

Modern Finance Theory

Paul Söderlind¹

16 November 2025

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

These lecture notes are for a first M.A. course in finance. The goal is to present financial theory in a way so that it can be used directly in quantitative/empirical projects that require numerical estimations and computations. The approach is therefore formal, but the mathematics is relatively easy, for instance, linear algebra is used, but stochastic calculus is not. Also, in terms of scope, the focus is on classical concept, for instance, the tangency portfolio plays an important role, but pricing kernels do not. Optional (often more advanced) material is denoted by a star (*).

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

When I first set up this course many years ago, I was inspired by the texts of Bodie, Danthine&Donaldson, Elton&Gruber, and Hull. Most likely that still shows.

My students at the MiQEF program at the University of St. Gallen have asked many good questions and pointed out mistakes. Also my teaching assistants did the same. Without that, these notes would have been worse.

Data Sources

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

1. The website of Kenneth French,

<http://mba.tuck.dartmouth.edu/pages/faculty/ken.french>

2. Bloomberg

3. Datastream

4. Federal Reserve Bank of St. Louis (FRED),

<http://research.stlouisfed.org/fred2/>

5. The website of Robert Shiller,

<http://www.econ.yale.edu/~shiller/data.htm>

6. yahoo! finance, <http://finance.yahoo.com/>

7. OlsenData, <http://www.olsendata.com>

Contents

1 The Basics of Return Calculations	7
1.1 Asset Returns	7
1.2 Portfolio Returns	15
1.3 Asset Classes	19
1.4 Markets, Instruments and Some Key Terms	19
1.8 Appendix – Matrix Algebra*	20
2 The Basics of Portfolio Choice	27
2.1 Expected Portfolio Return and Variance	27
2.2 Leverage	28
2.3 Diversification	30
2.4 Covariances Do Matter	34
2.5 Appendix – Statistics*	36
3 The Mean-Variance Frontier	44
3.1 The Mean-Variance Frontier of Risky Assets	44
3.2 The Mean-Variance Frontier of Risk-Free and Risky Assets	53
3.3 The Tangency Portfolio	56
3.4 Appendix – Calculus*	59
3.5 Appendix – Optimization*	62
4 The Inputs to MV Calculations	65
4.1 The Market Model: Betas	66
4.2 Estimation of the Covariance Matrix of the Asset Returns	70
4.3 Covariance Matrix with Time-Varying Parameters	71
4.4 Covariance Matrix with Average Correlations	71
4.5 Covariance Matrix from a Single-Index Model	72

4.6	Covariance Matrix from a Multi-Index Model	73
4.7	Covariance Matrix From A Shrinkage Estimator	73
4.8	An Evaluation of Different Approaches	74
4.9	Estimating Expected Returns	75
5	Portfolio Choice	76
5.1	Portfolio Choice with MV Preferences	76
5.2	A Single Risky Asset and a Risk-Free Asset	77
5.3	Several Risky Assets and a Risk-Free Asset	79
5.4	MV Preferences Gives a Portfolio on the MV Frontier	83
5.7	Appendix – Numerical Optimization Routines*	85
6	CAPM	90
6.1	Beta Representation of Expected Returns	90
6.2	More Properties of CAPM	95
6.3	Testing CAPM	98
6.3	Appendix – Discounted Cash Flow*	102
7	Downside Risk Measures	109
7.1	Value at Risk	109
7.2	Expected Shortfall	116
7.3	Target Semi-Variance	118
7.4	Maximum Drawdown	121
7.5	Empirical Return Distributions	122
8	Utility-Based Portfolio Choice	125
8.1	Utility Functions and Risky Investments	125
8.2	Utility-Based Portfolio Choice and MV Frontiers	129
8.3	Behavioural Finance	138
8.4	Appendix – Risk Aversion and the Level of Wealth*	140
8.5	Appendix – Portfolio Choice with $N()$ Returns*	143
9	Multi-Factor Models	145
9.1	Factor Investment	145
9.2	An Overview of Multi-Factor Models	148
9.3	Portfolio Choice with Background Risk	149

