_{t+1} - \mu = \rho(\Delta c_t - \mu) - \rho \sigma \eta_t + \varphi_e \sigma e_t + \sigma \eta_{t+1}$. With $e_t = \eta_t$, the error terms simplify to $(\varphi_e \sigma - \rho \sigma) \eta_t + \sigma \eta_{t+1}$, so we have an ARMA(1,1).

Notice that the one-period forecasts of Δc_{t+s} are driven by x_t only, making x_t a natural “state variable” of the model. Also notice that if x_t is a constant (because $\varphi_e = 0$), then there is no predictability (assuming x_t starts at its long-run mean of 0).

The properties of this consumption growth process are illustrated in Figure 9.8, using the calibration of BY. The key points are that (a) shocks to the state variable x_t are small, but they accumulate in case the process is persistent (ρ is high); (b) iid shocks to consumption growth are large (but have no persistence); (c) the implied autocorrelations of consumption are small, but they decline very slowly. Together this means that autocorrelations of annual consumption growth (as reported by BY) are high (0.25 or higher).

9.6.3 The Risk Premium I

If all variables are lognormally distributed, then we can use the standard result that the risk premium is

$$E_{t-1} r_t - r_{ft} + \text{Var}_{t-1}(r_t)/2 = -\text{Cov}_{t-1}(r_t, \ln m_t). \quad (9.33)$$

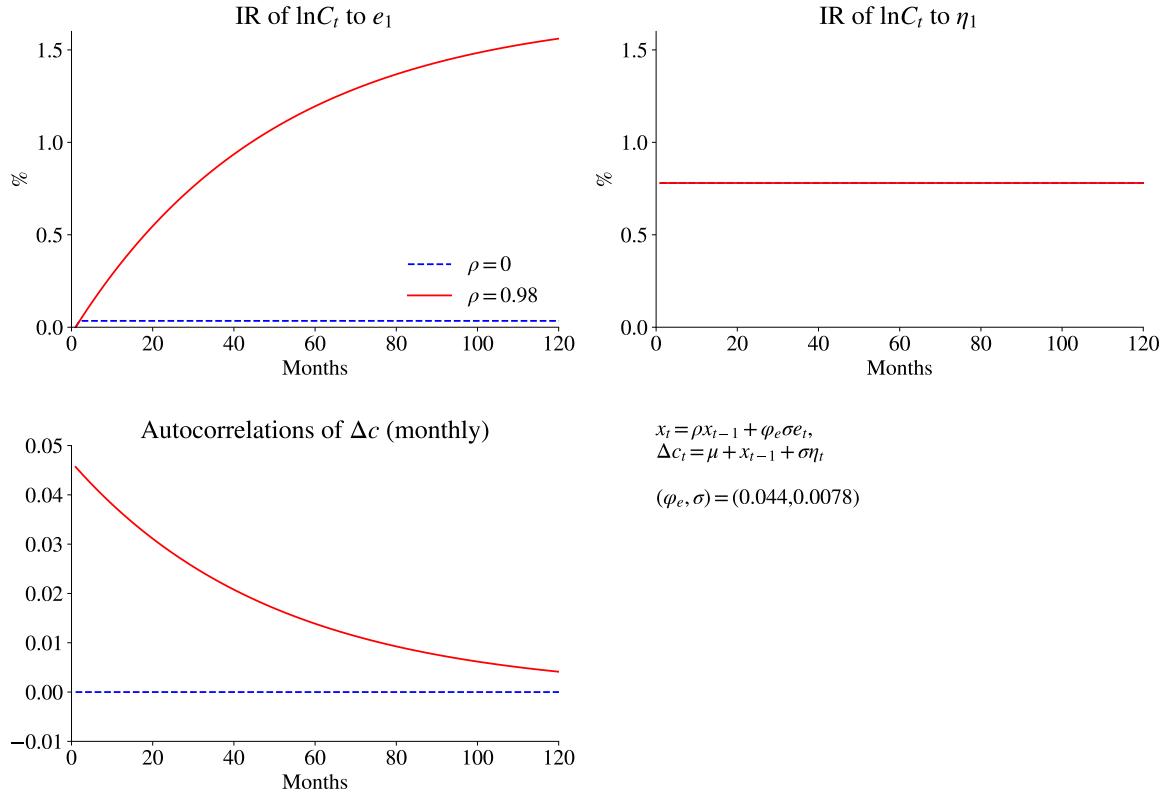


Figure 9.8: The consumption process in the Bansal-Yaron (2004) model

(The variance term on the right hand side is just a Jensen's inequality term—and would disappear if we instead considered $E_{t-1} R_t / R_{ft}$.) Similarly, the log riskfree rate is

$$r_{ft} = -E_{t-1} \ln m_t - \text{Var}_{t-1}(\ln m_t)/2. \quad (9.34)$$

(This riskfree rate is for an investment between $t - 1$ and t , but it is known already in $t - 1$.)

Proof. (of (9.33)). Write $E m R = 1$ as $E \exp(\ln m + r) = 1$. If $\ln m + r$ is normally distributed, then the expectation can be written $\exp[E \ln m + E r + \text{Var}(\ln m + r)/2]$. Take logs and rearrange to get $E r + \text{Var}(r)/2 = -E \ln m - \text{Var}(\ln m)/2 - \text{Cov}(r, \ln m)$. For a riskfree asset this simplifies to $r_f = -E \ln m - \text{Var}(\ln m)/2$. The difference of the two expressions gives the risk premium as (9.33). ■

Taking logs of the pricing kernel in (9.27) gives

$$\ln m_t = \theta \ln \delta - \theta/\psi \times \Delta c_t + (\theta - 1)r_{at}, \quad (9.35)$$

where $r_{at} = \ln R_{at}$. (Warning: BY uses m_t to denote the *log* pricing kernel, whereas these notes use $\ln m_t$.)

Combining (9.33) and (9.35) gives the risk premium on any asset as

$$E_{t-1} r_t - r_{ft} + \text{Var}_{t-1}(r_t)/2 = \theta/\psi \times \text{Cov}_{t-1}(r_t, \Delta c_t) + (1-\theta) \text{Cov}_{t-1}(r_t, r_{at}). \quad (9.36)$$

In the basic case with constant uncertainty (the focus of these notes), the risk premium is constant across time (since the variances and covariances are). Also, when $\theta = 1$ (that is, when $\gamma = 1/\psi$), then (9.36) simplifies to the CRRA case.

Combining (9.34) and (9.35) shows that the riskfree rate is

$$r_{ft} = -\theta \ln \delta + \theta/\psi \times E_{t-1} \Delta c_t + (1-\theta) E_{t-1} r_{at} - \text{Var}_{t-1}[-\theta/\psi \times \Delta c_t + (\theta-1)r_{at}]/2. \quad (9.37)$$

9.6.4 Simplified Expressions for the Risk Premia and Riskfree Rate

Since all variances and covariances are constant, we can write the risk premium (9.36) as

$$E_{t-1} r_t - r_{ft} = \pi, \quad (9.38)$$

where π is a constant which depends on the (constant) variances and covariances. This applies also to the wealth asset so equation (9.37) can be simplified (using $E_{t-1} r_{at} = r_{ft} + \pi_a$) as

$$r_{ft} = \pi_f + 1/\psi \times E_{t-1} \Delta c_t. \quad (9.39)$$

9.6.5 Solving the Model I

The “wealth” asset (with log return r_{at}) is not directly observable, but we could back it out from the model, since its dividend equals (aggregate) consumption.

The **Campbell and Shiller (1988)** approximation (see also the lecture notes on predictability) gives the log return of the wealth asset as

$$r_{at} \approx \kappa_0 + \kappa_1 (p_t - c_t) - (p_{t-1} - c_{t-1}) + \Delta c_t, \quad (9.40)$$

where κ_0 is a constant and κ_1 depends on the average dividend-price (that is, consumption-price) ratio and is calibrated to be around 0.997 for monthly data. (Strictly speaking, κ_1 should be endogenous since the price level is, but we disregard this here.) Notice that c_t shows up in this equation since the dividend of this asset equals aggregate consumption.

Since x_t is the only natural state variable of the model, it is reasonable to conjecture that the price-consumption ratio has the solution

$$p_t - c_t = A_0 + A_1 x_t, \quad (9.41)$$

where A_0 and A_1 are coefficients that we need to find.

9.6.6 Solving the Model II

Use the proposed solution for $p_t - c_t$ (9.41) in the approximation (9.40) of r_{at} . Then take expectations, using $E_{t-1} x_t = \rho x_{t-1}$ from (9.31), to get

$$E_{t-1} r_{at} = \pi_a + (\kappa_1 A_1 \rho - A_1 + 1)x_{t-1}. \quad (9.42)$$

Now, substitute for $E_{t-1} \Delta c_t$ in the riskfree rate (9.39) by using (9.32) to get

$$r_{ft} = \tilde{\pi}_f + 1/\psi \times x_{t-1}. \quad (9.43)$$

Finally, combine (9.42) and (9.43) to express the risk premium as

$$E_{t-1} r_{at} - r_{ft} = \tilde{\pi}_a + (\kappa_1 A_1 \rho - A_1 + 1 - 1/\psi)x_{t-1}. \quad (9.44)$$

This risk premium should be constant (see equation (9.38)) which requires that the term multiplying x_{t-1} is zero, that is,

$$\begin{aligned} \kappa_1 A_1 \rho - A_1 + 1 - 1/\psi &= 0, \text{ so} \\ A_1 &= \frac{1 - 1/\psi}{1 - \kappa_1 \rho}. \end{aligned} \quad (9.45)$$

This is the key parameter of the model. The A_1 values are illustrated in Figure 9.7.

Notice that $\psi = 1$ makes $p_t - d_t$ a constant (compare with the results in (9.30)). Also notice that high persistence (ρ close to one) makes $p_t - c_t$ very sensitive to movements in x_t (and positively so if the intertemporal elasticity of substitution ψ is larger than 1).

9.6.7 The Risk Premium II

Notice from (9.32) that the surprise in consumption growth is

$$\Delta c_t - E_{t-1} \Delta c_t = \sigma \eta_t. \quad (9.46)$$

Then, notice from (9.40) than the surprise in the r_{at} depends on the surprises in $p_t - c_t$ and in Δc_t . Therefore, use (9.41) to substitute for $p_t - c_t$ and (9.31)–(9.32) to express the result in terms of the shocks

$$r_{at} - E_{t-1} r_{at} = \kappa_1 A_1 \varphi_e \sigma e_t + \sigma \eta_t. \quad (9.47)$$

It follows directly that

$$\begin{aligned} \text{Var}_{t-1}(r_{at}) &= \kappa_1^2 A_1^2 \varphi_e^2 \sigma^2 + \sigma^2 \\ \text{Var}_{t-1}(\Delta c_t) &= \sigma^2 \\ \text{Cov}_{t-1}(r_{at}, \Delta c_t) &= \sigma^2. \end{aligned} \quad (9.48)$$

Applying (9.36) to $r_t = r_{at}$ then gives the risk premium

$$E_{t-1} r_{at} - r_{ft} + \text{Var}_{t-1}(r_{at})/2 = \theta/\psi \times \sigma^2 + (1-\theta)(\kappa_1^2 A_1^2 \varphi_e^2 + 1)\sigma^2. \quad (9.49)$$

The risk premia are illustrated in Figure 9.7. It is clear that it takes a combination of a high risk aversion (γ), a high intertemporal elasticity of substitution (ψ) and a high autocorrelation in growth (ρ) to generate a high risk aversion.

Remark 9.7 (“Market return”) *BY also analyse the risk premium of a market return (r_{mt}). This equity has the dividend process $\Delta d_t = \mu_d + \phi x_{t-1} + \varphi_d \sigma u_t$, where $\phi = 3$ in the base case model (which represents that equity dividends are more sensitive to long-run growth than consumption). Using the same approach as before, the price dividend ratio is $p_{mt} - d_t = A_{0m} + A_{1m} x_t$, where $A_{1m} = (\phi - 1/\psi)/(1 - \kappa_{1m}\rho)$, where κ_{1m} is similar to κ_1 (compare with (9.45)). Similarly, the equity risk premium is $E_{t-1} r_{mt} - r_{ft} + \text{Var}_{t-1}(r_{mt})/2 = (1-\theta)\kappa_{1m} A_{1m} \kappa_1 A_1 \varphi_e^2 \sigma^2$ (compare with (9.48)). The risk premia are illustrated in Figure 9.7.*

9.6.8 Estimation

The perhaps most straightforward way to estimate the model is to use data on both the riskfree rate and consumption growth rate. Notice that (9.43) gives

$$x_{t-1} = \psi(r_{ft} - \bar{r}_{ft}), \quad (9.50)$$

where \bar{r}_{ft} is the sample average of r_{ft} . Assuming a value of ψ thus allows us to reconstruct the x_t variable. We can then estimate $(\rho, \varphi_e \sigma)$ by OLS of (9.31), and σ as the

standard deviation of $\Delta c_t - x_{t-1}$ (see (9.32)). Combining these results gives φ_e .

Figure 9.9 shows results from a Monte Carlo simulation of this approach (using the same parameters as in the other figures). The left subfigure is for a “medium-sized” sample (420 months, that is, 35 years) while the right subfigure is for a very long sample.

Two results stand out. First, the estimated values (of ρ, φ_e, σ) imply a downward bias of the market risk premium. This is mostly driven by the usual downward bias of the autocorrelation coefficient (ρ). Second, the dispersion is considerable. A zero risk premium is just slightly outside the simulated 90% confidence band.

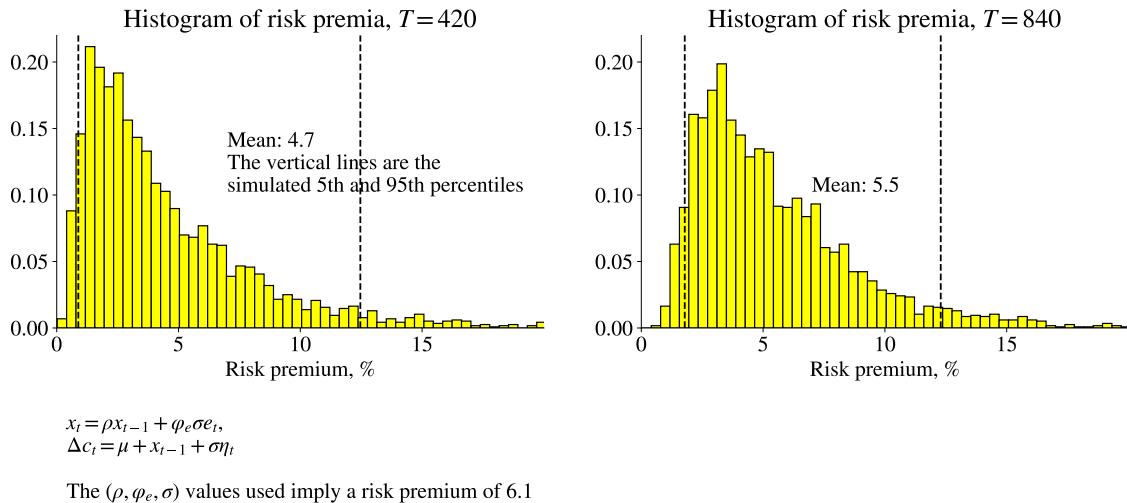


Figure 9.9: Monte Carlo simulation of the Bansal-Yaron model.

Chapter 10

Financial Panel Data

References: Verbeek (2012) 10, Baltagi (2008), Hoechle (2007), Driscoll and Kraay (1998), Wooldridge (2010) and Petersen (2009).

10.1 Introduction to Panel Data

A panel data set (also called a longitudinal data set) has data on a cross-section ($i = 1, 2, \dots, N$, individuals or firms) for many time periods ($t = 1, 2, \dots, T$). Our aim is to estimate a linear relation between the dependent variable and the regressors

$$y_{it} = \alpha_i + x'_{it}\beta_i + u_{it}, \quad (10.1)$$

where the coefficients (α_i, β_i) may or may not be different for different individuals (this is discussed in detail below). As examples of such applications, we may want to evaluate if alphas or betas of different mutual funds are related to fund characteristics, for instance, costs or trading activity. Alternatively, we want to investigate whether firms with different types of board compositions perform differently. Sometimes it will be convenient to put the constant in the constant in the x_{it} vector to write the model as $y_{it} = x'_{it}\beta_i + u_{it}$. (This should be clear from the context.)

Data on the dependent variable has this structure:

$$\begin{array}{cccc} & \underline{i=1} & \underline{i=2} & \cdots & \underline{i=N} \\ t=1: & y_{11} & y_{21} & & y_{N1} \\ t=2: & y_{12} & y_{22} & & y_{N2} \\ & \vdots & & & \\ t=T: & y_{1T} & y_{2T} & & y_{NT} \end{array} \quad (10.2)$$

The structure for each of the regressors is similar, although it can also be the case that (some of) the regressors are the same for all N investors (for instance, when the regressors are pricing factors like the market excess return). When needed for clarity we will use the $y_{i,t}$ notation instead of y_{it} .

The structure in (10.2) implicitly assumes that we have a *balanced panel*, that is, have data for all the cells. However, it is often the case that the panel is *unbalanced* in the sense that some data is missing. For instance, we may not have data on regressor 3 for $i = 7$ and $t = 3$. If data is *missing in a random way*, then we can simply exclude (y_{it}, x_{it}) for the missing (i, t) . In our example that means just excluding $(y_{7,3}, x_{7,3})$ but keeping all other data. In contrast, if data is missing in a non-random way (for instance, depending on the value of y_{it}), then we have to apply more sophisticated sample-selection models (not discussed in this chapter).

10.2 An Overview of Different Panel Data Models

A *pooled model* assumes that all individuals have the same coefficients (no subscript on β), so (10.1) becomes

$$y_{it} = \alpha + x'_{it}\beta + u_{it}. \quad (10.3)$$

This model can be estimated by pooled OLS (see below).

A *fixed effects model* assumes that all individuals have the same slope coefficients, but that their intercepts might differ

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}. \quad (10.4)$$

An extension of the fixed effects model is to also allow for *time fixed effects*

$$y_{it} = \lambda_t + \alpha_i + x'_{it}\beta + u_{it}. \quad (10.5)$$

Estimation of these models is discussed below.

A *random effects model* is similar to a fixed effects model, except that the individual “mean” α_i now contains a common component (α) and a random individual component (μ_i). We can then write the model as

$$y_{it} = \alpha + x'_{it}\beta + u_{it} \text{ where } u_{it} = \mu_i + \varepsilon_{it}. \quad (10.6)$$

The ε_{it} is typically assumed to be uncorrelated across time and individuals, but the μ_i

terms make the u_{it} residuals correlated over time (for the same individual). The estimation of this model is discussed later.

The *unrestricted model* (10.1) allows all individuals to have different coefficients (hence a subscript i on β_i). These regressions could be estimated by OLS for each individual separately. Alternatively, a GLS approach can be applied to enhance the efficiency by exploiting the correlation (of the residuals) across individuals. This approach is not discussed in these notes, since it is basically very similar to the SURE model used for testing CAPM and other linear factor models. (See the CAPM notes.)

10.3 Pooled OLS

Consider the regression model

$$y_{it} = x'_{it}\beta + u_{it}, \quad (10.7)$$

where x_{it} is an $k \times 1$ vector. For notational convenience, this section assumes that any constant is included in the x_{it} vector along with the other regressors. Notice that the coefficients are the same across individuals (and time), but that the regressors may vary along both the time series and cross-sectional dimensions. We assume that u_{jt} is uncorrelated with x_{it} (across all i and j).

Define the matrices

$$\Sigma_{xx} = \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N x_{it}x'_{it} \text{ (a } k \times k \text{ matrix)} \quad (10.8)$$

$$\Sigma_{xy} = \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N x_{it}y_{it} \text{ (a } k \times 1 \text{ vector).} \quad (10.9)$$

The LS estimator (stacking all TN observations) is then

$$\hat{\beta} = \Sigma_{xx}^{-1} \Sigma_{xy}. \quad (10.10)$$

In case u_{it} is uncorrelated across time and also across individuals, then the usual expressions for $\text{Std}(\hat{\beta})$ apply. However, it is often the case that there are *clusters* of individuals (all small firms, say) that have correlated residuals. This would require handling those correlations.

Recall that we can (conceptually) decompose the point estimate $\hat{\beta}$ by using (10.7) to substitute for y_{it} in Σ_{xy} (10.9) and then in (10.10). The result is

$$\hat{\beta} = \beta + \Sigma_{xx}^{-1} \left(\frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N h_{it} \right), \text{ where } h_{it} = x_{it}u_{it}. \quad (10.11)$$

The variance-covariance matrix ($k \times k$) can then be written

$$\text{Var}(\sqrt{TN}\hat{\beta}) = \Sigma_{xx}^{-1} \text{Var} \left(\frac{1}{\sqrt{TN}} \sum_{t=1}^T \sum_{i=1}^N h_{it} \right) \Sigma_{xx}^{-1}. \quad (10.12)$$

(Clearly, if $\text{Var}(\sqrt{TN}\hat{\beta}) = A$, then $\text{Var}(\hat{\beta}) = A/(TN)$.) Notice that the middle matrix in (10.12) is the variance-covariance matrix of a sum of the $k \times 1$ vector $x_{it}u_{it}$ (divided by \sqrt{TN}). This sum looks like

$$\sum_{t=1}^T \sum_{i=1}^N h_{it} = \underbrace{h_{1,1}}_{i=1, t=1} + \underbrace{h_{2,1}}_{i=2, t=1} + \dots + \underbrace{h_{N-1,T}}_{i=N-1, t=T} + \underbrace{h_{N,T}}_{i=N, t=T}. \quad (10.13)$$

The variance of this sum depends on how the elements are correlated. Different *cluster methods* would account for a non-zero covariance across individuals within the same period (for instance, between h_{it} and h_{jt}), or across time for the same individual (for instance, between $h_{i,t}$ and $h_{i,t-1}$) and sometimes for both.

Remark 10.1 (*Panel regression vs average coefficient) Consider the regression for investor i

$$y_{it} = x_t' \beta_i + \varepsilon_{it}, \quad i = 1 \dots N,$$

where the regressors are the same in all regressions—but where the coefficients might be different across investors. Clearly, we have for each i

$$\hat{\beta}_i = \tilde{S}_{xx}^{-1} \tilde{S}_{xy_i},$$

where $\tilde{S}_{xx} = \sum_{t=1}^T x_t x_t' / T$ and $\tilde{S}_{xy_i} = \sum_{t=1}^T x_t y_{it} / T$.

The cross-sectional average of the regression coefficients is therefore

$$\frac{1}{N} \sum_{i=1}^N \hat{\beta}_i = \tilde{S}_{xx}^{-1} \frac{1}{N} \sum_{i=1}^N \tilde{S}_{xy_i}.$$

Compare that to (10.8) and notice that since x_t is repeated N times, we have $\tilde{S}_{xx} = \Sigma_{xx}$. Similarly, comparing with (10.9) gives

$$\frac{1}{N} \sum_{i=1}^N \tilde{S}_{xy_i} = \Sigma_{xy}.$$

This shows that $\frac{1}{N} \sum_{i=1}^N \hat{\beta}_i = \hat{\beta}$, where the latter is from the panel regression (10.10).

10.4 The Within Estimator (“Fixed Effects Estimator”)

In the fixed effects model, we allow for different individual intercepts

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}. \quad (10.14)$$

There are several ways to estimate this model. The conceptually most straightforward is to include individual dummies (N) where dummy i takes the value of one if the data refers to individual i and zero otherwise and estimate the model with pooled OLS. (Clearly, the regression can then not include any intercept. Alternatively, include an intercept but only $N - 1$ dummies, for $i = 2 - N$.) However, this approach can be difficult to implement since it may involve a very large number of regressors.

As an alternative (which gives the same point estimates as pooled OLS with dummies) consider the following approach. First, take average across time (for a given i) of y_{it} and x_{it} in (10.14). That is, think (but do not run any estimation yet...) of forming the cross-sectional regression

$$\bar{y}_i = \alpha_i + \bar{x}'_i\beta + \bar{u}_i, \text{ where} \quad (10.15)$$

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it} \text{ and } \bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}. \quad (10.16)$$

Second, transform the data as

$$y_{it}^* = y_{it} - \bar{y}_i \quad (10.17)$$

$$x_{it}^* = x_{it} - \bar{x}_i. \quad (10.18)$$

Use this to express the difference between (10.14) and (10.15) as

$$y_{it}^* = x'_{it}\beta + u_{it}^*. \quad (10.19)$$

At this stage, estimate β by running pooled OLS on all observations of (10.19). There is no intercept in this regression. We denote this estimate $\hat{\beta}_{FE}$ (FE stands for fixed effects) and it is also often called the *within estimator*. The interpretation of this approach is that we estimate the slope coefficients by using only the movements around individual means (not how the individual means differ). Notice that it gives the same results as OLS with dummies. Third and finally, get estimates of individual intercepts as

$$\alpha_i = \bar{y}_i - \bar{x}'_i \hat{\beta}_{FE}. \quad (10.20)$$

Clearly, the within estimator wipes out all regressors that are constant across time for a given individual (say, gender and schooling): they are effectively merged with the individual means (μ_i). In practice, such variables must be excluded from the x_{it} vector since otherwise there will be some transformed variables, $x_{it} - \bar{x}_i$, that are always zero—causing numerical problems.

We can apply the usual tests on the pooled OLS results from (10.19)—provided the residuals are uncorrelated across time and individuals. Otherwise, we need to apply a cluster method.

Remark 10.2 (*Lagged dependent variable as regressor**) *If $y_{i,t-1}$ is among the regressors x_{it} , then the within estimator (10.19) is biased in short samples (that is, when T is small)—and increasing the cross-section (that is, N) does not help. To see the problem, suppose that the lagged dependent variable is the only regressor ($x_{it} = y_{i,t-1}$). The within estimator (10.19) is then*

$$y_{it} - \sum_{t=1}^T y_{it}/T = [y_{i,t-1} - \sum_{t=2}^T y_{i,t-1}/(T-1)]\beta + [u_{it} - \sum_{t=1}^T u_{it}/T].$$

The problem is that the regressor ($y_{i,t-1} - \dots$) is correlated with Σu_{it} since the latter contains $u_{i,t-1}$ which affects $y_{i,t-1}$ directly. In addition, $\Sigma y_{i,t-1}$ contains $y_{i,t}$ which is correlated with u_{it} . It can be shown that this bias can be substantial for panels with small T .

10.4.1 The Within Estimator with Time Fixed Effects

When we allow for both time fixed effects and individual fixed effects

$$y_{it} = \lambda_t + \alpha_i + x'_{it}\beta + u_{it}, \quad (10.21)$$

then we could once again introduce dummies (now for both time periods and individuals) and apply pooled OLS.

As before, it is often easier to transform the data before estimating with pooled OLS. In this case, we run the regression on transformed variables

$$y_{it}^* = x_{it}^{*\prime}\beta + u_{it}^*. \quad (10.22)$$

	LS	Fixed eff	Between	GLS
exper/100	7.84 (8.25)	4.11 (6.21)	10.64 (4.05)	4.57 (7.12)
exper2/100	-0.20 (-5.04)	-0.04 (-1.50)	-0.32 (-2.83)	-0.06 (-2.37)
tenure/100	1.21 (2.47)	1.39 (4.25)	1.25 (0.90)	1.38 (4.32)
tenure2/100	-0.02 (-0.85)	-0.09 (-4.36)	-0.02 (-0.20)	-0.07 (-3.77)
south	-0.20 (-13.51)	-0.02 (-0.45)	-0.20 (-6.67)	-0.13 (-5.70)
union	0.11 (6.72)	0.06 (4.47)	0.12 (3.09)	0.07 (5.57)

Table 10.1: Panel estimation of log wages for women, $T = 5$ and $N = 716$, from NLS (1982,1983,1985,1987,1988). Example of fixed and random effects models, Hill et al (2008), Table 15.9. Numbers in parentheses are t-stats.

The transformations are

$$y_{it}^* = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} \quad (10.23)$$

$$x_{it}^* = x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}, \quad (10.24)$$

where \bar{x}_i is defined in (10.16) and

$$\begin{aligned} \bar{x}_t &= \sum_{i=1}^N x_{it}/N \text{ and} \\ \bar{x} &= \sum_{t=1}^T \sum_{i=1}^N x_{it}/(TN). \end{aligned}$$

(Similarly for the transformation of y_{it} .) The last terms (\bar{y}, \bar{x}) makes sure that the grand mean of the transformed variable is zero. (If we instead add an intercept to (10.22), then this is not important for the slope coefficients.)

The estimation and testing of (10.22) is the same as for the standard within estimator (see above).

10.5 The First-Difference Estimator

An another way of estimating the fixed effects model is to difference away the μ_i by taking *first-differences* (in time)

$$\Delta y_{it} = \Delta \lambda_t + \Delta x'_{it} \beta + u_{it}^*, \quad (10.25)$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and similarly for the regressors. Notice that

$$u_{it}^* = u_{it} - u_{i,t-1}, \quad (10.26)$$

so there are reasons to suspect that u_{it}^* is (negatively) autocorrelated.

Notice that the first-difference approach focuses on how changes in the regressors (over time, for the same individual) affect changes in the dependent variable. Also this method wipes out all regressors that are constant across time (for a given individual).

Regression (10.26) can be estimated by pooled OLS. However, unadjusted standard errors are likely to overstate the uncertainty. This suggests that using the unadjusted standard errors is a conservative approach (harder to reject the null hypothesis). The reason is that if u_{it} is iid, then $\text{Cov}(u_{it}^*, u_{i,t-1}^*) = -\text{Var}(u_{it})$.

Remark 10.3 (*Lagged dependent variable as regressor**) *If $y_{i,t-1}$ is among the regressors x_{it} , then the first-difference method (10.25) does not work (OLS is inconsistent and a larger sample does not help). The reason is that the (autocorrelated) residual is then correlated with the lagged dependent variable. This model cannot be estimated by OLS (the instrumental variable method might work).*

10.6 Differences-in-Differences Estimator

Consider the first-difference model (10.25) when one of the regressors is a dummy variable indicating whether individual i was “treated” (for instance, received investment advise) in period t . We can estimate this as before—and interpret the coefficient as the effect of the “treatment” (conditional on all other variables)

In the classical difference-in-difference estimator there are only two periods ($T = 2$): before and after the treatment. *If there are no other regressors*, then (10.25) can be written

$$\Delta y_{it} = \Delta \lambda_t + \beta Q_{it} + u_{it}^*, \quad (10.27)$$

where Q_{it} is the dummy variable. (The restriction that all individuals have the same $\Delta\lambda_t$ term is the so called “parallel trend assumption.”) In this case β can be estimated by the difference between the average Δy_{it} among the treated ($\Delta \bar{y}_{B2}$) and the average Δy_{it} among the non-treated ($\Delta \bar{y}_{A2}$)

$$\hat{\beta} = \Delta \bar{y}_{B2} - \Delta \bar{y}_{A2}. \quad (10.28)$$

(Notice that the change of the average is the same as the average of the change.)

10.7 Random Effects Model*

The random effects model allows for *random* individual “intercepts” (μ_i)

$$y_{it} = \beta_0 + x'_{it}\beta + \mu_i + \varepsilon_{it}, \text{ where} \quad (10.29)$$

$$\varepsilon_{it} \text{ is iid } N(0, \sigma_\varepsilon^2) \text{ and } \mu_i \text{ is iid } N(0, \sigma_\mu^2). \quad (10.30)$$

Notice that μ_i is random (across agents) but constant across time, while ε_{it} is just random noise. Hence, μ_i can be interpreted as the permanent “luck” of individual i .

It is sometimes argued that the random effect only makes sense if the data is a sample from a larger population—and then captures the peculiar (relative to the population) features of the individuals that end up in the sample. It is then convenient to merge μ_i with ε_{it} , because it gives fewer parameters to estimate (and thus, saves degrees of freedom). In contrast, if the cross-section effectively contains the population (all mutual funds on a market, say), then a fixed effect is perhaps more reasonable.

Clearly, if we regard μ_i as non-random, then we are back in the fixed-effects model. (The choice between the two models is not always easy, so it may be wise to try both—and compare the results.)

We could write the regression as

$$y_{it} = \beta_0 + x'_{it}\beta + u_{it}, \text{ where } u_{it} = \mu_i + \varepsilon_{it}. \quad (10.31)$$

and we typically assume that ε_{jt} and μ_i are not correlated with each other or with x_{it} . Notice that u_{it} is autocorrelated even if ε_{it} is not: $\text{Cov}(u_{it}, u_{i,t-1}) = \text{Var}(\mu_i)$.

There are several ways to estimate the random effects model. First, the methods for fixed effects (the within and first-difference estimators) all work—so the “fixed effect” can actually be a random effect. Second, the *between estimator* using only individual

time averages (from (10.16))

$$\bar{y}_i = \beta_0 + \bar{x}'_i \beta + \underbrace{\mu_i + \bar{\varepsilon}_i}_{\text{residual}_i}, \quad (10.32)$$

is also consistent, but discards all time-series information. Third, LS on

$$y_{it} = \beta_0 + x'_{it} \beta + \underbrace{\mu_i + \varepsilon_{it}}_{\text{residual}_{it}} \quad (10.33)$$

is consistent (but not really efficient). However, in this case we may need to adjust $\text{Cov}(\hat{\beta})$ since the covariance matrix of the residuals is not diagonal.

In the random effects model, the μ_i variable can be thought of as an *excluded variable*. Excluded variables typically give a bias in the coefficients of all included variables—unless the excluded variable is uncorrelated with all of them. This is the assumption in the random effects model (recall: we assumed that μ_i is uncorrelated with x_{jt}). If this assumption is wrong, then we cannot estimate the RE model by either OLS or GLS, but the within-estimator (compare with the FE model) works, since it effectively eliminates the excluded variable from the system.

Remark 10.4 (*Generalized least squares**) *GLS is an alternative estimation method that exploits correlation structure of residuals to increase the efficiency. In this case, it can be implemented by running OLS on*

$$y_{it} - \vartheta \bar{y}_i = \beta_0(1 - \vartheta) + (x_{it} - \vartheta \bar{x}_i)' \beta + v_{it}, \text{ where} \\ \vartheta = 1 - \sqrt{\sigma_u^2 / (\sigma_u^2 + T\sigma_\mu^2)}.$$

In this equation, σ_u^2 is the variance of the residuals in the “within regression” as estimated in (10.19) and $\sigma_\mu^2 = \sigma_B^2 - \sigma_u^2 / T$, where σ_B^2 is the variance of the residuals in the “between regression” (10.32). Here, σ_μ^2 can be interpreted as the variance of the random effect μ_i . However, watch out for negative values of σ_μ^2 and notice that when $\vartheta \approx 1$, then GLS is similar to the “within estimator” from (10.19). This happens when $\sigma_\mu^2 \gg \sigma_u^2$ or when T is large. The intuition is that when σ_μ^2 is large, then it is smart to get rid of that source of noise by using the within estimator, which disregards the information in the differences between individual means.

10.8 Fama-MacBeth

The Fama and MacBeth (1973) approach (called FMB below) is a different method for handling panel data. The method has two main steps, described below.

First, estimate λ_t and β_t

$$y_{it} = \lambda_t + x'_{it}\beta_t + u_{it} \quad (10.34)$$

period by period (using the cross section $i = 1 - N$). The FMB has the nice properties of easily handling unbalanced data sets (the cross-sectional regressions (10.34) are run on the available cross section for each time period).

Second, estimate the time averages

$$\hat{\beta} = \frac{1}{T} \sum_{t=1}^T \hat{\beta}_t. \quad (10.35)$$

Remark 10.5 (*Step 0**) *The FMB can also be used to test CAPM (or other linear factor models). In this case, y_{it} in (10.34) are the excess returns on asset i in period t (R_{it}^e) and x_{it} are the loadings (γ_{it}) of the excess return on the market excess return (or other factors) according to the regression $R_{it}^e = \alpha + f'_t \gamma_{it} + \varepsilon_{it}$. In many cases, the γ_{it} values used as x_{it} are estimated during a previous sample, for instance, during the five years up to and including $t - 1$. In other cases, the γ_{it} values are estimated from the full sample, and are thus constant across periods. The latter has the advantage of being more precise estimates, provided the assumption of constant loadings is correct.*

Fama and MacBeth (1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\beta}_t$. In particular, they suggest that the variance of $\hat{\beta}_t$ (notice, not $\hat{\beta}$) can be estimated by the (average) squared variation around its mean

$$\text{Var}(\hat{\beta}_t) = \frac{1}{T} \sum_{t=1}^T (\hat{\beta}_t - \hat{\beta})^2. \quad (10.36)$$

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

$$\text{Var}(\hat{\beta}) = \frac{1}{T} \text{Var}(\hat{\beta}_t) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\beta}_t - \hat{\beta})^2. \quad (10.37)$$

When x_{it} are constant across time, then FMB and pooled OLS give the same point estimates (provided (10.34) is estimated without an intercept, effectively setting $\lambda_t = 0$). However, FMB's $\text{Var}(\hat{\beta})$ automatically handles the cross sectional correlations between residuals, while the pooled OLS would require applying a cluster method.

It can be noticed that when x_{it} is time-varying, then the FMB approach is not the same as OLS on pooled data. In fact, FMB is focused on the average cross-sectional affect, not on the time-series effect. For instance, regressions where all fixed effects have been taken out by demeaning are the same in FMB and pooled OLS.

See Table 10.2 shows estimates of the factor risk premia from several methods based on the 25 FF portfolios.

	Data	CR	FMB1	FMB2
Rm	6.34 (1.95)	5.98 (2.05)	5.98 (1.97)	-8.51 (3.55)
SMB	2.46 (1.33)	2.30 (1.41)	2.30 (1.37)	1.98 (1.37)
HML	4.38 (1.22)	4.97 (1.36)	4.97 (1.26)	4.53 (1.26)

Table 10.2: Different estimates of factor risk premia, annualized %. Numbers in (parentheses) are standard deviations. The 25 FF portfolios 1957:1-2016:12. Data are the mean excess returns of the factors; CR are estimates of the factor risk premia from a cross-sectional regression; FMB1 are from Fama-MacBeth without intercept in the cross-sectional regression; FMB2 are from Fama-MacBeth with intercept in the cross-sectional regression. In both FMB regressions, the betas are estimated from the full sample.

10.9 Calendar Time and Cross Sectional Regression

10.9.1 Calendar Time Approach

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)

$$\bar{y}_{jt} = \frac{1}{N_j} \sum_{i \in \text{Group } j} y_{it}, \quad (10.38)$$

where N_j is the number of individuals in group j . Notice that \bar{y}_{jt} is just one time series of the equally-weighted portfolio return for investors in group j .

Then, we run a factor model

$$\bar{y}_{jt} = x_t' \beta_j + v_{jt}, \text{ for } j = 1, 2, \dots, M \quad (10.39)$$

where x_t typically includes a constant and various return factors, for instance, excess returns on equity and bonds. Notice that x_t is the same for all groups. By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the "alpha") is higher for the M th group than for the first group.

Example 10.6 (*CalTime with two investor groups*) *With two investor groups, estimate the following SURE system*

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

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.

	Inactive	Active	Higly Active
coef	-0.76	3.08	8.65
t-tstat NW	-0.69	1.77	2.73

Table 10.3: Annualised alphas and t-stats from Table 10, in Dahlquist et al (RFS 2017). Three EW portfolios based on 62640 individuals, 2116 days. The dependent variables are the returns of the EW portfolio based on the activity indicators. The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

See Table 10.3 for results on a ten-year panel of some 60,000 Swedish pension savers from Dahlquist, Martinez, and Söderlind (2016). In this case, the dependent variable is the return of a pension investment portfolio (on day t , individual i). Each individual is sorted into one EW portfolio based of her/his trading activity over the last year (inactive active, very active).

The regressors include a constant, 7 risk factors (global and Swedish market, SMB, HML as a well as a bond factor) on ± 2 days ($1 + 7 \times 5$ regressors).

10.9.2 Cross Sectional Regression

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

$$y_{it} = x_i' \beta_i + \varepsilon_{it}, \text{ for } i = 1, 2, \dots, N \quad (10.40)$$

and to run cross-sectional regressions of the (estimated) betas (for instance, for the p th factor) on the investor characteristics

$$\hat{\beta}_{p,i} = z_i' c + u_i. \quad (10.41)$$

In this second-stage regression, the investor characteristics z_i could be a dummy variable (for age group, say) or a continuous variable (age, say). Notice that using a continuous 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 the cross-sectional correlation of the residuals (u_i). For instance, we may have a very large cross-sectional (N is large), but it so happens that many of the investors follow very similar investment strategies. Notice also that this approach can only handle the (time) average characteristics.

For an empirical illustration, see Table 10.4 where the t-stats look massively inflated.

	coef	t-tstat W	t-tstat C1	t-tstat C2	t-tstat C3
Inactive	-1.14	-17.27	-38.04	-13.39	-6.06
Active	5.95	42.13	6.70	18.77	11.62
Higly Active	10.25	25.80	9.91	11.64	25.71
Age	0.00	2.70	2.09	2.54	1.79
Male	0.43	20.66	7.83	23.61	27.88
Pension rights	-0.06	-8.55	-15.26	-6.76	-1.77

Table 10.4: Annualised regression coefficients and t-stats from a cross-sectional regression on the same data as in Dahlquist et al (RFS 2017). The dependent variable is the alpha of the 62640 individual portfolios. The alphas are from time series regressions which control for 7 risk factors over 5 days (2 lags, 2 leads). The t-stats are from White (W), clustering on activity (C1) age (C2) and pension rights (C3).

10.9.3 Cross Sectional Regression with Clustering of Residuals

When there is only one time period, then set $T = 1$ in (10.8)–(10.9), so (10.12) becomes

$$\text{Var}(\sqrt{N}\hat{\beta}) = \Sigma_{xx}^{-1} \text{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^N h_i\right) \Sigma_{xx}^{-1}. \quad (10.42)$$

When we assume that the residuals (or here, $h_i = x_i u_i$) are iid, then $\text{Var}(\sum_{i=1}^N h_i)$ is just the sum of the variances of each term, $\sum_{i=1}^N \text{Var}(h_i)$. With clustering, this is different. The following example illustrates this.

Example 10.7 (*Cluster method on $N = 4$*) Assume that individuals 1 and 2 form cluster 1 and that individuals 3 and 4 form cluster 2—and disregard correlations across clusters. This means setting the covariances across clusters to zero,

$$\begin{aligned} \text{Var}(\sum_{i=1}^N h_i) &= E(h_1^2 + h_2^2 + h_3^2 + h_4^2, \\ &\quad 2h_1h_2 + \underbrace{2h_1h_3}_0 + \underbrace{2h_1h_4}_0 + \underbrace{2h_2h_3}_0 + \underbrace{2h_2h_4}_0 + 2h_3h_4). \end{aligned}$$

(Recall that $E h_i = 0$, so $E h_i h_j = \text{Cov}(h_i, h_j)$.) Notice that this can be written

$$\text{Var}(\sum_{i=1}^N h_i) = E(h_1 + h_2)^2 + E(h_3 + h_4)^2.$$

Suppose there are C different clusters (and that we know which cluster i belongs to). Then, the previous example suggests that we can estimate the middle term of (10.42) as

$$S_C = \frac{1}{N} \sum_{c=1}^C h^c (h^c)', \text{ where } h^c = \sum_{i \in \text{cluster } c} h_i. \quad (10.43)$$

The iid case is when each i is her/his own cluster. In contrast, we cannot allow everyone to be in the same cluster, since this would give $h^c = 0$.

It is often argued that replacing h_i by $h_i C / (C - 1)$ improves the small properties of S_C .

10.10 Panel Regressions, Driscoll-Kraay and Cluster Methods

10.10.1 Pooled OLS as GMM

The k sample moment conditions for the pooled LS estimator (10.10) are

$$\bar{h} = \frac{1}{T} \sum_{t=1}^T \frac{1}{N} \sum_{i=1}^N h_{it} = \mathbf{0}_{k \times 1}, \text{ where} \quad (10.44)$$

$$h_{it} = x_{it} u_{it} = x_{it} (y_{it} - x'_{it} \beta). \quad (10.45)$$

Remark 10.8 (*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.

To apply this remark to the pooled OLS, notice that the Jacobian D_0 corresponds to (the probability limit of) the $-\Sigma_{xx}$ matrix (see (10.8)) so we get

$$\text{Var}(\sqrt{TN}\hat{\beta}) = \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1}, \text{ where } S_0 = \text{Cov}(\sqrt{TN}\bar{h}). \quad (10.46)$$

This is the same expression as in (10.12), except that we express the variance of $\sqrt{TN}\hat{\beta}$, not $\hat{\beta}$. Clearly, the value of $\text{Cov}(\sqrt{TN}\bar{h})$ depends on how the elements in the average \bar{h} are correlated (across time and across individuals).

10.10.2 The Effect of Cross-Sectional Correlations

To simplify the exposition we first focus on the cross-sectional correlations by assuming that there are no autocorrelations. In this case, we can simplify as

$$S_0 = \text{Cov} \left(\frac{1}{\sqrt{TN}} \sum_{t=1}^T \sum_{i=1}^N h_{it} \right) \quad (10.47)$$

$$= \frac{1}{TN} \sum_{t=1}^T \text{Cov}(h_t), \text{ where } h_t = \sum_{i=1}^N h_{it}. \quad (10.48)$$

In this expression, h_t is the $k \times 1$ vector of cross-sectional ($i = 1, 2, \dots, N$) sums in period t . Since we use the covariance matrix of the moment conditions, heteroskedasticity is accounted for (as in White's method).

In general, $\text{Cov}(h_t)$ involves all the cross-sectional covariances. For instance, with $N = 2$ we have

$$\text{Cov}(h_{1t} + h_{2t}) = \text{Cov}(h_{1t}, h_{1t}) + \text{Cov}(h_{2t}, h_{2t}) + \text{Cov}(h_{1t}, h_{2t}) + \text{Cov}(h_{2t}, h_{1t}), \quad (10.49)$$

where each term is a $k \times k$ matrix. An iid assumption would assume that covariances across individuals (firms) are zero ($\text{Cov}(h_{1t}, h_{2t}) = 0$). In contrast, a cluster method may assume that such covariances are zero unless the two individuals belong to the same cluster (town, football club,...). The Driscoll-Kraay method makes no such assumptions.