9.4	Asset Pricing Implications	153
9.5	Joint Portfolio and Savings Choice	156
9.6	Testing Multi-Factors Models	161
9.7	Appendix – The Asset Pricing Implications*	162
10	Efficient Markets	167
10.1	The Efficient Market Hypothesis	167
10.2	Autocorrelations and Autoregressions	168
10.3	Other Predictors and Methods	171
10.4	Out-of-Sample Forecasting Performance	173
10.5	Security Analysts	180
11	Performance Analysis	184
11.1	Performance Evaluation	184
11.2	Holdings-Based Performance Measurement	192
11.3	Performance Attribution	193
11.4	Style Analysis	194
12	Investment for the Long Run	196
12.1	Time Diversification	196
12.2	Mean-Variance Portfolio Choice	200
12.3	Appendix – The Conditional Variances*	206
13	Dynamic Portfolio Choice	208
13.1	Logarithmic Utility	208
13.2	CRRA Utility	210
13.3	Intertemporal Hedging	212
14	Foreign Exchange	218
14.1	Investing in Foreign Currency	218
14.2	Exchange Rate Quotation*	223
14.3	Currency Risk in Foreign Investments	226
14.4	Hedging Exchange Rate Movements	227
14.5	Explaining Exchange Rates	229

15 Forwards and Futures	231
15.1 Derivatives	231
15.2 Present Value	231
15.3 Forward Contracts	232
15.4 Forwards versus Futures	238
15.5 Swap Contracts	240
16 Interest Rate Calculations	241
16.1 Zero Coupon Bonds	241
16.2 Forward Rates	246
16.3 Coupon Bonds	249
16.4 Other Credit Instruments	255
16.5 Appendix – Estimating the Yield Curve*	258
16.6 Appendix – Conventions on Important Markets*	265
16.7 Appendix – More Proofs and Details*	269
17 Hedging Bonds	272
17.1 Bond Hedging	272
17.2 Duration: Definitions	273
17.3 Duration to Hedge a Bond Portfolio	277
17.4 Addressing Issues in Duration Hedging	284
18 Interest Rate Models	288
18.1 Empirical Properties of Yield Curves	288
18.2 Yield Curve Models	289
18.3 The Vasicek Model: Hedging a Bond	295
18.4 Interest Rates and Macroeconomics*	299
18.5 Forecasting Interest Rates*	305
18.6 Risk Premia on Fixed Income Markets	305
18.7 Appendix – Formal Derivation of the Vasicek Model*	305
19 Basic Properties of Options	309
19.1 Derivatives	309
19.2 Introduction to Options	310
19.3 Financial Engineering	315
19.4 Prices of Options	318

19.5 Put-Call Parity for European Options	319
19.6 Definition of American Calls and Puts	323
19.7 Basic Properties of Option Prices	325
19.8 Pricing Bounds and Convexity	327
19.9 Early Exercise of American Options	330
20 The Binomial Option Pricing Model	335
20.1 Overview of Option Pricing	335
20.2 The Basic Binomial Model	335
20.3 The Risk Neutral Probabilities	342
20.4 Multi-Period Trees I: Basic Setup	343
20.5 Multi-Period Trees II: Calibrating the Tree	349
20.6 Appendix – Continuous Dividends*	355
21 The Black-Scholes Model	359
21.1 The Black-Scholes Model	359
21.2 Deriving B-S I: Risk Neutral Pricing	363
21.3 Deriving B-S II: Convergence of the BOPM	364
21.4 Testing the B-S Model	369
21.5 Appendix – Details on the B-S Model*	371
21.6 Appendix – Probabilities in the BOPM and B-S Models*	373
21.7 Appendix – Statistical Tables	375
22 Hedging Options	378
22.1 Hedging an Option	378
22.2 An Approximate Hedge	379
22.3 Higher-Order Hedging*	383
22.4 Appendix – Hedging in the Binomial Model*	386
22.5 Appendix – More Greeks*	389

Chapter 1

The Basics of Return Calculations

This chapter first defines *returns*, demonstrates how to summarise their statistical properties (descriptive statistics), and discusses how to accumulate them. It then shows how *portfolio returns* depend on the returns of the assets in the portfolio. Later sections summarise the basic statistical properties of important asset classes and some key markets and trading concepts.

1.1 Asset Returns

1.1.1 Definition of a Return

The *net (rate of) return* on an asset in period t is

$$R_t = \frac{V_t - V_{t-1}}{V_{t-1}} = \frac{V_t}{V_{t-1}} - 1, \quad (1.1)$$

where V_t is the value of the asset in period t .