10.10.3 From Driscoll-Kraay to Standard OLS (no autocorrelations)

We initially rule out autocorrelations. The methods summarised below all aim at estimating S_0 in (10.48) in a consistent way.

The Driscoll and Kraay (1998) (DK) estimates S_0 by

$$S_{DK} = \frac{1}{TN} \sum_{t=1}^T h_t h'_t, \quad (10.50)$$

where h_t is the $k \times 1$ vector of the cross-sectional sum of h_{it} in period t , as defined in (10.48).

Remark 10.9 (Relation to the notation in Hoechle (2007)) Hoechle writes $\text{Cov}(\hat{\beta}) = (X'X)^{-1} \hat{S}_T (X'X)^{-1}$, where $\hat{S}_T = \sum_{t=1}^T h_t h'_t$. Clearly, $X'X/(TN) = \Sigma_{xx}$ and $\hat{S}_T/TN = S$. Combining gives (10.46).

Example 10.10 (DK on $N = 4$) As an example, suppose there is one regressor ($k = 1$) and $N = 4$. Then, (10.48) gives the cross-sectional sum in period t

$$h_t = h_{1t} + h_{2t} + h_{3t} + h_{4t},$$

and the covariance matrix (10.50)

$$\begin{aligned}
TN \times S_{DK} &= \sum_{t=1}^T h_t h'_t \\
&= \sum_{t=1}^T (h_{1t} + h_{2t} + h_{3t} + h_{4t})^2 \\
&= \sum_{t=1}^T (h_{1t}^2 + h_{2t}^2 + h_{3t}^2 + h_{4t}^2 \\
&\quad + 2h_{1t}h_{2t} + 2h_{1t}h_{3t} + 2h_{1t}h_{4t} + 2h_{2t}h_{3t} + 2h_{2t}h_{4t} + 2h_{3t}h_{4t})
\end{aligned}$$

The term in parentheses is the sum of all the elements in this matrix of cross products ($h_{it}h_{jt}$):

$$\begin{array}{ccccc}
i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\
\underline{1} & h_{1t}^2 & h_{1t}h_{2t} & h_{1t}h_{3t} & h_{1t}h_{4t} \\
\underline{2} & h_{2t}h_{1t} & h_{2t}^2 & h_{2t}h_{3t} & h_{2t}h_{4t} \\
\underline{3} & h_{3t}h_{1t} & h_{3t}h_{2t} & h_{3t}^2 & h_{3t}h_{4t} \\
\underline{4} & h_{4t}h_{1t} & h_{4t}h_{2t} & h_{4t}h_{3t} & h_{4t}^2
\end{array}$$

This means that all cross-sectional covariances are allowed to be non-zero. In case h_{it} is a $k \times 1$ vector, replace (the scalar) $h_{it}h_{jt}$ by the ($k \times k$ matrix) $h_{it}h'_{jt}$.

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 C clusters ($c = 1, \dots, C$)

$$S_C = \frac{1}{TN} \sum_{t=1}^T \sum_{c=1}^C h_t^c (h_t^c)' \text{, where } h_t^c = \sum_{i \in \text{cluster } c} h_{it}. \quad (10.51)$$

(Clearly, with only one cluster, then we are back in the DK method (10.50).)

Example 10.11 (*Cluster method on $N = 4$, changing Example 10.10 directly*) Reconsider Example 10.10, 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,

$$TN \times S_C = \sum_{t=1}^T (h_{1t}^2 + h_{2t}^2 + h_{3t}^2 + h_{4t}^2, \\ 2h_{1t}h_{2t} + \underbrace{2h_{1t}h_{3t}}_0 + \underbrace{2h_{1t}h_{4t}}_0 + \underbrace{2h_{2t}h_{3t}}_0 + \underbrace{2h_{2t}h_{4t}}_0 + 2h_{3t}h_{4t}).$$

In this case, the term in parentheses sums all the elements in this matrix:

$$\begin{array}{ccccc} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & h_{1t}^2 & h_{1t}h_{2t} & 0 & 0 \\ \underline{2} & h_{1t}h_{2t} & h_{2t}^2 & 0 & 0 \\ \underline{3} & 0 & 0 & h_{3t}^2 & h_{3t}h_{4t} \\ \underline{4} & 0 & 0 & h_{3t}h_{4t} & h_{4t}^2 \end{array}$$

This disregards any cross-sectional correlations across clusters.

Instead, we get *White's covariance matrix* by excluding all cross-sectional cross terms. This can be accomplished by defining

$$S_W = \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N h_{it}h'_{it}. \quad (10.52)$$

(This can be interpreted as a cluster method (10.51) where each i is its own cluster.) Notice that this disregards any cross-sectional correlations.

Example 10.12 (*White's method on $N = 4$*) With only one regressor (10.52) gives

$$TN \times S = \sum_{t=1}^T (h_{1t}^2 + h_{2t}^2 + h_{3t}^2 + h_{4t}^2),$$

so the term in parentheses sums all the elements in this matrix:

$$\begin{array}{ccccc} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & h_{1t}^2 & 0 & 0 & 0 \\ \underline{2} & 0 & h_{2t}^2 & 0 & 0 \\ \underline{3} & 0 & 0 & h_{3t}^2 & 0 \\ \underline{4} & 0 & 0 & 0 & h_{4t}^2 \end{array}$$

Finally, the *traditional LS covariance matrix* assumes that White's estimate (10.52) can be simplified by exploiting the fact that $x_{it}x'_{it}$ and u_{it}^2 are not correlated. This changes

S_W to $\Sigma_{xx}s^2$ where $s^2 = \sum_{t=1}^T \sum_{i=1}^N u_{it}^2 / TN$. Using in (10.46) gives

$$\text{Cov}_{LS}(\sqrt{TN}\hat{\beta}) = \Sigma_{xx}^{-1}s^2. \quad (10.53)$$

10.10.4 Reintroducing Autocorrelations

The previous analysis disregarded autocorrelations. We now reintroduce this possibility.

For the DK estimator, first define the estimate of the p th autocovariance matrix by

$$S_{DK,p} = \frac{1}{TN} \sum_{t=1}^T h_t h'_{t-p}. \quad (10.54)$$

When $p = 0$, then this is clearly the same as S_{DK} in (10.50). If we allow for P lags, then

the estimate of S_0 is

$$S_{DK} = S_{DK,0} + \sum_{p=1}^P w_p (S_{DK,p} + S'_{DK,p}), \quad (10.55)$$

where $w_p = 1 - p/(P + 1)$ in case we use the Bartlett weights (as in Newey-West), but also $w_p = 1$ can be motivated (see Petersen (2009) for a discussion).

The cluster method is to first define

$$S_{C,p} = \frac{1}{TN} \sum_{t=1}^T \sum_{c=1}^C h_t^c (h_{t-p}^c)' \quad (10.56)$$

and then use

$$S_C = S_{C,0} + \sum_{p=1}^P w_p (S_{C,p} + S'_{C,p}). \quad (10.57)$$

Finally, if we rule out all correlations across individuals, then we set

$$S_{W,p} = \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N h_{it} h'_{i,t-p} \quad (10.58)$$

and use

$$S_W = S_{W,0} + \sum_{p=1}^P w_p (S_{W,p} + S'_{W,p}). \quad (10.59)$$

10.10.5 A Monte Carlo Experiment

Reference: Dahlquist, Martinez, and Söderlind (2016)

Basic Setup

This section reports results from a simple Monte Carlo experiment. We use the model

$$y_{it} = \alpha + \beta f_t + \gamma g_i + u_{it}, \quad (10.60)$$

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 $t = 1, \dots, 2000$ and $i = 1, \dots, 1665$. 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 u_{it} is 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.

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

$$u_{it} = \varepsilon_{it} + v_{jt} + w_t, \quad (10.61)$$

where $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, $v_{jt} \sim N(0, \sigma_v^2)$ and $w_t \sim N(0, \sigma_w^2)$ —and the three components are uncorrelated. This implies that

$$\begin{aligned} \text{Var}(u_{it}) &= \sigma_\varepsilon^2 + \sigma_v^2 + \sigma_w^2, \\ \text{Cov}(u_{it}, u_{kt}) &= \begin{cases} \sigma_v^2 + \sigma_w^2 & \text{if individuals } i \text{ and } k \text{ belong to the same group} \\ \sigma_w^2 & \text{otherwise.} \end{cases} \end{aligned} \quad (10.62)$$

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 individual shocks ε_{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.

Results from the Monte Carlo Simulations

Table 10.5 reports the fraction of times the absolute value of a t-statistic for a true null hypothesis is higher than 1.96. The table has three panels for different correlation patterns the residuals (u_{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) show that with a completely random cluster specification (unknown to the econometricians), 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).

Remark 10.13 (*Interpretation of the simulations results**) To understand this last result, consider a stylised case where $y_{it} = \delta g_i + u_{it}$ where $\delta = 0$ and $u_{it} = w_t$ so all residuals

are due to an (excluded) time fixed effect. In this case, the matrix above becomes

$$\begin{bmatrix} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & w_t^2 & \underline{w_t^2} & -w_t^2 & -w_t^2 \\ \underline{2} & \underline{w_t^2} & w_t^2 & -w_t^2 & -w_t^2 \\ \underline{3} & -w_t^2 & -w_t^2 & w_t^2 & \underline{w_t^2} \\ \underline{4} & -w_t^2 & -w_t^2 & \underline{w_t^2} & w_t^2 \end{bmatrix}$$

(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 sum 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 $S = 0$.

10.10.6 An Empirical Illustration

Based on Table 4, regression [2] in Karnaukh, Ranaldo, and Söderlind (2015)Table 10.6 shows point estimates and Table 10.7 four different sets of t-stats.

10.11 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

$$y_{it} = x_t' \beta_1 + u_{it}^{(1)} \text{ for } i \in \text{group 1} \quad (10.63)$$

$$y_{it} = x_t' \beta_2 + u_{it}^{(2)} \text{ for } i \in \text{group 2}. \quad (10.64)$$

More interestingly, these regression equations can be combined into one *single* 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)

$$\begin{aligned}
y_{it} &= (z_{1i}x_t)' \beta_1 + (z_{2i}x_t)' \beta_2 + u_{it} \\
&= \left(\begin{bmatrix} z_{1i}x_t \\ z_{2i}x_t \end{bmatrix} \right)' \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u_{it} \\
&= (z_i \otimes x_t)' \beta + u_{it}, \text{ where } z_i = \begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix}.
\end{aligned} \tag{10.65}$$

This is estimated with LS by stacking all NT observations.

To see why the CalTime approach implicitly handles correlations within the groups (clusters), notice that the CalTime approach (10.39) and the panel approach (10.65) give the same coefficients. This makes it clear that the errors in CalTime are just group averages of the errors in the panel regressions

$$v_{jt} = \frac{1}{N_j} \sum_{i \in \text{Group } j} u_{it}^{(j)}. \tag{10.66}$$

We know that

$$\text{Var}(v_{jt}) = \frac{1}{N_j} (\bar{\sigma}_{ii} - \bar{\sigma}_{ih}) + \bar{\sigma}_{ih}, \tag{10.67}$$

where $\bar{\sigma}_{ii}$ is the average $\text{Var}(u_{it}^{(j)})$ and $\bar{\sigma}_{ih}$ 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).

Proof. (*of (10.67)) Write (10.66) as $v = \mathbf{1}'u/N$, where $\mathbf{1}$ is an $N \times 1$ vector of ones. It follows that $\text{Var}(v) = \mathbf{1}'\Sigma\mathbf{1}/N^2$, where Σ is the covariance matrix of u . Clearly, $\mathbf{1}'\Sigma\mathbf{1}$ is just the sum of all elements of Σ . First, the sum of all elements along the diagonal divided by N is the average variance, $\bar{\sigma}_{ii}$. Second, the sum of all off-diagonal elements divided by $N(N - 1)$ is the average covariance, $\bar{\sigma}_{ih}$. Therefore, $\mathbf{1}'\Sigma\mathbf{1}/N^2 = \bar{\sigma}_{ii}/N + \bar{\sigma}_{ih}N(N - 1)/N^2$, which can be rearranged as (10.67). ■

We could also consider the case *when the characteristics are not dummies* (like young or old), but rather continuous variable (for instance, age measured in years). For this case,

write the model as

$$y_{it} = (z_{it} \otimes x_t)'d + v_{it} \quad (10.68)$$

$$= ([1, z_{1it}, \dots, z_{mit}] \otimes [1, x_{1t}, \dots, x_{kt}])'d + u_{it}, \quad (10.69)$$

where z_{jit} measures characteristics j of investor i in period t and where x_{pt} is the p th regressor. 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 young). In other cases, $z_{ jit}$ is time invariant and contains static information about investor i .

This model is estimated with LS (stacking all NT observations), but the standard errors could be 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}).

Example 10.14 (*One investor characteristic and one pricing factor*). In this case (10.68) is

$$\begin{aligned} y_{it} &= \begin{bmatrix} 1 \\ x_t \\ z_{it} \\ z_{it}x_{1t} \end{bmatrix}' d + u_{it}, \\ &= d_0 + d_1 x_t + d_2 z_{it} + d_3 z_{it} x_{1t} + u_{it}. \end{aligned}$$

In case we are interested in how the investor characteristics (z_{it}) affect the intercept (alpha), then d_2 is the key coefficient. To see that, rearrange as

$$y_{it} = \underbrace{d_0 + d_2 z_{it}}_{\text{intercept}} + \underbrace{(d_1 + d_3 z_{it})x_t}_{\text{slope}} + u_{it}.$$

Clearly, d_2 shows how the characteristics z_{it} affects the intercept and d_3 how it affects the slope.

10.11.1 An Empirical Illustration

See Tables 10.8 for results on a ten-year panel of some 60,000 Swedish pension savers from Dahlquist, Martinez, and Söderlind (2016). In this case, the dependent variable is the return of a pension investment portfolio (on day t , individual i). The regressors include a constant, 7 risk factors (global and Swedish market, SMB, HML as a well as a

bond factor) on ± 2 days ($1 + 7 \times 5$ regressors), an indicator of trading activity of the individual over the last year (inactive active, very active).

The table illustrates the distinct difference in t-stats obtained by using different ways of handling the cross-sectional correlations (of the residuals). Notice, in particular, that the DK t-stats are the same as in calendar time approach (using Newey-West) in Table 10.3, although the estimation method is very different (here: a panel regression).

Table 10.9 is similar, but includes more regressors (age, gender and pension rights). This would be difficult to handle in a calendar time approach, and thus illustrates that a panel regression can handle more general cases.

10.12 The Results in Hoechle, Schmid and Zimmermann

Hoechle, Schmid, and Zimmermann (2015) (HSZ) prove the following two propositions about (10.68)–(10.69).

Proposition 10.15 *If the z_{it} vector in (10.68) consists of dummy variables indicating exclusive and constant group membership ($z_{1it} = 1$ means that investor i belongs to group 1, so $z_{j it} = 0$ for $j = 2, \dots, m$), then the LS estimates and DK standard errors of (10.68) are the same as LS estimates and Newey-West standard errors of the CalTime approach (10.39). (See HSZ for a proof.)*

This proposition basically says that panel regression is as good as the CT approach. So why use a panel regression, then? A. Because it allows for (a) many characteristics (poor, old, men) without having to define a very large set of dummies (poor&old&men, poor&old&female, poor&young&men,...); (b) a finer (continuous) characteristics grid (age in years, months, days and...).

Proposition 10.16 *(When z_{it} is a measure of investor characteristics, eg number of fund switches) The LS estimates and DK standard errors of (10.68) are the same as the LS estimates of CrossReg approach (10.41), but where the standard errors account for the cross-sectional correlations, while those in the CrossReg approach do not. (See HSZ for a proof.)*

	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 10.5: **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 $r_{it} = \alpha + \beta f_t + \gamma g_i + u_{it}$, 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 $t = 1, \dots, 2000$ and $i = 1, \dots, 1665$. Each individual belongs to one of five different clusters. The error term is constructed as $u_{it} = \varepsilon_{it} + v_{jt} + w_t$, where ε_{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 $(\varepsilon_{it}, v_{jt}, w_t)$ 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.

	Poor	Rich
cap flow	-4.2	-4.4
VIX	-8.6	-6.8
TED	-4.6	-6.8
MSCIw	-4.9	-2.0
FXvol	-24.1	-37.7
Stockvol	-4.2	-1.9
StockLiq	-1.7	-8.6
BondLiq	9.7	7.3
lag	-6.5	-1.3

Table 10.6: Regression coefficients (in %) from Table 4, regression [2] in Karnaukh et al (RFS 2015), 1995:01–2009:12. Panel regressions of 30 FX liquidity time-series on (common) drivers.

	OLS		White's		Cluster		DK	
	Poor	Rich	Poor	Rich	Poor	Rich	Poor	Rich
cap flow	-2.14	-2.12	-2.12	-2.47	-1.50	-1.51	-1.26	-1.35
VIX	-2.63	-1.97	-2.26	-1.81	-1.79	-1.22	-1.64	-1.12
TED	-2.46	-3.43	-2.16	-3.42	-1.79	-2.37	-1.67	-2.19
MSCIw	-2.20	-0.86	-1.95	-0.74	-1.32	-0.46	-1.13	-0.46
FXvol	-10.46	-15.60	-6.61	-9.95	-4.86	-5.26	-4.43	-5.11
Stockvol	-1.61	-0.71	-1.25	-0.48	-0.83	-0.33	-1.03	-0.31
StockLiq	-0.75	-3.71	-0.60	-2.63	-0.43	-2.00	-0.40	-1.90
BondLiq	4.68	3.33	3.94	2.83	2.49	1.93	2.06	1.89
lag	-3.38	-0.62	-2.93	-0.53	-1.98	-0.34	-1.72	-0.30

Table 10.7: Different t-stats for Table 4, regression [2] in Karnaukh et al (RFS 2015), 1995:01–2009:12. Clustering is done according rich/poor (on average). All methods, except OLS, allow for first-order autocorrelation.

	coef	t-tstat W	tstat DK
Inactive	-0.76	-56.89	-0.69
Active	3.08	37.48	1.77
Higly Active	8.65	28.73	2.73

Table 10.8: Annualised regression coefficients and different t-stats from Table 10, regressions I and II in Dahlquist et al (RFS 2017). Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t , individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

	coef	t-tstat W	tstat DK
Inactive	-1.10	-1.63	-0.69
Active	3.10	34.61	1.79
Higly Active	8.69	28.44	2.74
Age	0.00	0.19	0.11
Male	0.62	2.94	2.22
Pension rights	-0.03	-0.39	-0.33

Table 10.9: Annualised regression coefficients and different t-stats from Table 10, regressions I and II in Dahlquist et al (RFS 2017). Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t , individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

Chapter 11

Expectations Hypothesis of Interest Rates

11.1 Term (Risk) Premia

Term risk premia can be defined in several ways and this section will define three of them. All these premia have zero (or at least constant) expected values under the *expectations hypothesis* (EH).

Remark 11.1 (*Prices and yields on zero-coupon bonds**) Consider an m -period zero coupon bond. Its price $B(m)$ and continuously compounded interest rate $y(m)$ are related according to

$$B(m) = \exp[-my(m)].$$

Similarly, the continuously compounded forward rate for an m -period investment that starts in k periods ahead is

$$f(k, k+m) = \frac{1}{m} \ln \frac{B(k)}{B(k+m)} = \frac{(k+m)y(k+m) - ky(k)}{m}.$$

The (realized) *yield term premium* is defined as the difference between a long (n -period) interest rate and the average future short (m -period) rates over the same period

$$\varphi_{t+n}^y = y_{nt} - \frac{1}{k} \sum_{s=0}^{k-1} y_{m,t+sm}, \text{ with } k = n/m. \quad (11.1)$$

The EH says that its expected value should be constant

$$E_t \varphi_{t+n}^y = \alpha, \quad (11.2)$$

which means that rolling over short bonds is expected to yield the same (plus a constant) as holding a long bond. Figure 11.1 illustrates the timing.

Example 11.2 (Yield term premium, rolling over 3-month rates for a year). Let $y_{1y,t}$ be the current 1-year rate and $y_{3M,t+sM}$ the 3-month rate s months ahead

$$\varphi_{t+1y}^y = y_{1y,t} - \frac{1}{4} E_t (y_{3M,t} + y_{3M,t+3M} + y_{3M,t+6M} + y_{3M,t+9M}).$$

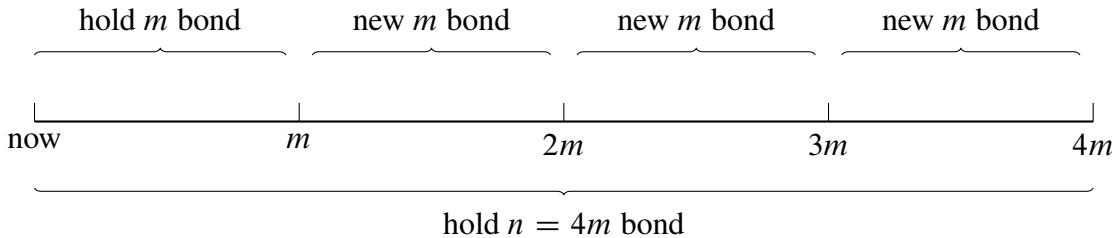


Figure 11.1: Timing for yield term premium

Let $f_t(k, k+m)$ be the forward rate that applies for the future period $t+k$ to $t+k+m$. The (realized) *forward term premium* is the difference between a forward rate for an m -period investment (starting in k periods ahead) and the short interest rate for the same period

$$\varphi_{t+k+m}^f = f_t(k, k+m) - E_t y_{m,t+k}. \quad (11.3)$$

The EH says that its expected value should be constant

$$E_t \varphi_{t+k+m}^f = \alpha, \quad (11.4)$$

which means that the expected profit from a forward contract is always the same (possibly zero). Figure 11.2 illustrates the timing.

Example 11.3 (Forward term premium, 1-month investment starting 2 months from now) Let $f_t(2M, 3M)$ be the 1-month forward rate starting 2 months ahead and let $y_{1M,t+2M}$ be the one-month interest rate over the same period. Then,

$$\varphi_{t+3M}^f = f_t(2M, 3M) - E_t y_{1M,t+2M}.$$

Finally, the (realized) *holding-period premium* is the 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

$$\varphi_{t+m}^h = \frac{1}{m} \ln(P_{n-m,t+m}/P_{nt}) - y_{mt}. \quad (11.5)$$

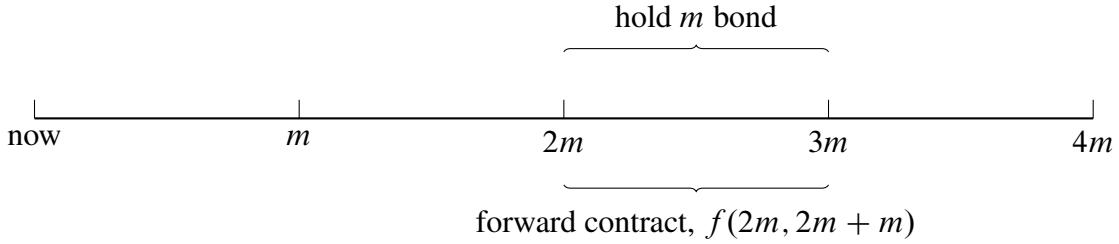


Figure 11.2: Timing for forward term premium

As before, the EH says that its expected value should be constant

$$E_t \varphi_{t+m}^h = \alpha, \quad (11.6)$$

which means that the expected profit from holding a long bond over a short period is expected to yield the same (plus a constant) as holding a short bond. This version is perhaps the most similar to the definition of risk premia of other assets (for instance, equity). Figure 11.3 illustrates the timing.

Example 11.4 (*Holding-period premium, holding a 10-year bond for one year*).

$$\begin{aligned} \varphi_t^h(10y, 1y) &= E_t \ln(P_{9y,t+1}/P_{10y,t}) - y_{1y,t}. \\ &= [10y_{10y,t} - 9 E_t y_{9y,t+1}] - y_{1y,t}. \end{aligned}$$

The second line just rewrites the bond prices in terms of the interest rates.

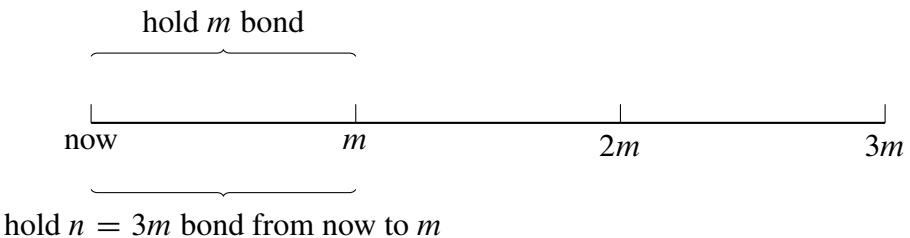


Figure 11.3: Timing for holding-period premium

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(er) rate 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.

11.2 Testing the Expectations Hypothesis of Interest Rates

11.2.1 Basic Tests

The expectations hypothesis (see (11.2), (11.4) and (11.6)) says that the *expected* term premia should be constant. This implies that the *realized* term premia should be unpredictable by any information today (denoted x_t below). That is, the slopes (b) in the following regression should be zero

$$\varphi_{t+s} = a + b'x_t + u_{t+s}, \quad (11.7)$$

where φ_{t+s} denotes any of the realized term premia in (11.1), (11.3) or (11.5). The intercepts in this regression picks out a constant term premia. A non-zero slope would indicate that the changes of the term premia are predictable—which is at odds with the expectations hypothesis.

11.2.2 The Role of Rational Expectations in the Basic Test

Notice that we use realized (instead of expected) values on the left hand side of the regression (11.7). This is valid—under the assumption that expectations can be well approximated by the properties of the sample data, that is, *rational expectations*.

To see that, notice that we really would like to estimate a regression model where the *expected* term premium is the dependent variable

$$E_t \varphi_{t+s} = a_0 + b'_0 x_t + \eta_{t+s}, \quad (11.8)$$

since this would tell us if market participants “price” risk when valuing bonds. Such a test could be performed if we have (good) data for surveys (see Froot (1989) for an example). However, most tests are done using realized (also called *ex post*) premia as in (11.7).

To analyse this, let ε_{t+s} be the prediction error, that is, let it be defined by

$$\varphi_{t+s} = E_t \varphi_{t+s} + \varepsilon_{t+s}. \quad (11.9)$$

We can then rewrite the regression model (11.7) as

$$E_t \varphi_{t+s} + \varepsilon_{t+s} = a + b'x_t + u_{t+s}. \quad (11.10)$$

If the prediction errors come from a rational prediction model, then they cannot be correlated to x_t , assuming x_t was known in t . (The prediction errors from a rational model can-

not be predictable—since that would mean that the predictions are systematically wrong.) In this case, the estimated b captures only how x_t is correlated with $E_t \varphi_{t+s}$ (which is what we want).

In contrast, if $E_t \varphi_{t+s}$ is a non-rational prediction model, then it is possible that the estimated b has nothing to do with $E_t \varphi_{t+s}$, but that it captures how the prediction errors can be explained by x_t . By *assuming* rational expectations, we rule out this possibility.

11.2.3 A Single Factor for All Maturities?

Reference: Cochrane and Piazzesi (2005)

Cochrane and Piazzesi (2005) regress excess holding period returns on forward rates, that is, x_t in (11.7) contains only forward rates. They observe that the slope coefficients are very similar across different maturities of the bonds. It seems as if the coefficients (b) for one maturity are the same as the coefficients for another maturity—apart from a scaling factor. This means that if we construct a “forecasting factor”

$$ff_t = b' x_t \quad (11.11)$$

for one maturity (2-year bond, say), then the regressions

$$\frac{1}{m} \ln(P_{n-m,t+m}/P_{nt}) - y_{mt} = a_n + b_n ff_t \quad (11.12)$$

should work almost as well as using the full vector x_t . Figure 11.4 and Tables 11.1–11.2 illustrate some results.

	2y	3y	4y	5y
factor	1.00 (6.70)	1.87 (6.82)	2.68 (6.99)	3.46 (7.15)
constant	-0.00 (-0.00)	-0.00 (-0.34)	-0.00 (-0.68)	-0.00 (-1.02)
R2	0.14	0.14	0.15	0.16
obs	624.00	624.00	624.00	624.00

Table 11.1: 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. Numbers in parentheses are t-stats. U.S. data for 1964:1–2016:12.

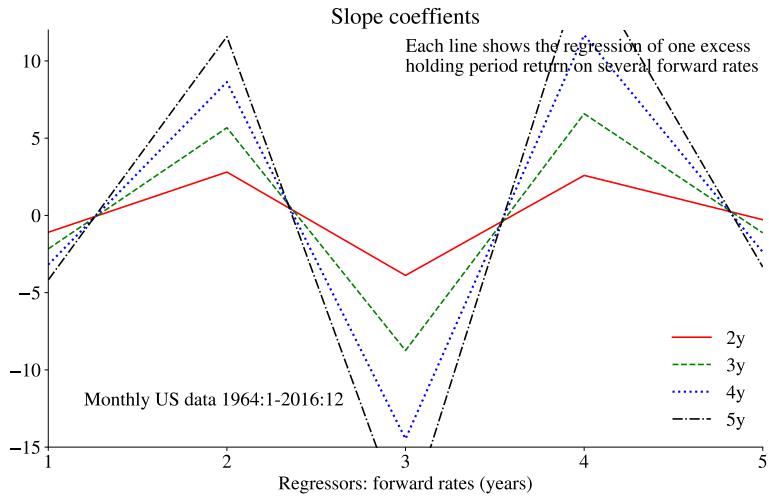


Figure 11.4: A single forecasting factor for bond excess hold period returns

	2y	3y	4y	5y
factor	1.00 (4.05)	1.87 (4.19)	2.68 (4.33)	3.46 (4.48)
constant	-0.00 (-0.00)	-0.00 (-0.16)	-0.00 (-0.32)	-0.00 (-0.48)
R2	0.14	0.14	0.15	0.16
obs	624.00	624.00	624.00	624.00

Table 11.2: 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 1964:1-2016:12. Numbers in parentheses are t-stats. Bootstrapped standard errors, with blocks of 10 observations.

11.3 Spread-Based Tests*

11.3.1 Basic Setup*

Many classical tests of the expectations hypothesis have only used interest rate spreads as predictors (x_t include only interest rates). Using spreads is a way to overcome issues with the long swings (being close to non-stationary) in interest rates. This is best illustrated by the yield term premium.

First, add and subtract y_{mt} (the current short m -period rate) from the yield term pre-

mium (11.1)

$$\begin{aligned}\varphi_{t+n}^y &= y_{nt} - \frac{1}{k} \sum_{s=0}^{k-1} y_{m,t+sm} + y_{mt} - y_{mt} \\ &= (y_{nt} - y_{mt}) - \frac{1}{k} \sum_{s=0}^{k-1} (y_{m,t+sm} - y_{mt}).\end{aligned}\quad (11.13)$$

If φ_{t+n}^y is constant across time, then the first and second term in (11.13) must move equally much—which means that $\beta = 1$ in a regression like

$$\frac{1}{k} \sum_{s=0}^{k-1} (y_{m,t+sm} - y_{mt}) = \alpha + \beta (y_{nt} - y_{mt}) + \varepsilon_{t+n}. \quad (11.14)$$

(A constant risk premium could be captured by $\alpha \neq 0$.) This regression tests if the current long-short spread is an unbiased predictor of the future changes in the short rate. See Figure 11.5 for an empirical example.

Example 11.5 (*Yield term premium, rolling over 3-month rates for a year*)

$$\frac{1}{4} [(y_{3M,t} - y_{3M,t}) + (y_{3M,t+3m} - y_{3M,t}) + (y_{3M,t+6M} - y_{3M,t}) + (y_{3M,t+9M} - y_{3M,t})]$$

regressed on $y_{1y,t} - y_{3M,t}$.

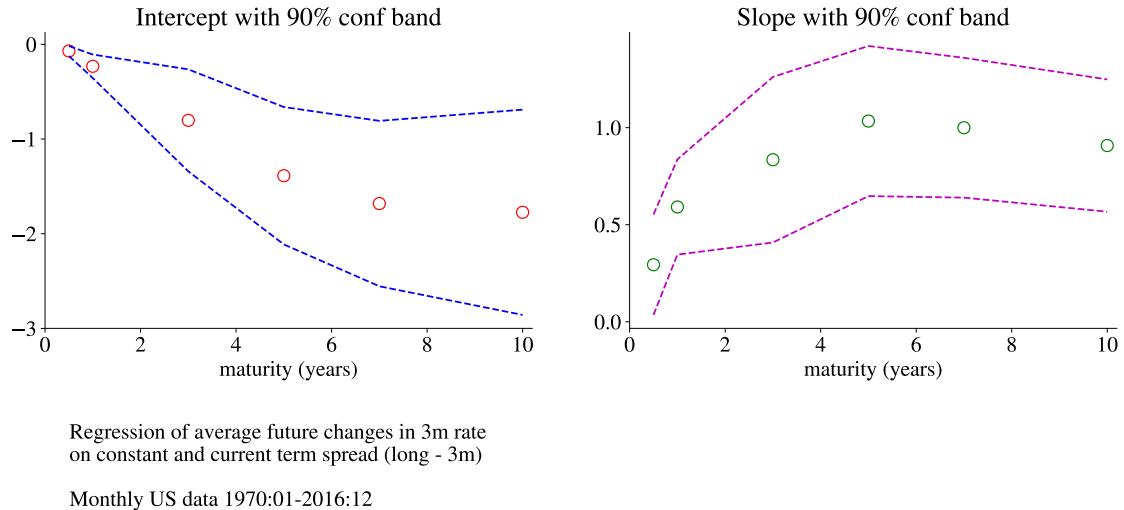


Figure 11.5: Testing the expectations hypothesis on US interest rates

Similarly, adding and subtracting y_{mt} to (11.3) and rearranging gives that EH requires $\beta = 1$ in

$$y_{m,t+k} - y_{mt} = \alpha + \beta [f_t(k, k+m) - y_{mt}] + \varepsilon_{t+k+m}. \quad (11.15)$$

This regression tests if the forward-spot spread is an unbiased predictor of the change of the spot rate.

Finally, rearrange (11.5), add and subtract my_{nt} and rewrite it in terms of interest rates show the EH requires $\beta = 1$ in

$$y_{n-m,t+m} - y_{nt} = \alpha + \beta \frac{m}{n-m} (y_{nt} - y_{mt}) + \varepsilon_{t+n}. \quad (11.16)$$

If the holding period (m) is short compared to the maturity (n), then this regression (almost) tests if the current spread, scaled by $m/(n - m)$, is an unbiased predictor of the change in the long rate.

11.3.2 The Properties of Spread-Based EH Tests*

Reference: Froot (1989)

The spread-based EH tests ((11.14), (11.15) and (11.16)), can be written

$$\Delta i_{t+1} = \alpha + \beta s_t + \varepsilon_{t+1}, \text{ where} \quad (11.17)$$

$$s_t = E_t^m \Delta i_{t+1} + \varphi_t, \quad (11.18)$$

where $E_t^m \Delta i_{t+1}$ 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 (11.17) is

$$\beta = 1 - \frac{\sigma(\sigma + \rho)}{1 + \sigma^2 + 2\rho\sigma} + \gamma, \quad \text{where} \quad (11.19)$$

$$\sigma = \frac{\text{Std}(\varphi)}{\text{Std}(E_t^m \Delta i_{t+1})}, \quad \rho = \text{Corr}(E_t^m \Delta i_{t+1}, \varphi), \quad \text{and}$$

$$\gamma = \frac{\text{Cov}[(E_t - E_t^m) \Delta i_{t+1}, E_t^m \Delta i_{t+1} + \varphi]}{\text{Var}(E_t^m \Delta i_{t+1} + \varphi)},$$

The second term in (11.19) captures the effect of the (time varying) risk premium and the third term (γ) captures any systematic expectations errors $((E_t - E_t^m) \Delta i_{t+1})$.

Figure 11.6 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

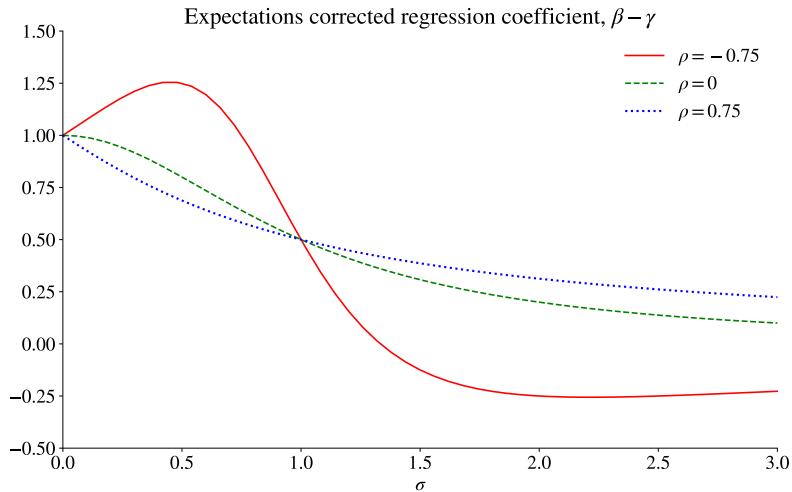


Figure 11.6: Regression coefficient in EH test

$(\rho = -\sigma)$, 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 (11.19)) Define

$$\begin{aligned}\Delta i_{t+1} &= E_t \Delta i_{t+1} + u_{t+1} \\ E_t \Delta i_{t+1} &= E_t^m \Delta i_{t+1} + \eta_{t+1}.\end{aligned}$$

The regression coefficient is

$$\begin{aligned}\beta &= \frac{\text{Cov}(s_t, \Delta i_{t+1})}{\text{Var}(s_t)} \\ &= \frac{\text{Cov}(E_t^m \Delta i_{t+1} + \varphi_t, E_t^m \Delta i_{t+1} + \eta_{t+1} + u_{t+1})}{\text{Var}(E_t^m \Delta i_{t+1} + \varphi_t)} \\ &= \frac{\text{Var}(E_t^m \Delta i_{t+1})}{\text{Var}(E_t^m \Delta i_{t+1} + \varphi_t)} + \frac{\text{Cov}(\varphi_t, E_t^m \Delta i_{t+1})}{\text{Var}(E_t^m \Delta i_{t+1} + \varphi_t)} + \frac{\text{Cov}(E_t^m \Delta i_{t+1} + \varphi_t, \eta_{t+1})}{\text{Var}(E_t^m \Delta i_{t+1} + \varphi_t)}\end{aligned}$$

The third term is γ . Write the first two terms as

$$\begin{aligned}\frac{\sigma_{mm} + \sigma_{m\varphi}}{\sigma_{mm} + \sigma_{\varphi\varphi} + 2\sigma_{m\varphi}} &= 1 + \frac{\sigma_{mm} + \sigma_{m\varphi} - (\sigma_{mm} + \sigma_{\varphi\varphi} + 2\sigma_{m\varphi})}{\sigma_{mm} + \sigma_{\varphi\varphi} + 2\sigma_{m\varphi}} \\ &= 1 - \frac{\rho\sigma_m\sigma_\varphi + \sigma_\varphi^2}{\sigma_m^2 + \sigma_\varphi^2 + 2\rho\sigma_m\sigma_\varphi} \\ &= 1 - \frac{(\rho\sigma_m\sigma_\varphi + \sigma_\varphi^2) / \sigma_m^2}{(\sigma_m^2 + \sigma_\varphi^2 + 2\rho\sigma_m\sigma_\varphi) / \sigma_m^2} \\ &= 1 - \frac{\sigma(\sigma + \rho)}{1 + \sigma^2 + 2\rho\sigma}\end{aligned}$$

where the second line multiplies by σ_m^2/σ_m^2 and the third line uses the definition $\sigma = \sigma_\varphi/\sigma_m$. ■

Chapter 11

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

11.1 Describing Yield Curves

On average, yield curves tend to be upward sloping (see Figure 11.2), but there is also considerable time variation on both the level and shape of the yield curves.

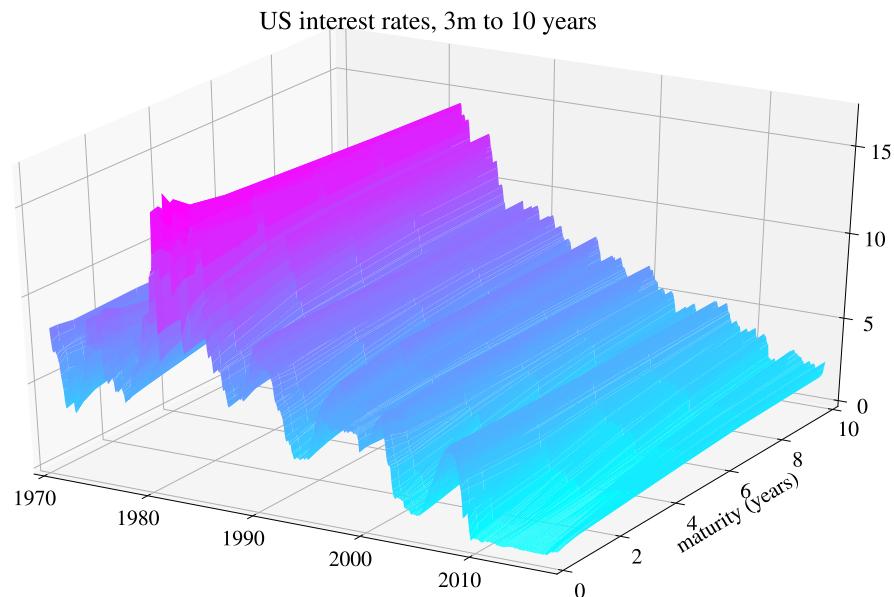


Figure 11.1: US yield curves

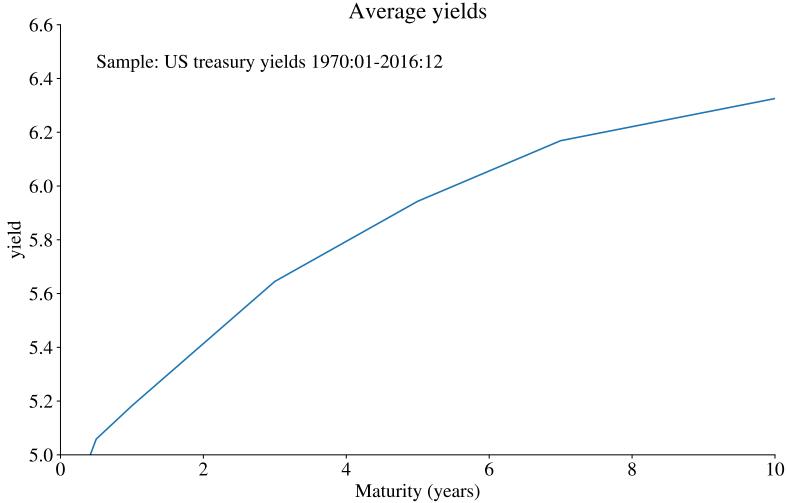


Figure 11.2: Average US yield curve

It is common to describe the movements in terms of three “factors”: level, slope, and curvature. One way of measuring these factors is by defining

$$\begin{aligned} \text{Level}_t &= y_{10y} \\ \text{Slope}_t &= y_{10y} - y_{3m} \\ \text{Curvature}_t &= (y_{2y} - y_{3m}) - (y_{10y} - y_{2y}). \end{aligned} \quad (11.1)$$

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 11.3 for an example.

An alternative way to study the yield curve factors is to use principal component analysis. See Figure 11.4 for an example.

Remark 11.1 (*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 eigenvector 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 $\lambda_1 = \text{Var}(w_1' z_t) = w_1' \Sigma w_1$. 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 w_j .*

tor associated with the j th largest eigenvalue (which equals $\lambda_j = \text{Var}(w'_j z_t) = w'_j \Sigma w_j$). If the rank of Σ is K , then only K eigenvalues are non-zero. From the properties of eigenvectors, it can be shown that $\Sigma = \sum_{i=1}^N \lambda_i w_i w'_i$. Taking the trace (summing the diagonal elements) gives $\sum_{i=1}^N \text{Var}(z_{it}) = \sum_{i=1}^N \lambda_i$. (This follows from that the sum of the diagonal elements of $\sum_{i=1}^N \lambda_i w_i w'_i$ equals $\sum_{i=1}^N \lambda_i (\sum_{j=1}^N w_{ji}^2)$ and that the inner sum is 1.) Together this means: (a) principal components are uncorrelated; (b) an eigenvalue equals the variance of the corresponding principal component (the first being most volatile etc); (c) the sum of the eigenvalues equals the total variance of the data ($\sum_{i=1}^N \text{Var}(z_{it})$). Dividing an eigenvalue with the sum of eigenvalues gives a measure of the relative importance of that principal component (in terms of variance).

Remark 11.2 (Principal component analysis 3) Let W be $N \times N$ matrix with w_i as column i . We can then calculate the $N \times 1$ vector of principal components as $pc_t = W' z_t$. Since $W^{-1} = W'$ (the eigenvectors are orthogonal), we can invert as $z_t = W pc_t$. 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.

Example 11.3 (PCA with 2 series) Let $w_i^{(j)}$ be the t th element in the i th eigenvector. With two series we have

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

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

$$\begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} = \begin{bmatrix} w_1^{(1)} & w_2^{(1)} \\ w_1^{(2)} & w_2^{(2)} \end{bmatrix} \begin{bmatrix} pc_{1t} \\ pc_{2t} \end{bmatrix}$$

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

Interest rates are strongly related to business cycle conditions, so it often makes sense to include macro economic data in the modelling. See Figure 11.5 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.

Diebold and Li (2006) use a more modern approach to study the yield curve factors.

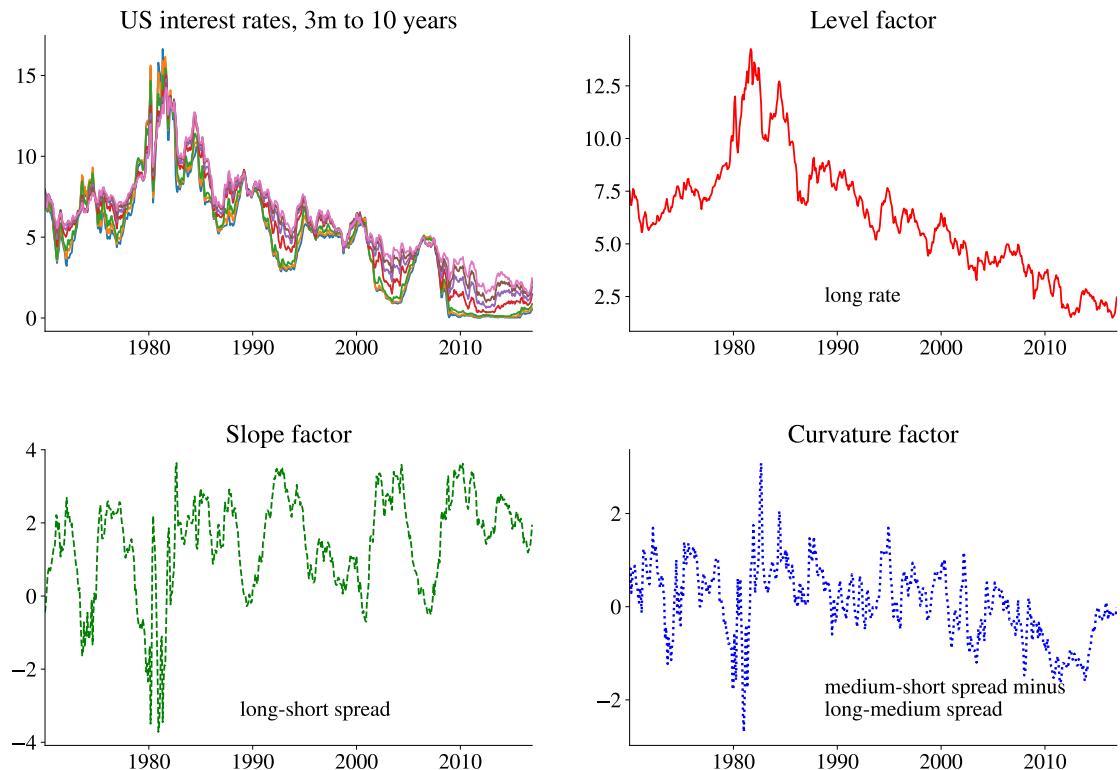


Figure 11.3: US yield curves: level, slope and curvature

The use the Nelson-Siegel model for an m -period interest rate, which says that

$$y(m) = \beta_0 1 + \beta_1 \frac{1 - \exp(-m/\tau_1)}{m/\tau_1} + \beta_2 \left[\frac{1 - \exp(-m/\tau_1)}{m/\tau_1} - \exp\left(-\frac{m}{\tau_1}\right) \right], \quad (11.2)$$

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 (β_0 is the level factor, β_1 the slope factor and β_2 the curvature factor).

Remark 11.4 (*Diebold and Li (2006) vs (11.1)*) Use (T_0, T_1, T_2) to label the terms in (11.2). Observe that we have (approximately)

$$\begin{bmatrix} m & T_0 & T_1 & T_2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 0.5 & 0.25 \\ \infty & 1 & 0 & 0 \end{bmatrix}$$

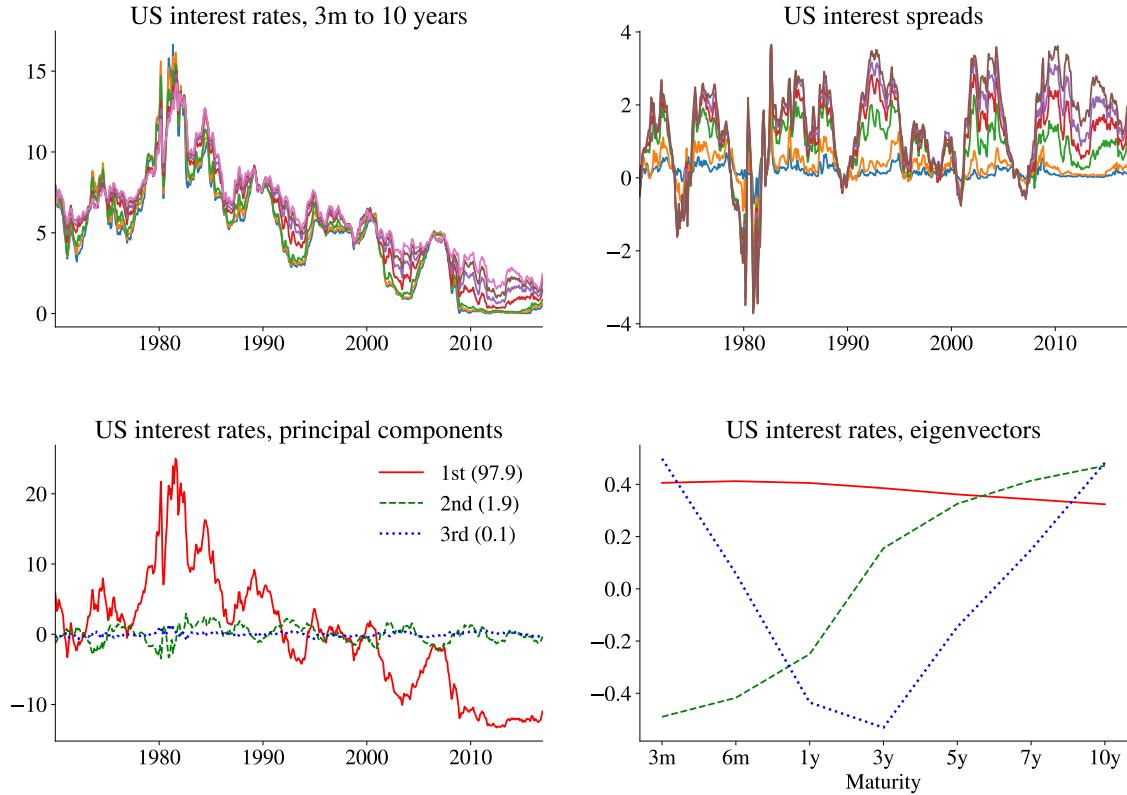


Figure 11.4: US yield curves and principal components

Therefore, $y(0) = \beta_0 + \beta_1$, $y(2) = \beta_0 + 0.5\beta_1 + 0.25\beta_2$ and $y(\infty) = \beta_0$. Comparing with (11.1) shows that the long rate corresponds to β_0 , that the spread between a very long and a very short rate corresponds to β_1 and that $[y(2) - y(0)] - [y(\infty) - y(2)]$ corresponds to $0.5\beta_2$.

See Figure 11.6 for an example. The results are very similar to the factors calculated directly from yields (cf. Figure 11.3).

11.2 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.

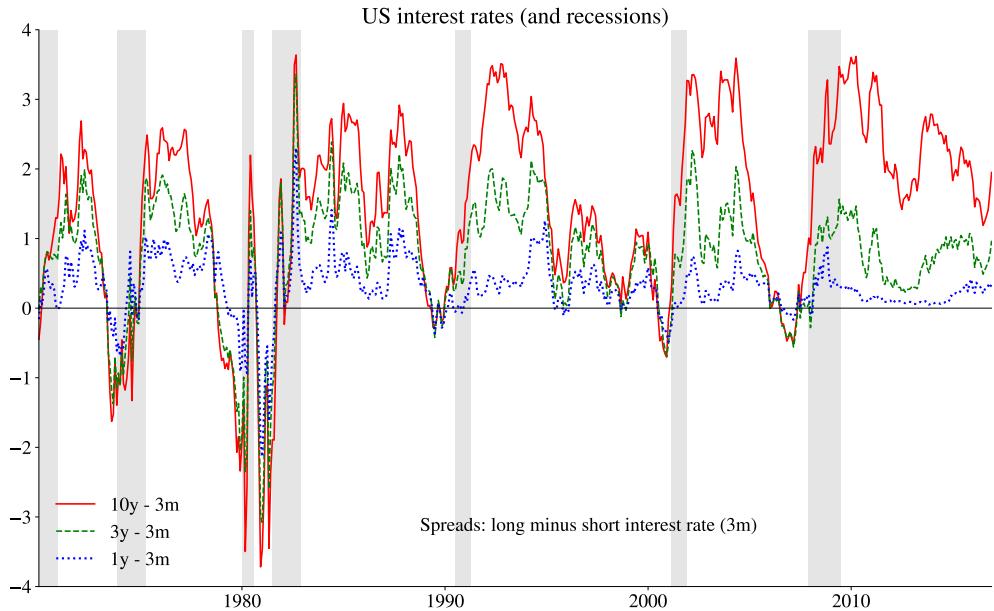


Figure 11.5: US term spreads (over a 3m T-bill)

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 11.7–11.9 provide some examples.

11.3 Summary of the Solutions of Some Affine Yield Curve Models

An affine yield curve model implies that the yield on an n -period *discount* (zero coupon) bond can be written

$$y_{nt} = a_n + b'_n x_t, \text{ where} \quad (11.3)$$

$$a_n = A_n/n \text{ and } b_n = B_n/n,$$

where x_t is an $K \times 1$ vector of state variables. The A_n (a scalar) and the B_n (a $K \times 1$ vector) are functions of the model parameters—discussed below.

To set up a consistent model (which does not create internal arbitrage opportunities) recall that the price of an n -period bond equals the cross-moment between the pricing kernel (denoted M_{t+1} , since m_{t+1} will be used for the logarithm) and the value of the

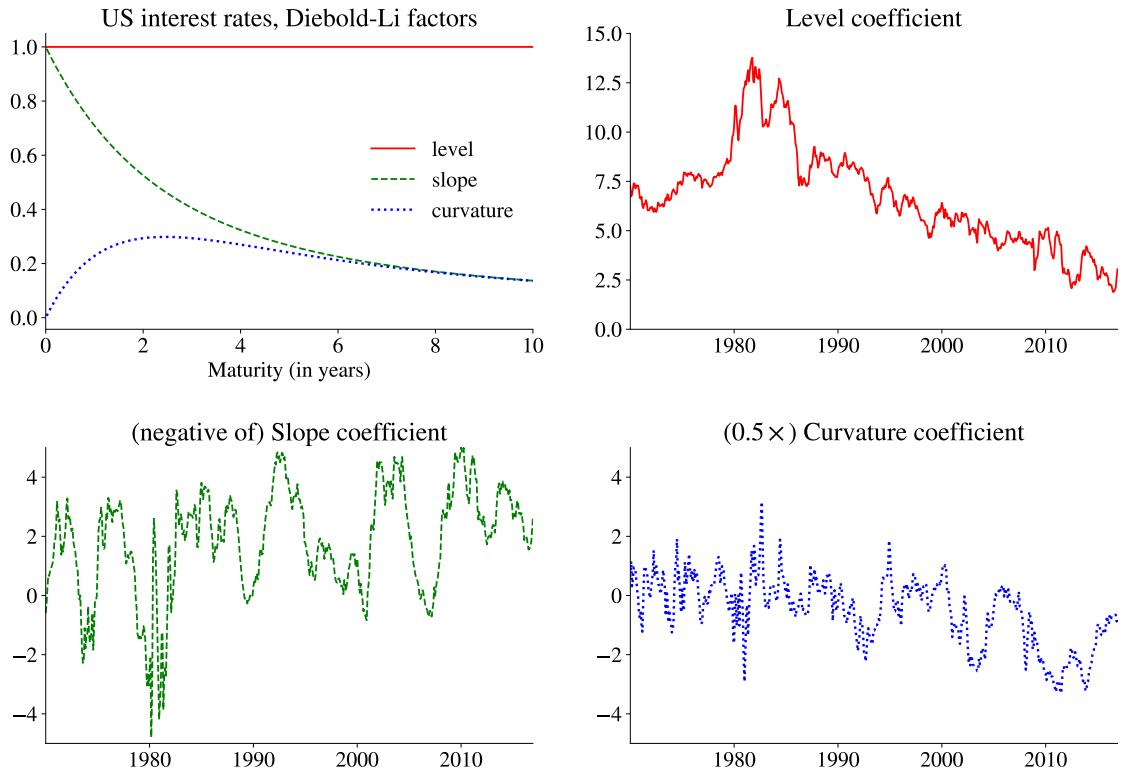


Figure 11.6: US yield curves: level, slope and curvature, Diebold-Li approach

same bond next period (then an $n - 1$ -period bond)

$$P_{nt} = E_t M_{t+1} P_{n-1,t+1}. \quad (11.4)$$

The **Vasicek (1977)** model assumes that the log SDF (m_{t+1}) is an affine function of a single AR(1) state variable x_t . It is convenient to write this as

$$-m_{t+1} = x_t + \lambda \sigma \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0, 1) \text{ and} \quad (11.5)$$

$$x_{t+1} = (1 - \rho) \mu + \rho x_t + \sigma \varepsilon_{t+1}. \quad (11.6)$$

To extend to a multifactor model, specify

$$-m_{t+1} = \mathbf{1}' x_t + \lambda' S \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0, I) \text{ and} \quad (11.7)$$

$$x_{t+1} = (I - \Psi) \mu + \Psi x_t + S \varepsilon_{t+1}, \quad (11.8)$$

where S and Ψ are matrices while λ and μ are (column) vectors, and $\mathbf{1}$ is a vector of ones.

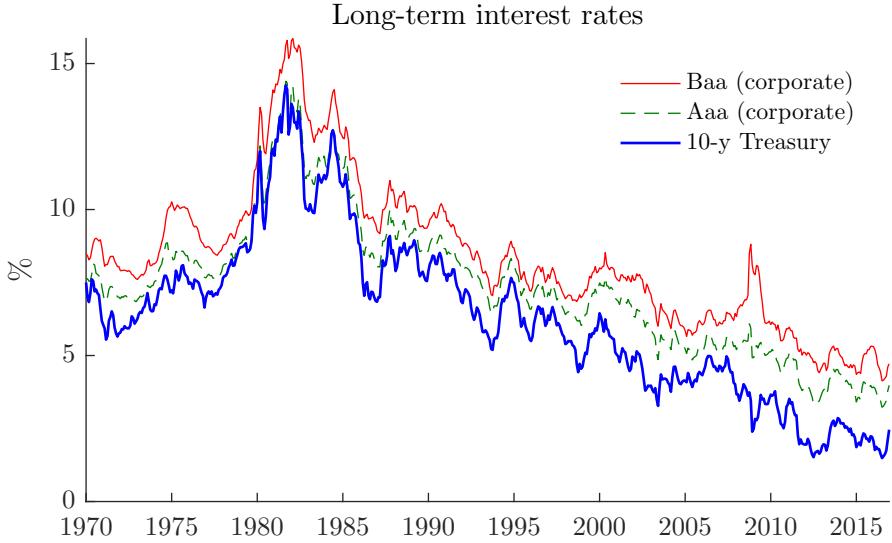


Figure 11.7: US interest rates

Example 11.5 (*Two-factor Vasicek model, independent factors*) A multivariate Vasicek model with two independent factors is

$$-m_{t+1} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \end{bmatrix}, \text{ and}$$

$$\begin{bmatrix} x_{1,t+1} \\ x_{2,t+1} \end{bmatrix} = \begin{bmatrix} 1 - \rho_1 & 0 \\ 0 & 1 - \rho_2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \end{bmatrix}.$$

For the multivariate Vasicek model we have the solution (see Appendix 11.6 for a proof)

$$B_n = \mathbf{1} + \Psi' B_{n-1}, \text{ and} \quad (11.9)$$

$$A_n = A_{n-1} + B_{n-1}' (I - \Psi) \mu - (\lambda' + B_{n-1}') S S' (\lambda + B_{n-1}) / 2, \quad (11.10)$$

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. Clearly, the single-factor Vasicek model is a special case. See Figure 11.10 for an illustration of the a_n and b_n coefficients for a single-factor model and Figure 11.11 for some typical yield curve shapes.

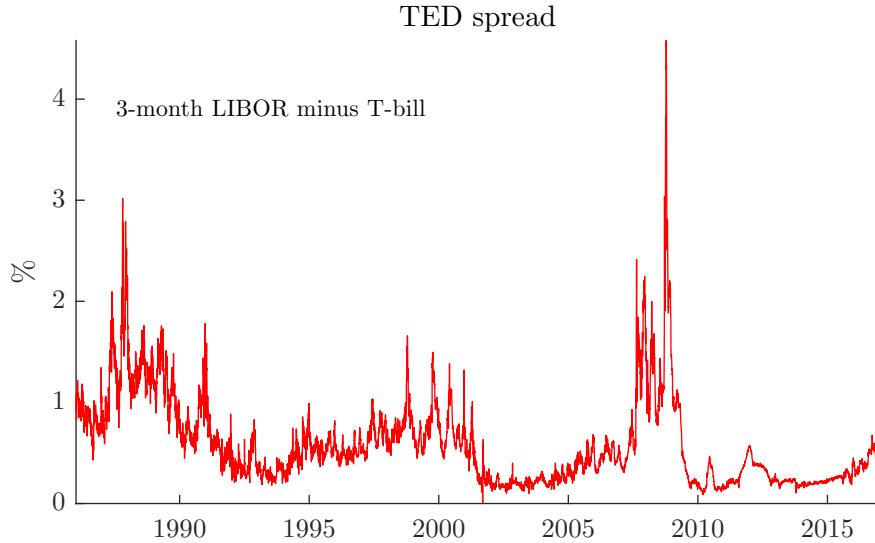


Figure 11.8: TED spread

Example 11.6 (A_n and B_n in the single-factor Vasicek model) (11.9–(11.10) give

$$B_0 = 0 \text{ and } A_0 = 0$$

$$B_1 = 1 \text{ and } A_1 = -\lambda^2 \sigma^2 / 2$$

$$B_2 = 1 + \rho \text{ and } A_2 = (1 - \rho) \mu - [\lambda^2 + (1 + \lambda)^2] \sigma^2 / 2.$$

If $\rho < 1$, then long rates are less sensitive to x_t than short rates are.

The interest rates on long bonds are mostly driven by the expectation of future short rates. In addition, the uncertainty/persistence/price of risk all affect the risk premia. See Figures 11.12–11.13 for an illustration.

Figure 11.14 illustrates the a_n and b_n coefficients in a 2-factor Vasicek model. Figure 11.15 shows that the 2-factor model is capable of generating many more shapes of the yield curve than a 1-factor model is.

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

$$\begin{aligned} y_{1t} &= a_1 + b'_1 x_t, \\ -m_{t+1} &= y_{1t} - \chi_{t+1}, \\ \chi_{t+1} &= -\theta'_t \theta_t / 2 - \theta'_t \varepsilon_{t+1}, \text{ with } \varepsilon_{t+1} \sim N(0, I). \end{aligned} \tag{11.11}$$

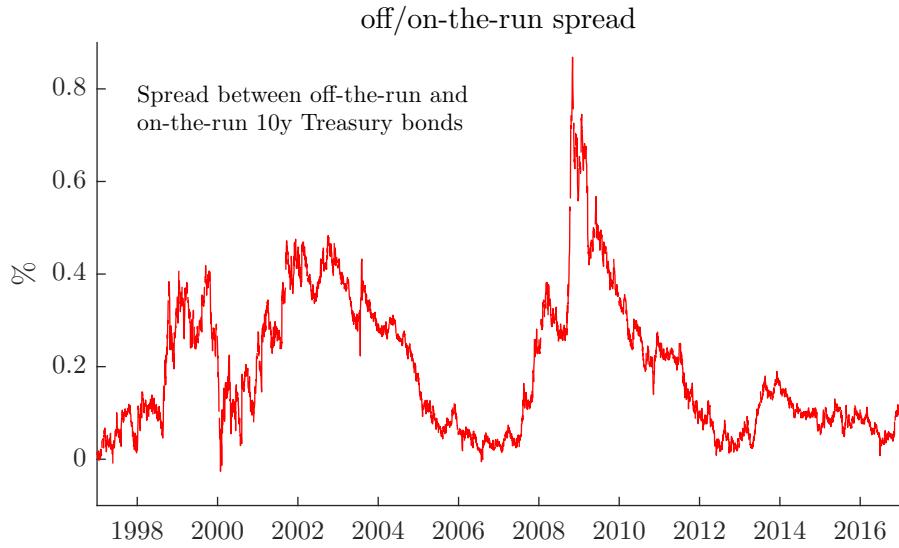


Figure 11.9: Off-the-run liquidity premium

The $K \times 1$ vector of market prices of risk (θ_t) is affine in the state vector

$$\theta_t = \theta^0 + \theta^1 x_t, \quad (11.12)$$

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 (11.8).

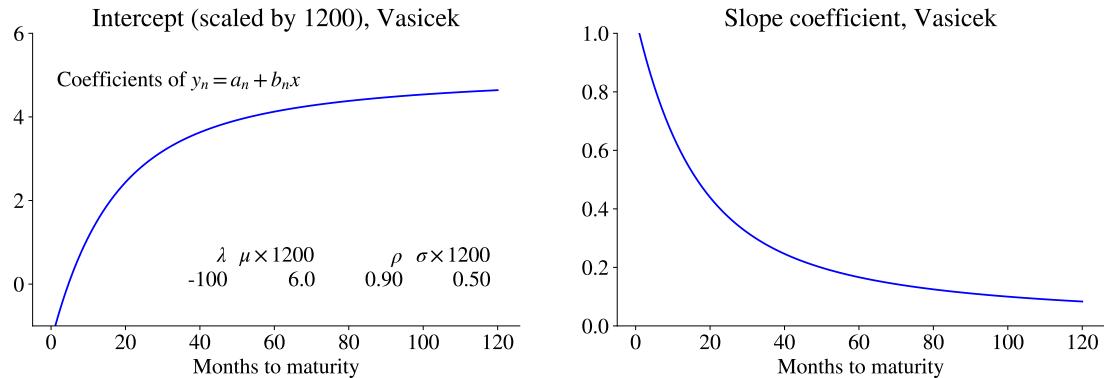


Figure 11.10: a_n and b_n in the Vasicek model

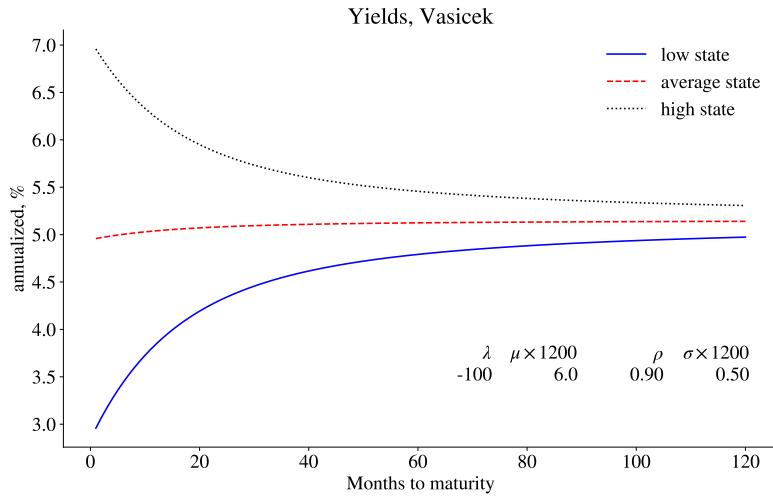


Figure 11.11: Yield curves in the Vasicek model

Example 11.7 (*Two factor model*) With $K = 2$, (11.11)–(11.12) are

$$\begin{aligned} y_{1t} &= a_1 + \begin{bmatrix} b_{1,1} \\ b_{1,2} \end{bmatrix}' \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix}, \\ \chi_{t+1} &= -\begin{bmatrix} \theta_{1t} \\ \theta_{2t} \end{bmatrix}' \begin{bmatrix} \theta_{1t} \\ \theta_{2t} \end{bmatrix} / 2 - \begin{bmatrix} \theta_{1t} \\ \theta_{2t} \end{bmatrix}' \begin{bmatrix} \varepsilon_{1t+1} \\ \varepsilon_{2t+1} \end{bmatrix} \\ \begin{bmatrix} \theta_{1t} \\ \theta_{2t} \end{bmatrix} &= \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix} + \begin{bmatrix} \theta_{11}^1 & \theta_{12}^1 \\ \theta_{21}^1 & \theta_{22}^1 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \end{aligned}$$

For this model, the coefficients are (see Appendix 11.6 for a proof)

$$B'_n = B'_{n-1} (\Psi - S\theta^1) + b'_1 \quad (11.13)$$

$$A_n = A_{n-1} + B'_{n-1} [(I - \Psi)\mu - S\theta^0] - B'_{n-1} SS'B_{n-1}/2 + a_1, \quad (11.14)$$

where the recursion starts at $B_0 = 0$ and $A_0 = 0$ (or $B_1 = b_1$ and $A_1 = a_1$).

Example 11.8 (A_n and B_n in the model with affine market price of risk) In the univariate case we have

$$B_0 = 0 \text{ and } A_0 = 0,$$

$$B_1 = b_1 \text{ and } A_1 = a_1$$

$$B_2 = b_1 (1 + \rho - \sigma\theta^1) \text{ and } A_2 = b_1 [(1 - \rho)\mu - \sigma\theta^0] - b_1^2\sigma^2/2 + 2a_1.$$

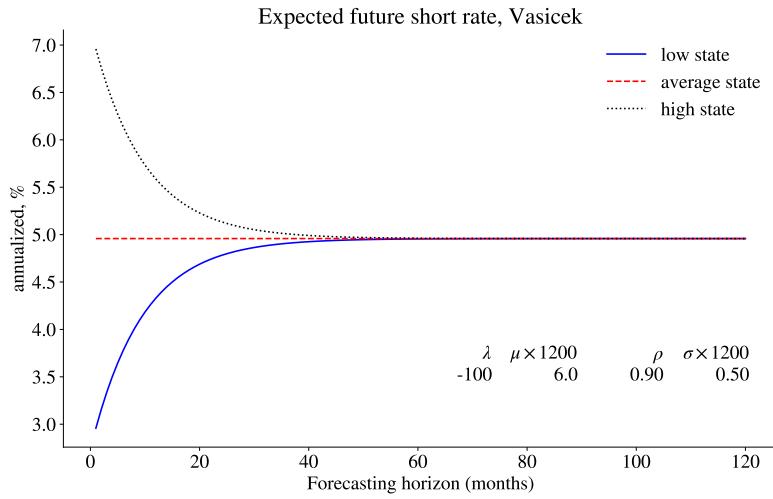


Figure 11.12: Expected future short rate in the Vasicek model

If $\theta^1 = 0$ (so the price of risk is constant), then we have $b_1(1 + \rho)$ which is similar to the Vasicek model. Indeed, with the (free) choices $\theta^0 = \lambda\sigma$, $b_1 = 1$ and $a_1 = -\lambda^2\sigma^2/2$ we get exactly the same expressions as in the Vasicek model (see Example 11.6).

Figure 11.16 illustrates the risk premia in three different one-factor models: the Vasicek model the affine model and the Cox, Ingersoll, and Ross (1985) model. (The CIR modifies the Vasicek model by letting the variance of the shock (ε) be proportional to the state variable (x).) Note that the Vasicek model has constant risk premia across time (although it differs across maturities).

11.4 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.

11.4.1 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

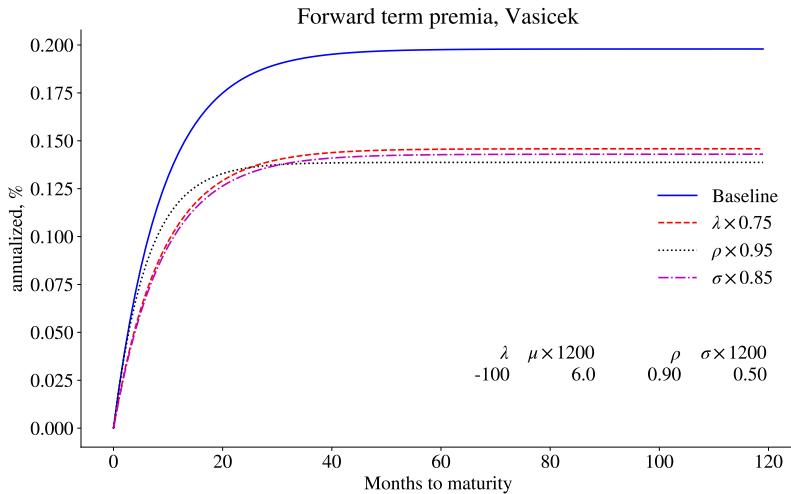


Figure 11.13: Risk premia in the Vasicek model

(times maturity) with errors (u of “unobserved”, although that is something of a misnomer), y_{ut} ,

$$y_{ot} = a_o + b'_o x_t \text{ so } x_t = b'^{-1}_o (y_{ot} - a_o), \text{ and} \quad (11.15)$$

$$y_{ut} = a_u + b'_u x_t + \epsilon_t \quad (11.16)$$

where ϵ_t are the measurement errors. The vector a_o and matrix b_o stack 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. Clearly, the a vectors and b matrices *depend on the parameters of the model*, and need to be recalculated in every iteration of 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 yield curve models themselves do not assume *rational expectations*: we could think of the state dynamics as reflecting what the market participants believed in (and therefore influenced bond prices). However, in the econometrics we estimate this by using the actual dynamics in the historical sample.

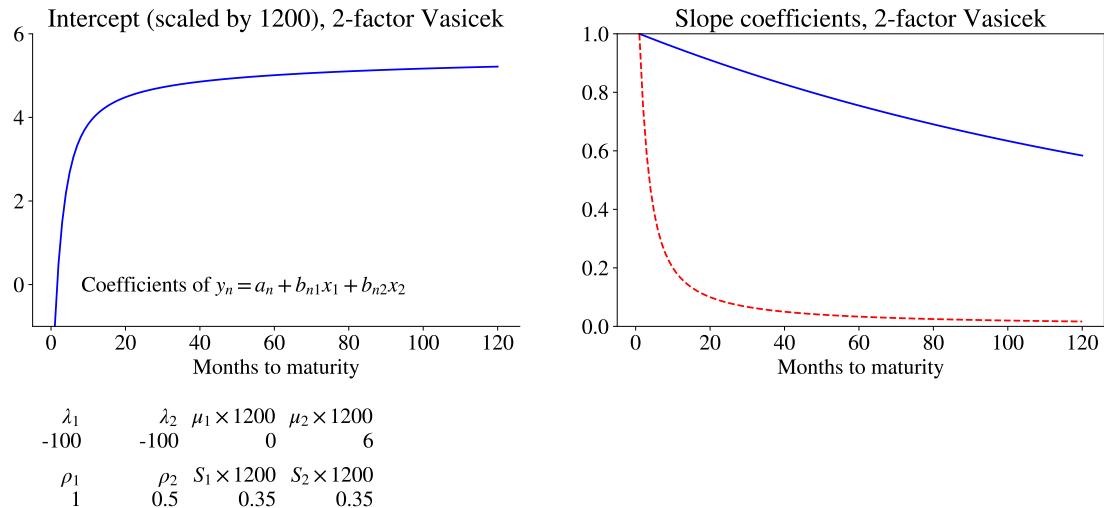


Figure 11.14: a_n and b_n in a two-factor Vasicek model

Remark 11.9 (*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 11.10 (*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 (11.15)–(11.16) 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 11.11 (*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

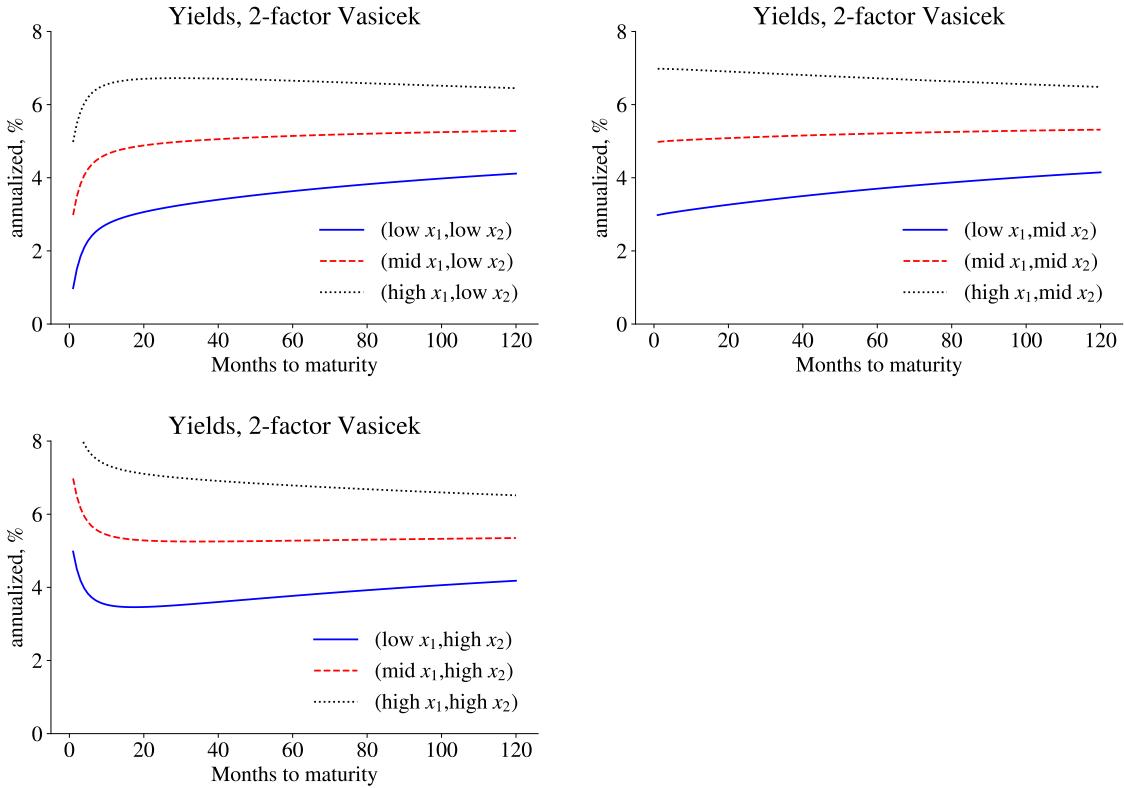


Figure 11.15: Yield curves in a two-factor model, the SDF is a sum of two AR(1)

we want to analyse. 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 11.12 (*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 $P_2^c = cP_1 + (1+c)P_2$, where P_1 and P_2 are the prices of zero coupon bonds. This can be written $c \exp(-A_1 - B'_1 x_t) + (1+c) \exp(-A_2 - B'_2 x_t)$, which is not difficult to handle numerically. 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 11.13 (*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

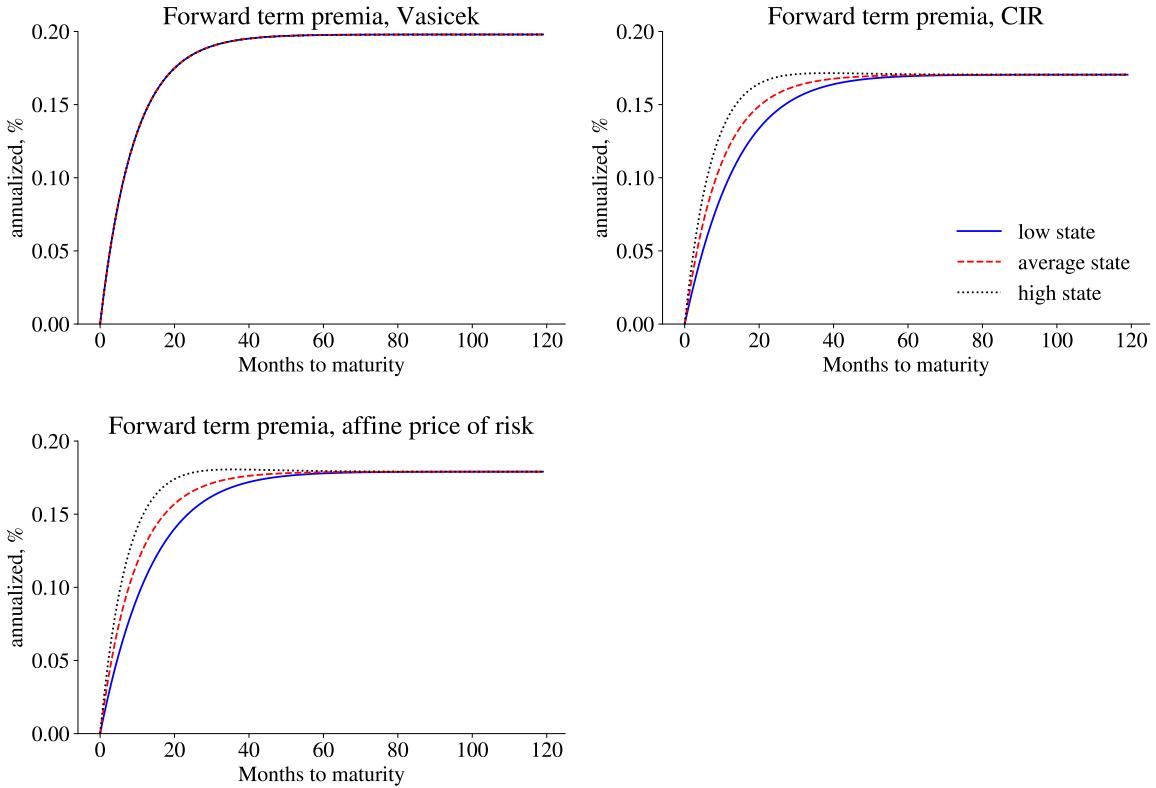


Figure 11.16: Yield curves and risk premia in different models. Same parameters as in Figure 11.11.

not work. Instead, the estimation problem is embedded into a Kalman filter that treats the states are unobservable. In this case, the state dynamics are 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 11.14 (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 11.5.3.*

11.4.2 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 (11.15) we then have

$$\underbrace{\begin{bmatrix} y_{ot} \\ F_t \end{bmatrix}}_{\tilde{y}_{ot}} = \underbrace{\begin{bmatrix} a_o \\ \mathbf{0}_{K_F \times 1} \end{bmatrix}}_{\tilde{a}_0} + \underbrace{\begin{bmatrix} b'_o \\ \begin{bmatrix} \mathbf{0}_{K_F \times K_y} & I_{K_F} \end{bmatrix} \end{bmatrix}}_{\tilde{b}_0} x_t \text{ so } x_t = \tilde{b}'_o^{-1} (\tilde{y}_{ot} - \tilde{a}_0). \quad (11.17)$$

Clearly, the last K_F elements of x_t are identical to F_t .

Example 11.15 (*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 (11.17) 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, $x_{3t} = F_t$.

11.4.3 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 for other maturities are 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

$$\ln \mathcal{L}_o = \sum_{t=1}^T \ln L_{ot}, \text{ with } \ln L_{ot} = \ln \text{pdf}(y_{ot} | y_{o,t-1}). \quad (11.18)$$

Notice that the relation between x_t and y_{ot} in (11.15) 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 $x_t | x_{t-1} \sim N [\mathbb{E}_{t-1} x_t, \text{Cov}_{t-1}(x_t)]$ we have

$$y_{ot} | y_{o,t-1} \sim N \left[\underbrace{a_o + b'_o \mathbb{E}_{t-1} x_t}_{\mathbb{E}_{t-1} y_{ot}}, \underbrace{b'_o \text{Cov}_{t-1}(x_t) b_o}_{\text{Var}_{t-1}(y_{ot})} \right], \text{ with} \quad (11.19)$$

$$x_t = b_o'^{-1} (y_{ot} - a_o).$$

To calculate this expression, we must use the relevant expressions for the conditional mean and covariance, which depends on the dynamics of the state vector in, for instance, (11.8).

See *Figure 11.17* for an illustration.

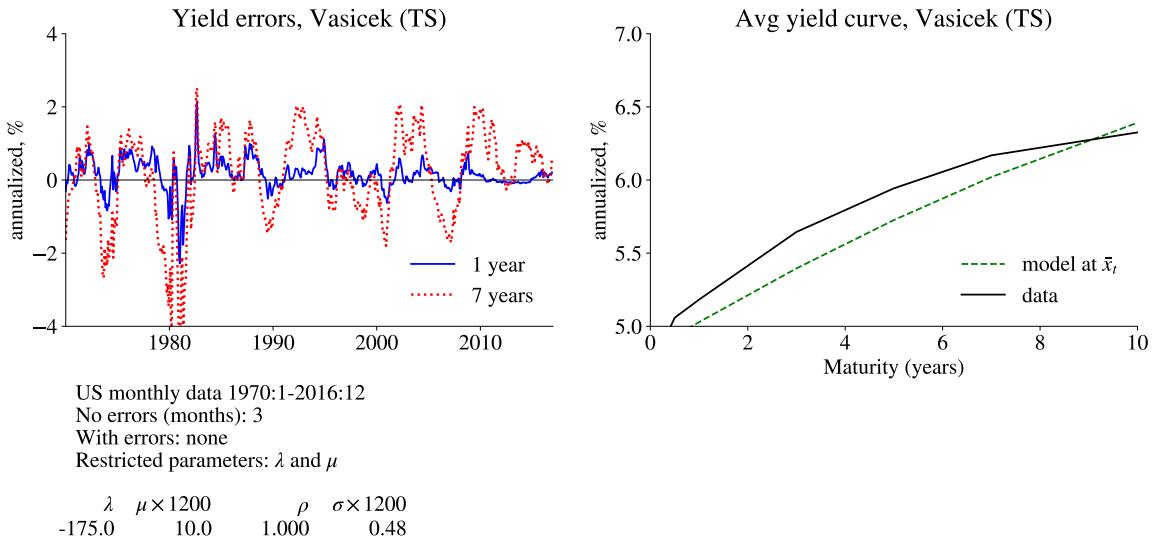


Figure 11.17: Estimation of Vasicek model, time-series approach

Example 11.16 (*Time series estimation of the Vasicek model*) In the one-factor Vasicek model (11.5)–(11.6) 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

$$x_t | x_{t-1} \sim N [(1 - \rho) \mu + \rho x_{t-1}, \sigma^2].$$

Similarly, the distribution of y_{1t} conditional on $y_{1,t-1}$ is

$$y_{1t}|y_{1,t-1} \sim N[a_1 + b_1[(1 - \rho)\mu + \rho x_t], b_1\sigma^2 b_1] \text{ with}$$

$$a_1 = -\lambda^2\sigma^2/2, b_1 = 1, E_{t-1} x_t = (1 - \rho)\mu + \rho x_{t-1}.$$

Inverting the short rate equation (compare with (11.15)) gives

$$x_t = y_{1t} + \lambda^2\sigma^2/2.$$

Combining gives

$$y_{1t}|y_{1,t-1} \sim N[(1 - \rho)(\mu - \lambda^2\sigma^2/2) + \rho y_{1,t-1}, \sigma^2].$$

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.

Example 11.17 (Empirical results from the Vasicek model, time series estimation) Figure 11.17 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 value of λ is imposed (as λ is not separately identified by the data), 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.

11.4.4 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

$$\ln \mathcal{L}_u = \sum_{t=1}^T \ln L_{ut}, \text{ with } \ln L_{ut} = \ln \text{pdf}(y_{ut}|y_{ot}) \quad (11.20)$$

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)

$$y_{ut} | y_{ot} \sim N \left[\underbrace{a_u + b'_u x_t}_{E(y_{ut} | y_{ot})}, \underbrace{\text{diag}(\omega_i^2)}_{\text{Var}(y_{ut} | y_{ot})} \right], \text{ with} \quad (11.21)$$

$$x_t = b_o'^{-1} (y_{ot} - a_o).$$

Under this assumption (normal distribution with a diagonal covariance matrix), maximizing the likelihood function amounts to minimizing the weighted squared errors of the yields

$$\arg \max \ln \mathcal{L}_u = \arg \min \sum_{t=1}^T \sum_{n \in u} \left(\frac{y_{nt} - \hat{y}_{nt}}{\omega_i} \right)^2, \quad (11.22)$$

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 11.18* for an illustration.

Example 11.18 (*Cross-sectional likelihood for the Vasicek model*) In the Vasicek model in Example 11.16, the two-period rate is

$$y_{2t} = (1 - \rho) \mu / 2 + (1 + \rho) x_t / 2 - [\lambda^2 + (1 + \lambda)^2] \sigma^2 / 4.$$

The pdf of y_{2t} , conditional on y_{1t} , is therefore

$$y_{2t} | y_{1t} \sim N(a_2 + b_2 x_t, \omega^2), \text{ with } x_t = y_{1t} + \lambda^2 \sigma^2 / 2, \text{ where}$$

$$b_2 = (1 + \rho) / 2 \text{ and } a_2 = (1 - \rho) \mu / 2 - [\lambda^2 + (1 + \lambda)^2] \sigma^2 / 4.$$

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.

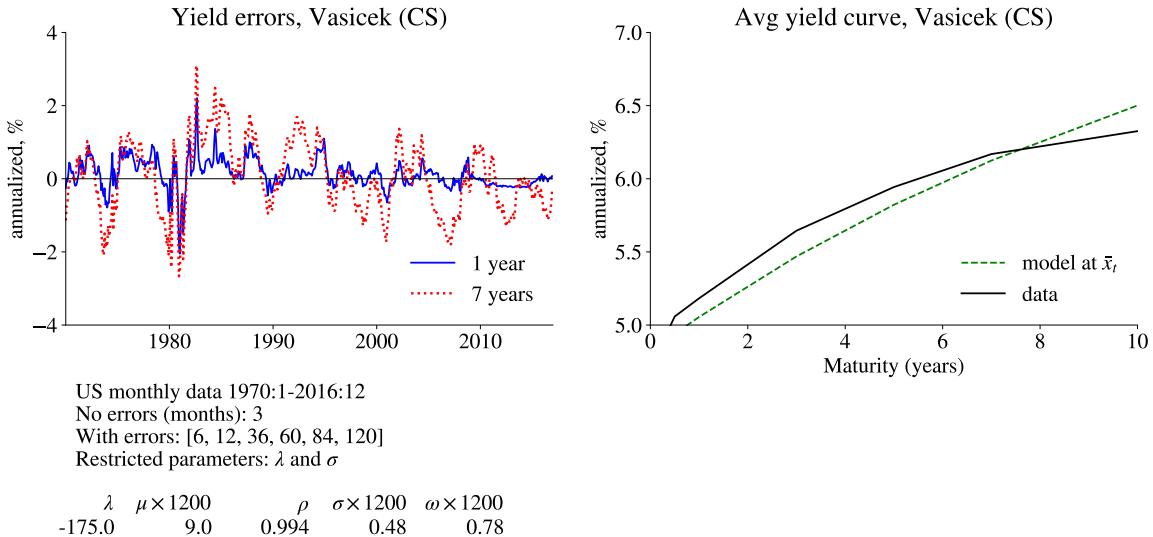


Figure 11.18: Estimation of Vasicek model, cross-sectional approach

Example 11.19 (*Empirical results from the Vasicek model, cross-sectional estimation*)
Figure 11.18 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 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).

11.4.5 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

$$\ln \mathcal{L} = \sum_{t=1}^T \ln L_t, \text{ with } \ln L_t = \ln \text{pdf}(y_{ut}, y_{ot} | y_{o,t-1}). \quad (11.23)$$

Notice that the joint density of (y_{ut}, y_{ot}) , conditional on $y_{o,t-1}$ can be split up as

$$\text{pdf}(y_{ut}, y_{ot} | y_{ot-1}) = \text{pdf}(y_{ut} | y_{ot}) \text{pdf}(y_{ot} | y_{o,t-1}), \quad (11.24)$$

since $y_{o,t-1}$ does not affect how y_{ut} is distributed once we have conditioned on y_{ot} . Taking logs gives

$$\ln L_t = \ln \text{pdf}(y_{ut} | y_{ot}) + \ln \text{pdf}(y_{ot} | y_{o,t-1}). \quad (11.25)$$

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 (11.23) is therefore just the sum of the log likelihoods from the pure cross-sectional and the pure time series estimations

$$\ln \mathcal{L} = \sum_{t=1}^T \ln L_{ut} + \ln L_{ot}. \quad (11.26)$$

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 11.20 (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 11.16 and 11.18.

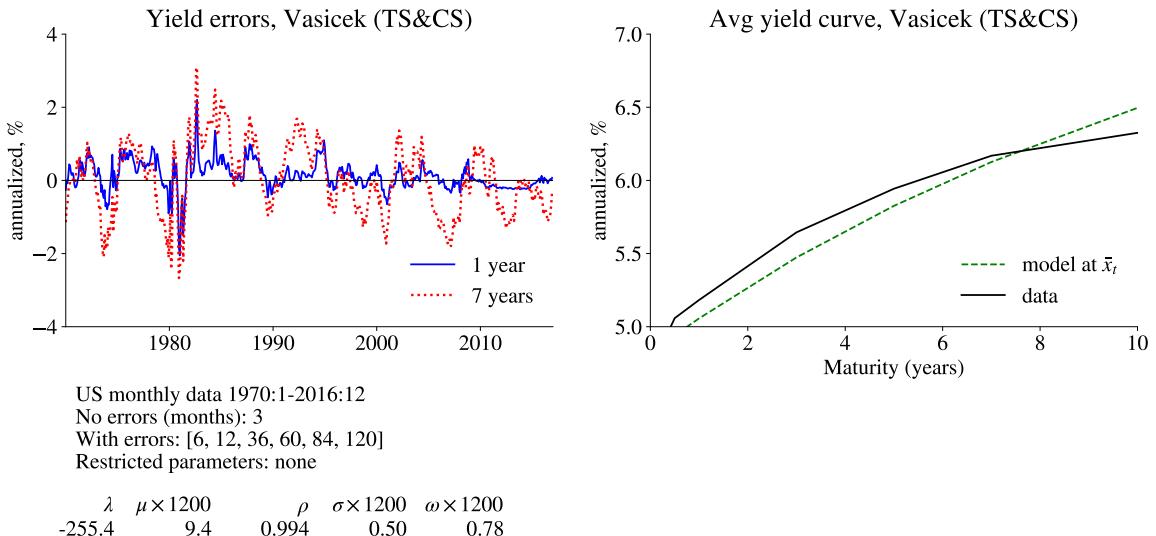


Figure 11.19: Estimation of Vasicek model, combined time series and cross-sectional approach

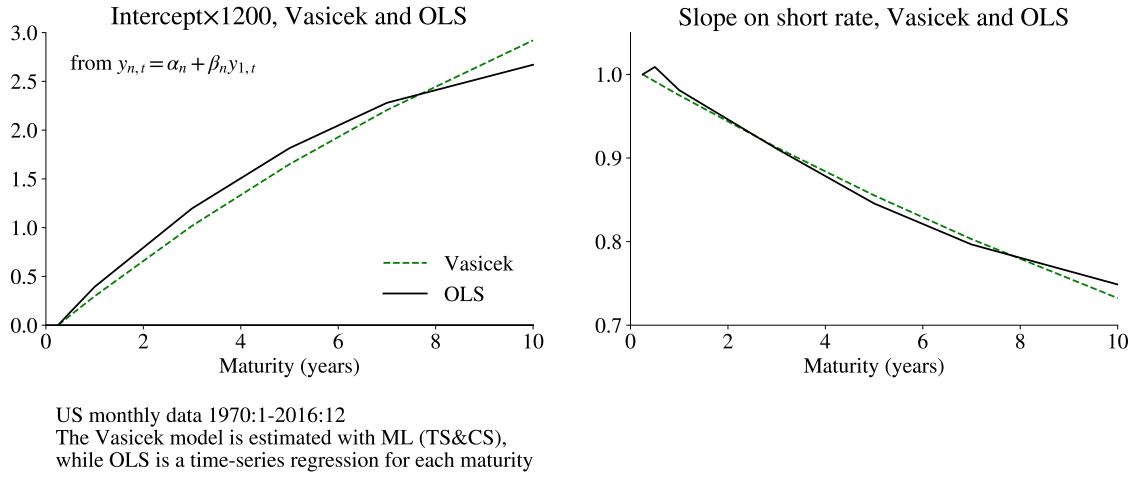


Figure 11.20: Loadings in a one-factor model: LS and Vasicek

Example 11.21 (*Empirical results from the Vasicek and affine models, combined time series and cross-sectional estimation*) Figure 11.19 reports results from a combined time series and cross-sectional estimation of the Vasicek model. All model parameters $(\lambda, \mu, \rho, \sigma^2)$ are estimated, along with the variance of the measurement errors. (All measurement errors are assumed to have the same variances, ω .) Figure 11.20 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) restrictions imposed by the Vasicek model are not at great odds with data. The estimated affine model in Figure 11.21 suggests that the risk premia are higher when short rates are higher, but otherwise the results look similar to the Vasicek model.

Remark 11.22 (*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 11.22 for an example.

Example 11.23 (*Empirical results from a two-factor Vasicek model*) Figure 11.22 reports results from a two-factor Vasicek model. We can only identify the mean of the SDF, not whether if it is due to factor one or two, so I restrict one of the μ_i values. The results indicate that there is one very persistent factor (affecting the yield curve level), and another

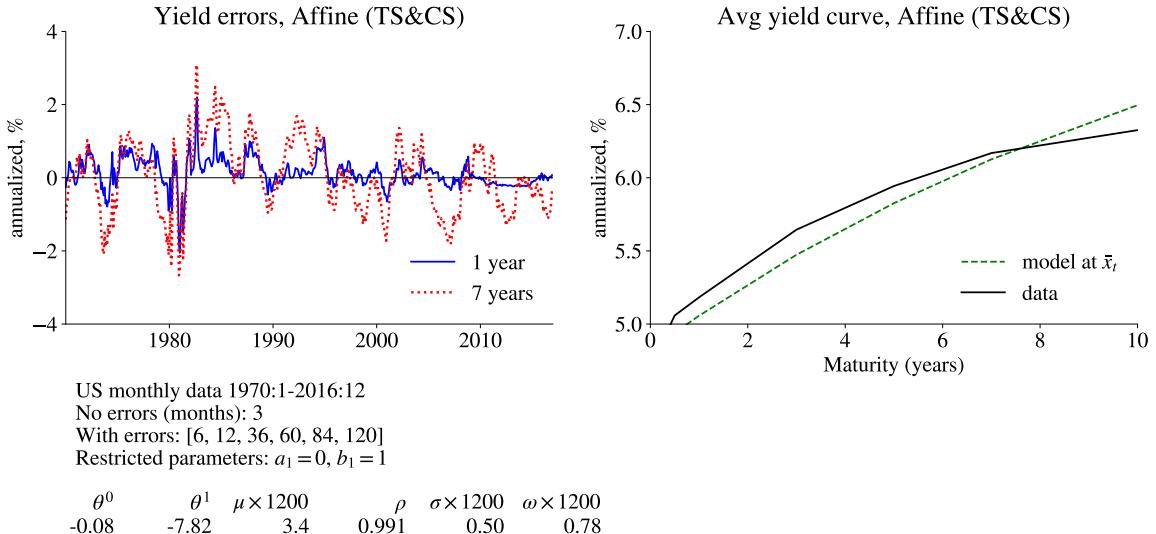


Figure 11.21: Estimation of an affine one-factor model, combined time series and cross-sectional approach

slightly less persistent factor (affecting the yield curve slope). As a practical matter, it turned out that a derivative-free method worked much better than standard optimization routines. The pricing errors are clearly smaller than in a one-factor Vasicek model, but still considerable in the early 1980s.

Example 11.24 (Empirical results from a two-factor affine model) Figure 11.23 reports results from a two-factor affine model. The pricing errors are fairly similar to the two-factor Vasicek model.

11.5 Summary of Some Empirical Findings

11.5.1 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

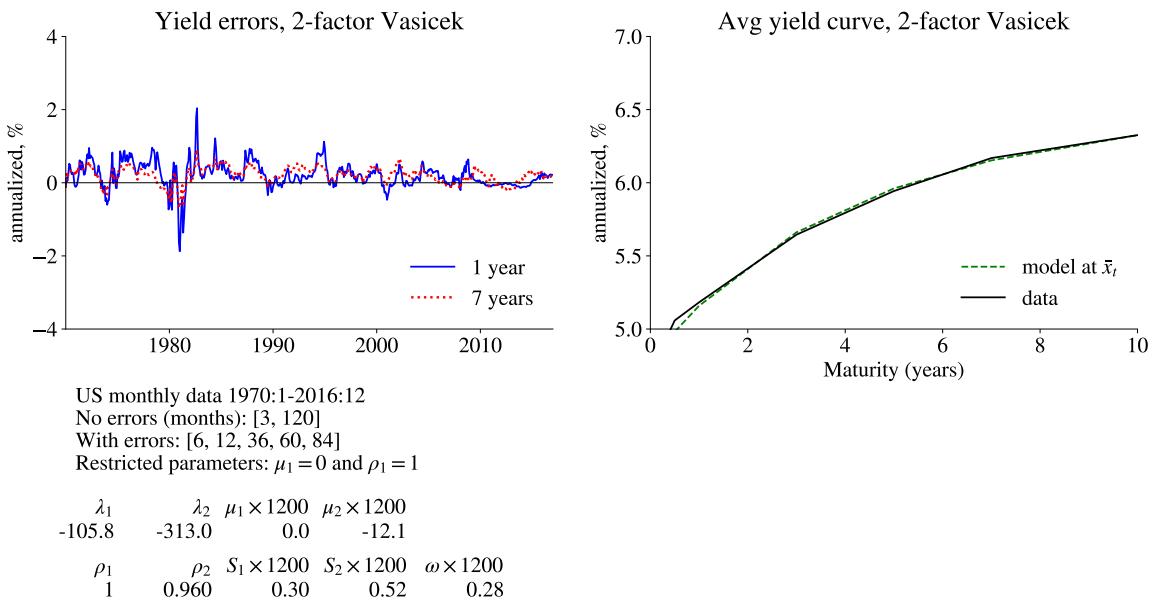


Figure 11.22: Estimation of 2-factor Vasicek model, time-series&cross-section approach

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.)

11.5.2 “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,

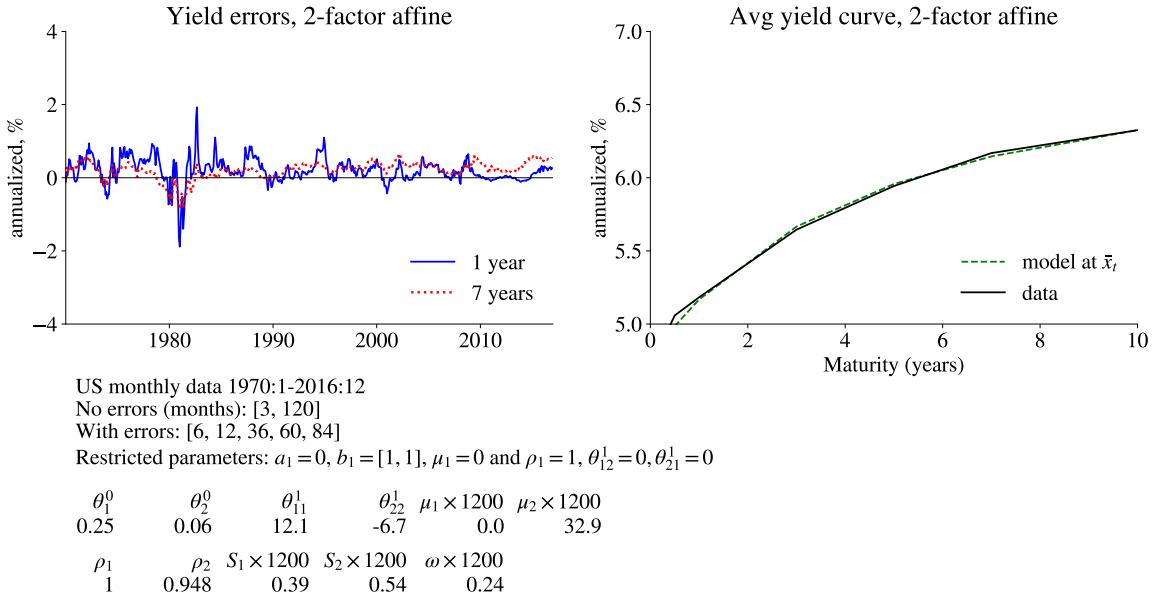


Figure 11.23: Estimation of 2-factor affine model

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.

11.5.3 “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}

$$y_{1,t+1} - y_{1t} = \alpha + \beta y_{1t} + \varepsilon_{t+1}, \text{ where} \quad (11.27)$$

$$E_t \varepsilon_{t+1} = 0 \text{ and } E_t \varepsilon_{t+1}^2 = \text{Var}_t(\varepsilon_{t+1}) = \sigma^2 y_{1t}^{2\gamma}.$$

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

$$dr_t = (\beta_0 + \beta_1 r_t)dt + \sigma r_t^\gamma dW_t, \quad (11.28)$$

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 (11.27) with GMM using the following moment conditions

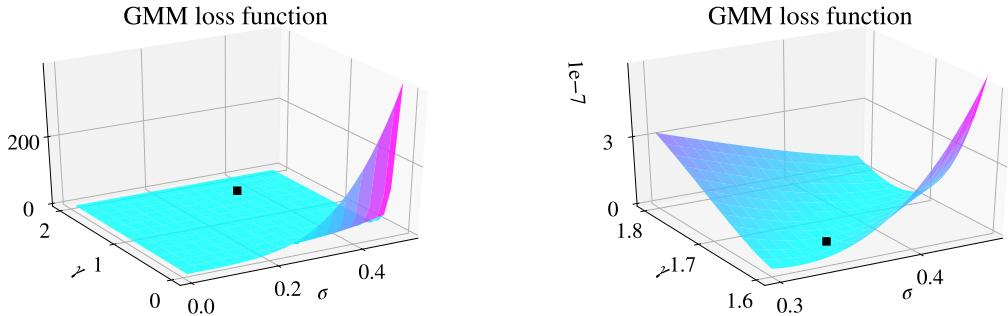
$$g_t(\alpha, \beta, \gamma, \sigma^2) = \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1}^2 - \sigma^2 y_{1t}^{2\gamma} \end{bmatrix} \otimes \begin{bmatrix} 1 \\ y_{1t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1} y_{1t} \\ \varepsilon_{t+1}^2 - \sigma^2 y_{1t}^{2\gamma} \\ (\varepsilon_{t+1}^2 - \sigma^2 y_{1t}^{2\gamma}) y_{1t} \end{bmatrix}, \quad (11.29)$$

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 [Figure 11.24](#) for an illustration.

Example 11.25 (Re-estimating the Chan et al model) Some results obtained from re-estimating the model on a longer data set are found in [Figure 11.24](#). In this figure, $\alpha = \beta = 0$ is imposed, but the results are very similar if this is relaxed. One of the first things 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 (11.29) are very strongly correlated: moment 1 and 2 have a very strong



Monthly federal funds rates 1954:07-2016:12
 Point estimates of γ and σ : 1.65 and 0.36

Correlations of moment conditions:

	1	2	3	4
1	1.00	0.93	-0.30	-0.34
2	0.93	1.00	-0.42	-0.47
3	-0.30	-0.42	1.00	0.99
4	-0.34	-0.47	0.99	1.00

Figure 11.24: Federal funds rate, monthly data, $\alpha = \beta = 0$ imposed

correlation, and 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).

11.6 Appendix: Details on Yield Curve Models

Proof. (of (11.9)–(11.10) in the one-factor case) First, use the dynamics to write

$$\begin{aligned}
 m_{t+1} + p_{n-1,t+1} &= \underbrace{-x_t - \lambda\sigma\varepsilon_{t+1}}_{m_{t+1}} - \underbrace{A_{n-1} - B_{n-1}x_{t+1}}_{p_{n-1,t+1}} \\
 &= -(1 + B_{n-1}\rho)x_t - (\lambda + B_{n-1})\sigma\varepsilon_{t+1} - A_{n-1} - B_{n-1}(1 - \rho)\mu,
 \end{aligned}$$

where the second line follows from using the time series process of x_t (see (11.6)). (We clearly don't need to transpose B_{n-1} since it is a scalar.) The conditional moments are

$$\begin{aligned}\mathrm{E}_t(m_{t+1} + p_{n-1,t+1}) &= -(1 + B_{n-1}\rho)x_t - A_{n-1} - B_{n-1}(1 - \rho)\mu \\ \mathrm{Var}_t(m_{t+1} + p_{n-1,t+1}) &= (\lambda + B_{n-1})^2\sigma^2.\end{aligned}$$

Second, use in

$$p_{nt} = \mathrm{E}_t(m_{t+1} + p_{n-1,t+1}) + \mathrm{Var}_t(m_{t+1} + p_{n-1,t+1})/2,$$

to get

$$-A_n - B_n x_t = -(1 + B_{n-1}\rho)x_t - A_{n-1} - B_{n-1}(1 - \rho)\mu + (\lambda + B_{n-1})^2\sigma^2/2$$

This equation must always hold (for any x_t), so matching coefficients gives (11.9)–(11.10). ■

Proof. (of (11.13)–(11.14)) First,

$$m_{t+1} + p_{n-1,t+1} = -y_{1t} - \theta'_t\theta_t/2 - \theta'_t\varepsilon_{t+1} - A_{n-1} - B'_{n-1}x_{t+1}$$

The conditional moments are

$$\begin{aligned}\mathrm{E}_t[m_{t+1} + p_{n-1,t+1}] &= -y_{1t} - \theta'_t\theta_t/2 - A_{n-1} - B'_{n-1}(I - \Psi)\mu - B'_{n-1}\Psi x_t \\ \mathrm{Var}_t[m_{t+1} + p_{n-1,t+1}] &= (\theta'_t + B'_{n-1}S)(\theta'_t + B'_{n-1}S)' \\ &= (\theta'_t + B'_{n-1}S)(\theta_t + S'B_{n-1}) \\ &= \theta'_t\theta_t + \theta'_t S' B_{n-1} + B'_{n-1} S \theta_t + B'_{n-1} S S' B_{n-1}\end{aligned}$$

Second,

$$\begin{aligned}-A_n - B'_n x_t &= [-y_{1t} - \theta'_t\theta_t/2 - A_{n-1} - B'_{n-1}(I - \Psi)\mu - B'_{n-1}\Psi x_t] + \\ &\quad [\theta'_t\theta_t + \theta'_t S' B_{n-1} + B'_{n-1} S \theta_t + B'_{n-1} S S' B_{n-1}]/2 \\ &= (-a_1 - b'_1 x_t) - A_{n-1} - B'_{n-1}(I - \Psi)\mu - B'_{n-1}\Psi x_t + \\ &\quad B'_{n-1} S (\theta^0 + \theta^1 x_t) + B'_{n-1} S S' B_{n-1}/2 \\ &= -a_1 - A_{n-1} + B'_{n-1} [S\theta^0 - (I - \Psi)\mu] + B'_{n-1} S S' B_{n-1}/2 - \\ &\quad b'_1 x_t - B'_{n-1} \Psi x_t + B'_{n-1} S \theta^1 x_t\end{aligned}$$

This equation must always hold (for any x_t), so matching coefficients gives (11.13)–

(11.14). ■

Chapter 12

Yield Curve Models: Nonparametric Estimation

12.1 Nonparametric Regression

12.1.1 “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

$$r_{t+1} = \alpha + \rho r_t + \varepsilon_{t+1}, \text{ where} \\ E_t \varepsilon_{t+1} = 0 \text{ and } \text{Var}_t(\varepsilon_{t+1}^2) = \sigma^2 r_t^{2\gamma}. \quad (12.1)$$

nests several well-known models. Subtract r_{t-1} from both sides to get

$$r_{t+1} - r_t = \alpha + \beta r_t + \varepsilon_{t+1}, \quad (12.2)$$

where $\beta = \rho - 1 \leq 0$. This is an approximation of the diffusion process

$$dr_t = (\beta_0 + \beta_1 r_t)dt + \sigma r_t^\gamma dW_t, \quad (12.3)$$

where W_t is a Wiener process. Recall that affine one-factor models require $\gamma = 0$ (Vasicek (1977)) or $\gamma = 0.5$ (Cox, Ingersoll, and Ross (1985)). Notice that this linear model implies that $E_t(r_{t+1} - r_t) = \alpha + \beta r_t$ is a downward sloping line as a function of r_t (flat if no mean reversion, $\rho = 1$ so $\beta = 0$). The model also implies that the conditional variance, $\sigma^2 r_t^{2\gamma}$, is flat if Vasicek and linear (upward sloping) if CIR.

This paper tests several models of the short interest rate by using a nonparametric technique.

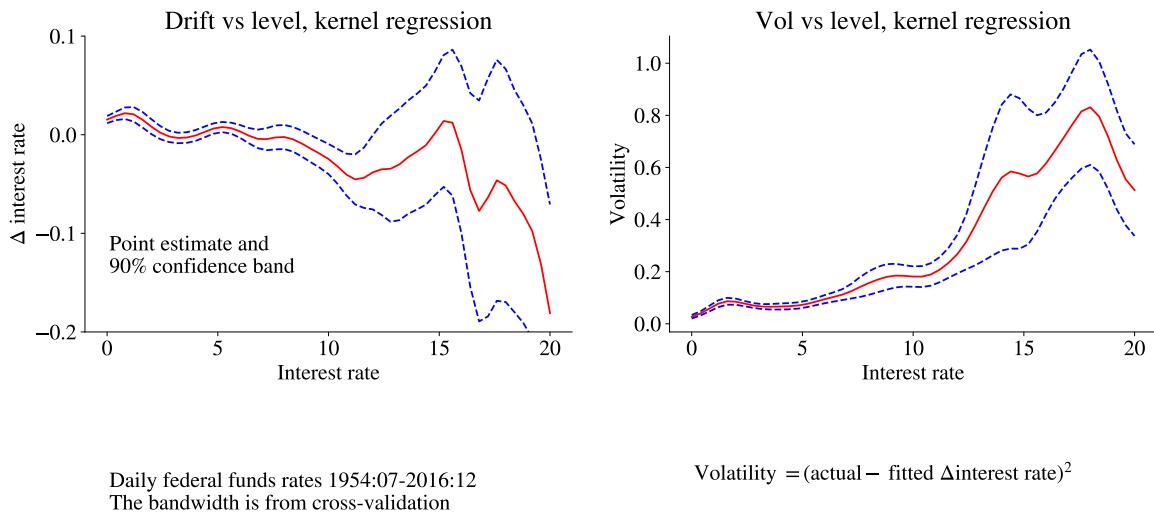


Figure 12.1: Kernel regression, confidence band

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

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{T} \sum_{t=1}^T [\pi(\theta, r_t) - \hat{\pi}_0(r)]^2. \quad (12.4)$$

3. The model is tested by using a scaled version of the minimized value of the right hand side of (12.4) 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 12.1 *The fairly non-linear mean reversion in Figure 12.1* seems to be the key reason for why [Ait-Sahalia \(1996\)](#) rejects most short rate models.

12.2 Approximating Non-Linear Regression Functions

12.2.1 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 coefficient at 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

$$y_t = b(x_t) + \varepsilon_t. \quad (12.5)$$

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

$$\begin{bmatrix} x_t \\ \max(x_t - \xi_1, 0) \\ \max(x_t - \xi_2, 0) \end{bmatrix} \quad (12.6)$$

and approximate as

$$b(x_t) = \beta_1 x_t + \beta_2 \max(x_t - \xi_1, 0) + \beta_3 \max(x_t - \xi_2, 0). \quad (12.7)$$

This can also be written

$$b(x_t) = \begin{bmatrix} \beta_1 x_t & \text{if } x_t < \xi_1 \\ \beta_1 x_t + \beta_2 (x_t - \xi_1) & \text{if } \xi_1 \leq x_t < \xi_2 \\ \beta_1 x_t + \beta_2 (x_t - \xi_1) + \beta_3 (x_t - \xi_2) & \text{if } \xi_2 \leq x_t \end{bmatrix}. \quad (12.8)$$

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

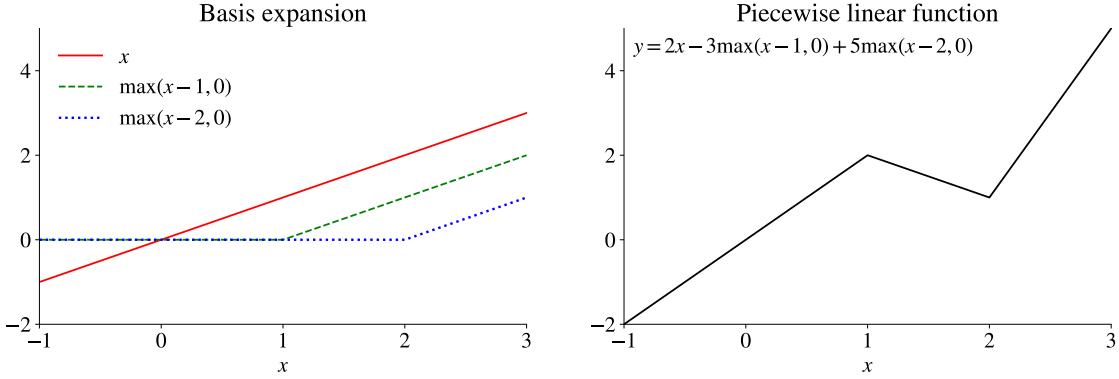


Figure 12.2: Example of piecewise linear function, created by basis expansion

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 we impose restrictions on the slopes and intercepts). Figure 12.2 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” (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 12.2 (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 (12.6)).*

With point estimates and their covariance matrix, we could, for instance, 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 continuous 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

$$\frac{\partial b(x_t)}{\partial \theta} = \begin{bmatrix} \partial b(x_t) / \partial \xi_1 \\ \partial b(x_t) / \partial \xi_2 \\ \partial b(x_t) / \partial \beta_1 \\ \partial b(x_t) / \partial \beta_2 \\ \partial b(x_t) / \partial \beta_3 \end{bmatrix} = \begin{bmatrix} -\beta_2 \delta(x_t - \xi_1 \geq 0) \\ -\beta_3 \delta(x_t - \xi_2 \geq 0) \\ x_t \\ \max(x_t - \xi_1, 0) \\ \max(x_t - \xi_2, 0) \end{bmatrix}, \quad (12.9)$$

where $\delta(q) = 1$ if q is true and 0 otherwise. The variance of $\hat{b}(x_t)$ is then

$$\text{Var}[\hat{b}(x_t)] = \frac{\partial b(x_t)}{\partial \theta'} V \frac{\partial b(x_t)}{\partial \theta}, \quad (12.10)$$

where V is the covariance of the parameters collected in the vector θ (here $\xi_1, \xi_2, \beta_1, \beta_2, \beta_3$).

Remark 12.3 (*The derivatives of $b(x_t)$*) From (12.8) we have the following derivatives

$$\begin{bmatrix} \partial b(x_t) / \partial \xi_1 \\ \partial b(x_t) / \partial \xi_2 \\ \partial b(x_t) / \partial \beta_1 \\ \partial b(x_t) / \partial \beta_2 \\ \partial b(x_t) / \partial \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ x_t \\ 0 \\ 0 \end{bmatrix} \text{ if } x_t < \xi_1, \quad \begin{bmatrix} -\beta_2 \\ 0 \\ x_t \\ x_t - \xi_1 \\ 0 \end{bmatrix} \text{ if } \xi_1 \leq x_t < \xi_2, \quad \begin{bmatrix} -\beta_2 \\ -\beta_3 \\ x_t \\ x_t - \xi_1 \\ x_t - \xi_2 \end{bmatrix} \text{ if } \xi_2 \leq x_t.$$

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)

$$b(x_t, z_t) = b_x(x_t) + b_z(z_t), \quad (12.11)$$

where both $b_x(x_t)$ and $b_z(z_t)$ are piecewise functions of the sort discussed in (12.8). 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 (12.10). See Figure 12.3 for an illustration.

12.3 Appendix: Partial Linear Model

A possible way out of the curse of dimensionality of the multivariate kernel regression is to specify a partially linear model

$$y_t = z'_t \beta + b(x_t) + \varepsilon_t, \quad (12.12)$$

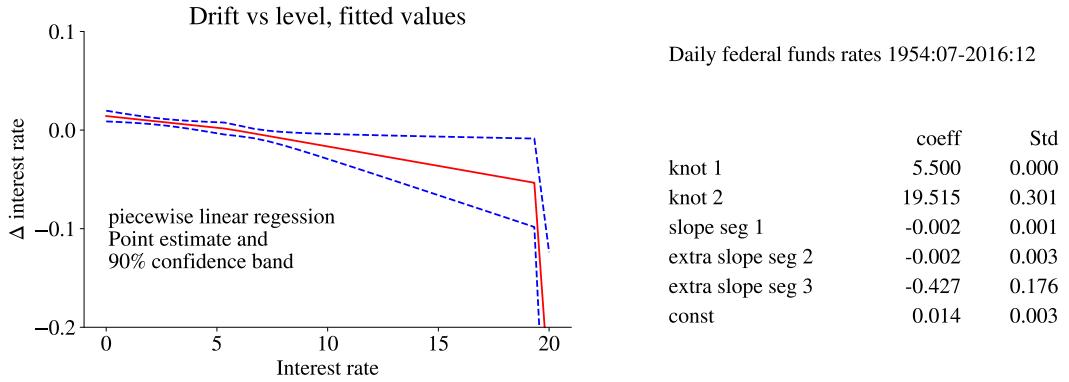


Figure 12.3: Federal funds rate, piecewise linear model

where ε_t is uncorrelated over time and where $E \varepsilon_t = 0$ and $E(\varepsilon_t | x_t, z_t) = 0$. 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 (12.12) conditional on x_t

$$E(y_t | x_t) = E(z_t | x_t)' \beta + b(x_t). \quad (12.13)$$

Subtract from (12.12) to get

$$y_t - E(y_t | x_t) = [z_t - E(z_t | x_t)]' \beta + \varepsilon_t. \quad (12.14)$$

The *double residual method* (see Pagan and Ullah (1999) 5.2) has several steps. *First*, estimate $E(y_t | x_t)$ by a kernel regression of y_t on x_t ($\hat{b}_y(x)$), and $E(z_t | x_t)$ by a similar kernel regression of z_t on x_t ($\hat{b}_z(x)$). *Second*, use these estimates in (12.14)

$$y_t - \hat{b}_y(x_t) = [z_t - \hat{b}_z(x_t)]' \beta + \varepsilon_t \quad (12.15)$$

and estimate β by least squares. *Third*, use these estimates in (12.13) to estimate $b(x_t)$ as

$$\hat{b}(x_t) = \hat{b}_y(x_t) - \hat{b}_z(x_t)' \hat{\beta}. \quad (12.16)$$

It can be shown that (under the assumption that y_t, z_t and x_t are iid)

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N[0, \text{Var}(\varepsilon_t) \text{Cov}(z_t | x_t)^{-1}]. \quad (12.17)$$

We can consistently estimate $\text{Var}(\varepsilon_t)$ by the sample variance of the fitted residuals in (12.12)—plugging in the estimated β and $b(x_t)$: and we can also consistently estimate

$\text{Cov}(z_t | x_t)$ 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).

Bibliography

- Ait-Sahalia, Y., 1996, “Testing continuous-time models of the spot interest rate,” *Review of Financial Studies*, 9, 385–426.
- Ait-Sahalia, Y., and A. W. Lo, 1998, “Nonparametric estimation of state-price densities implicit in financial asset prices,” *Journal of Finance*, 53, 499–547.
- Alexander, C., 2008, *Market Risk Analysis: Practical Financial Econometrics*, Wiley.
- Amemiya, T., 1985, *Advanced econometrics*, Harvard University Press, Cambridge, Massachusetts.
- Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold, 2005, “Volatility forecasting,” Working Paper 11188, NBER.
- Andrews, D. W. K., and J. C. Monahan, 1992, “An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator,” *Econometrica*, 60, 953–966.
- Ang, A., and J. Chen, 2002, “Asymmetric correlations of equity portfolios,” *Journal of Financial Economics*, 63, 443–494.
- Ang, 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.
- Back, K. E., 2010, *Asset Pricing and Portfolio Choice Theory*, Oxford University Press, Oxford.
- Backus, D., S. Foresi, and C. Telmer, 1998, “Discrete-time models of bond pricing,” Working Paper 6736, NBER.
- Baltagi, D. H., 2008, *Econometric Analysis of Panel Data*, Wiley, 4th edn.

- 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.
- Bansal, R., and A. Yaron, 2004, “Risks for the long run: a potential resolution of asset pricing puzzles,” *The Journal of Finance*, 59, 1481–1509.
- Bekaert, G., and M. S. Urias, 1996, “Diversification, integration and emerging market closed-end funds,” *Journal of Finance*, 51, 835–869.
- Berkowitz, J., and L. Kilian, 2000, “Recent developments in bootstrapping time series,” *Econometric-Reviews*, 19, 1–48.
- Bossaert, P., 2002, *The paradox of asset pricing*, Princeton University Press.
- Brandimarte, P., 2006, *Numerical Methods in Finance and Economics*, Wiley, Hoboken, NJ.
- Breeden, D., and R. Litzenberger, 1978, “Prices of State-Contingent Claims Implicit in Option Prices,” *Journal of Business*, 51, 621–651.
- Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger, 1989, “Empirical tests of the consumption-oriented CAPM,” *Journal of Finance*, 44, 231–262.
- Britten-Jones, M., and A. Neuberger, 2000, “Option prices, implied price processes, and stochastic volatility,” *Journal of Finance*, 55, 839–866.
- Brockwell, P. J., and R. A. Davis, 1991, *Time series: theory and methods*, Springer Verlag, New York, second edn.
- 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., 1993, “Intertemporal asset pricing without consumption data,” *American Economic Review*, 83, 487–512.
- 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.
- 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 R. J. Shiller, 1988, “The dividend-price ratio and expectations of future dividends and discount factors,” *Review of Financial Studies*, 1, 195–227.
- 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.
- Carhart, M., 1997, “On persistence in mutual fund performance,” *Journal of Finance*, 52, 57–82.
- 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.
- Chapman, D., and N. D. Pearson, 2000, “Is the short rate drift actually nonlinear?,” *Journal of Finance*, 55, 355–388.
- 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, 2011, “The time-varying systematic risk of carry trade strategies,” *Journal of Financial and Quantitative Analysis*, 46, 1107–1125.
- Clark, T. E., and M. W. McCracken, 2001, “Tests of equal forecast accuracy and encompassing for nested models,” *Journal of Econometrics*, 105, 85–110.
- Clark, T. E., and K. D. West, 2007, “Approximately normal tests for equal predictive accuracy in nested models,” *Journal of Ec*, 138, 291–311.

- Cochrane, J. H., 2001, *Asset pricing*, Princeton University Press, Princeton, New Jersey.
- Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.
- Cochrane, J. H., and M. Piazzesi, 2005, “Bond risk premia,” *American Economic Review*, 95, 138–160.
- Constantinides, G. M., and D. Duffie, 1996, “Asset pricing with heterogeneous consumers,” *The Journal of Political Economy*, 104, 219–240.
- Cox, J. C., J. E. Ingersoll, and S. A. Ross, 1985, “A theory of the term structure of interest rates,” *Econometrica*, 53, 385–407.
- Cox, J. C., and S. A. Ross, 1976, “The Valuation of Options for Alternative Stochastic Processes,” *Journal of Financial Economics*, 3, 145–166.
- Dahlquist, M., 1996, “On alternative interest rate processes,” *Journal of Banking and Finance*, 20, 1093–1119.
- Dahlquist, M., J. V. Martinez, and P. Söderlind, 2016, “Individual Investor Activity and Performance,” forthcoming in *The Review of Financial Studies*.
- Davidson, R., and J. G. MacKinnon, 1993, *Estimation and inference in econometrics*, Oxford University Press, Oxford.
- Davison, A. C., and D. V. Hinkley, 1997, *Bootstrap methods and their applications*, Cambridge University Press.
- de Jong, F., 2000, “Time series and cross-section information in affine term-structure models,” *Journal of Business and Economic Statistics*, 18, 300–314.
- DeGroot, M. H., 1986, *Probability and statistics*, Addison-Wesley, Reading, Massachusetts.
- DeMiguel, V., L. Garlappi, and R. Uppal, 2009, “Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?,” *Review of Financial Studies*, 22, 1915–1953.
- Diebold, F. X., 2001, *Elements of forecasting*, South-Western, 2nd edn.

- Diebold, F. X., and C. Li, 2006, “Forecasting the term structure of government yields,” *Journal of Econometrics*, 130, 337–364.
- Diebold, F. X., and R. S. Mariano, 1995, “Comparing predictive accuracy,” *Journal of Business and Economic Statistics*, 13, 253–265.
- Driscoll, J., and A. Kraay, 1998, “Consistent covariance matrix estimation with spatially dependent panel data,” *Review of Economics and Statistics*, 80, 549–560.
- Duan, J., 1995, “The GARCH option pricing model,” *Mathematical Finance*, 5, 13–32.
- Duffee, G. R., 2002, “Term premia and interest rate forecasts in affine models,” *Journal of Finance*, 57, 405–443.
- Duffee, G. R., 2005, “Time variation in the covariance between stock returns and consumption growth,” *Journal of Finance*, 60, 1673–1712.
- Efron, B., T. Hasti, I. Johnstone, and R. Tibshirani, 2004, “Least angle regression,” *The Annals of Statistics*, 32, 407–499.
- Efron, B., and R. J. Tibshirani, 1993, *An introduction to the bootstrap*, Chapman and Hall, New York.
- Elliot, G., and A. Timmermann, 2016, *Economic forecasting*, Princeton University Press, Princeton, New Jersey.
- 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, 1989, “Substitution, risk aversion, and the temporal behavior of asset returns: a theoretical framework,” *Econometrica*, 57, 937–969.
- 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., and J. MacBeth, 1973, “Risk, return, and equilibrium: empirical tests,” *Journal of Political Economy*, 71, 607–636.
- 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.
- 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., S. Sarkissian, and T. T. Simin, 2003, “Spurious regressions in financial economics,” *Journal of Finance*, 57, 1393–1413.
- 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.
- Froot, K. A., 1989, “New Hope for the Expectations Hypothesis of the Term Structure of Interest Rates,” *The Journal of Finance*, 44, 283–304.
- Gibbons, M., S. Ross, and J. Shanken, 1989, “A test of the efficiency of a given portfolio,” *Econometrica*, 57, 1121–1152.
- Glosten, L. R., R. Jagannathan, and D. Runkle, 1993, “On the relation between the expected value and the volatility of the nominal excess return on stocks,” *Journal of Finance*, 48, 1779–1801.
- Gourieroux, C., and J. Jasiak, 2001, *Financial econometrics: problems, models, and methods*, Princeton University Press.
- Goyal, A., and I. Welch, 2008, “A comprehensive look at the empirical performance of equity premium prediction,” *Review of Financial Studies* 2008, 21, 1455–1508.
- Greene, W. H., 2000, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

- Greene, W. H., 2003, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 5th edn.
- Greene, W. H., 2012, *Econometric analysis*, Pearson Education Ltd, Harlow, Essex, 7th edn.
- Hamilton, J. D., 1994, *Time series analysis*, Princeton University Press, Princeton.
- Hansen, L. P., and R. Jagannathan, 1991, “Implications of security market data for models of dynamic economies,” *Journal of Political Economy*, 99, 225–262.
- Härdle, W., 1990, *Applied nonparametric regression*, Cambridge University Press, Cambridge.
- Harvey, A. C., 1989, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.
- Hastie, T., R. Tibshirani, and J. Friedman, 2001, *The elements of statistical learning: data mining, inference and prediction*, Springer Verlag.
- Hentschel, L., 1995, “All in the family: nesting symmetric and asymmetric GARCH models,” *Journal of Financial Economics*, 39, 71–104.
- Heston, S. L., and S. Nandi, 2000, “A closed-form GARCH option valuation model,” *Review of Financial Studies*, 13, 585–625.
- Hoechle, D., 2007, “Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence,” *The Stata Journal*, 7, 281–312.
- Hoechle, D., M. M. Schmid, and H. Zimmermann, 2015, “Decomposing Performance,” Working paper, University of St. Gallen.
- Holm, S., 1979, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, 6, 65–70.
- Hördahl, P., O. Tristiani, and D. Vestin, 2006, “A joint econometric model of macroeconomic and term structure dynamics,” *Journal of Econometrics*, 131, 405–444.
- Horowitz, J. L., 2001, “The Bootstrap,” in J.J. Heckman, and E. Leamer (ed.), *Handbook of Econometrics* . , vol. 5, Elsevier.

- Huang, C.-F., and R. H. Litzenberger, 1988, *Foundations for financial economics*, Elsevier Science Publishing, New York.
- Jackwerth, J. C., 2000, “Recovering risk aversion from option prices and realized returns,” *Review of Financial Studies*, 13, 433–451.
- Jagannathan, R., and Z. Wang, 1996, “The conditional CAPM and the cross-section of expected 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.
- Jiang, G. J., and Y. S. Tian, 2005, “The model-free implied volatility and its information content,” *Review of Financial Studies*, 18, 1305–1342.
- Jondeau, E., S.-H. Poon, and M. Rockinger, 2007, *Financial Modeling under Non-Gaussian Distributions*, Springer.
- Karnaukh, N., A. Ranaldo, and P. Söderlind, 2015, “Understanding FX liquidity,” *The Review of Financial Studies*, 28, 3073–3108.
- Leitch, G., and J. E. Tanner, 1991, “Economic forecast evaluation: profit versus the conventional error measures,” *American Economic Review*, 81, 580–590.
- 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.
- Lo, A. W., and A. C. MacKinlay, 1990, “When are contrarian profits due to stock market overreaction?,” *Review of Financial Studies*, 3, 175–208.
- 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.

- Lustig, H. N., N. L. Roussanov, and A. Verdelhan, 2011, “Common risk factors in currency markets,” *Review of Financial Studies*, 24, 3731–3777.
- MacKinlay, C., 1995, “Multifactor models do not explain deviations from the CAPM,” *Journal of Financial Economics*, 38, 3–28.
- Mankiw, G. N., 1986, “The equity premium and the concentration of aggregate shocks,” *Journal of Financial Economics*, 17, 211–219.
- McNeil, A. J., R. Frey, and P. Embrechts, 2005, *Quantitative risk management*, Princeton University Press.
- Mehra, R., and E. Prescott, 1985, “The equity premium: a puzzle,” *Journal of Monetary Economics*, 15, 145–161.
- 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.
- Mittelhammer, R. C., 1996, *Mathematical statistics for economics and business*, Springer-Verlag, New York.
- Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric foundations*, Cambridge University Press, Cambridge.
- Nelson, D. B., 1991, “Conditional heteroskedasticity in asset returns,” *Econometrica*, 59, 347–370.
- Newey, W. K., and K. D. West, 1987, “A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix,” *Econometrica*, 55, 703–708.
- Pagan, A., and A. Ullah, 1999, *Nonparametric econometrics*, Cambridge University Press.
- Parker, J., and C. Julliard, 2005, “Consumption risk and the cross section of expected returns,” *Journal of Political Economy*, 113, 185–222.
- Pastor, L., and R. F. Stambaugh, 2003, “Liquidity risk and expected stock returns,” *Journal of Political Economy*, 111, 642–685.
- Pennacchi, G., 2008, *Theory of Asset Pricing*, Pearson Education.

- Petersen, M. A., 2009, “Estimating standard errors in finance panel data sets: comparing approaches,” *The Review of Financial Studies*, 22, 435–480.
- Priestley, M. B., 1981, *Spectral analysis and time series*, Academic Press.
- Ranaldo, A., and P. Söderlind, 2010, “Safe haven currencies,” *Review of Finance*, 14, 385–407.
- Ritchey, R. J., 1990, “Call option valuation for discrete normal mixtures,” *Journal of Financial Research*, 13, 285–296.
- Ruiz, E., 1994, “Quasi-maximum likelihood estimation of stochastic volatility models,” *Journal of Econometrics*, 63, 289–306.
- Silverman, B. W., 1986, *Density estimation for statistics and data analysis*, Chapman and Hall, London.
- Singleton, K. J., 2006, *Empirical dynamic asset pricing*, Princeton University Press.
- Söderlind, P., 1999, “An interpretation of SDF based performance measures,” *European Finance Review*, 3, 233–237.
- Söderlind, P., 2000, “Market expectations in the UK before and after the ERM crisis,” *Economica*, 67, 1–18.
- 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*, 16, 1005–1008.
- 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.
- Stekler, H. O., 1991, “Macroeconomic forecast evaluation techniques,” *International Journal of Forecasting*, 7, 375–384.
- Svensson, L. E. O., 1989, “Portfolio choice with non-expected utility in continuous time,” *Economics Letters*, 40, 313–317.

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

Treynor, J. L., and K. Mazuy, 1966, “Can Mutual Funds Outguess the Market?,” *Harvard Business Review*, 44, 131–136.

Vasicek, O. A., 1977, “An equilibrium characterization of the term structure,” *Journal of Financial Economics*, 5, 177–188.

Verbeek, M., 2012, *A guide to modern econometrics*, Wiley, 4th edn.

Weil, P., 1989, “The equity premium puzzle and the risk-free rate puzzle,” *Journal of Monetary Economics*, 24, 401–421.

Wooldridge, J. M., 2010, *Econometric analysis of cross section and panel data*, MIT Press, 2nd edn.