Remark 1.1 (*On notation*) A precise notation (which time, investment horizon, asset, units. ...) can be cumbersome. When needed, we will use $R_{i,t}$ (but $R_{i,t-1}$) to indicate the return of asset i in period t ($t - 1$). However, when dealing with a single asset where the time dimension is important, then we just keep the t subscript. Instead, when dealing with several assets but where the time dimension is less important, then we just keep the i subscript. Sometimes we drop all subscripts. The meaning should be clear from the context.

The gross return is

$$1 + R_t = \frac{V_t}{V_{t-1}}. \quad (1.2)$$

Example 1.2 (Returns)

$$R = \frac{110 - 100}{100} = 0.1 \text{ (or } 10\%)$$

$$1 + R = \frac{110}{100} = 1.1$$

Remark 1.3 (% and bp) Recall that 6% means $6/100 = 0.06$, and 400 bp (basis points) means $400/10000 = 0.04$. Warning: if you just drop the % symbol and thus effectively work with $100R$ (in this case getting 6), then you have to be careful, in particular, when accumulating returns over time and when calculating variances.

In many cases, the values are

$$V_{t-1} = P_{t-1} \text{ (price yesterday)}$$

$$V_t = D_t + P_t \text{ (dividend + price today)}, \quad (1.3)$$

so the return can be written

$$\begin{aligned} R_t &= \frac{D_t + P_t - P_{t-1}}{P_{t-1}} \\ &= \underbrace{\frac{D_t}{P_{t-1}}}_{\text{dividend yield}} + \underbrace{\frac{P_t - P_{t-1}}{P_{t-1}}}_{\text{capital gain yield}} \end{aligned} \quad (1.4)$$

Example 1.4 (Dividend yield ad capital gain yield)

$$R = \frac{2}{100} + \frac{108 - 100}{100} = 0.1$$

The *excess return* of an asset (compared to the risk-free rate R_f) is

$$R_t^e = R_t - R_{ft}. \quad (1.5)$$

In other cases, we use something else than a risk-free return as the reference rate.

Example 1.5 (Excess return) If $R_t = 0.08$ and $R_{ft} = 0.01$, then the excess return is $R_t^e = 0.07$ (7%).

Remark 1.6 (Approximating the risk-free return*) Suppose you have monthly equity returns and want to calculate excess returns. Do as follows. First, find a representative money market instrument (for instance, a T-bill or an interbank contract) with approximately one month to maturity. Second, use the interest rate on that instrument quoted

a month ago. divided by 1. (This is the rate you earn/pay on a loan between a month ago and now.) The result is an approximation since interest rates are quoted in different ways (simple, effective,...) and because the maturity may not be an exact match with the investment horizon.

1.1.2 Logarithmic Returns*

It is sometimes better to work with *log returns*, especially when we compare different investment horizons for the same asset. In contrast, log returns are somewhat inconvenient when the focus is on choosing the portfolio weights: the log portfolio return is *not* a weighted average of the log returns of the assets in the portfolio. (An approximation might work, as demonstrated in the chapter on dynamic portfolio choice.)

Anyhow, a log return is defined as

$$r_t = \ln(1 + R_t), \quad (1.6)$$

which clearly equals $\ln(V_t / V_{t-1})$. To convert from log returns to net returns, use $R_t = \exp(r_t) - 1$.

The corresponding excess log return is

$$r_e^e = \ln(1 + R_t) - \ln(1 + R_{ft}). \quad (1.7)$$

Assuming we invest an equal amount in both instruments in $t - 1$ ($V_{t-1} = V_{f,t-1}$), the excess log return equals $\ln(V_t / V_{ft})$. Notice that excess log return is *not* the log of the excess return. Rather, it is the log of $(1 + R_t)/(1 + R_{ft})$. Figure 1.1 illustrates that the difference between r^e and the possible approximation $\ln(1 + R^e)$ can be substantial.

Example 1.7 (Excess log return) If $R_t = 0.08$ and $R_{ft} = 0.01$, then the excess log return is 0.067.

1.1.3 Inflation and Real Returns

In most portfolio choice models, it is the *real* return (measured in units of “goods”) that matters, not the *nominal* return (measured in currency units). The reason is straightforward: utility depends on real goods and services, not on nominal price levels.

To see the link between real and nominal returns, let Γ_t be the nominal price level (price of the consumption basket of an investor, measured in currency units). If V in (1.1) is a nominal value (measured in currency units), then the real value is V/Γ .

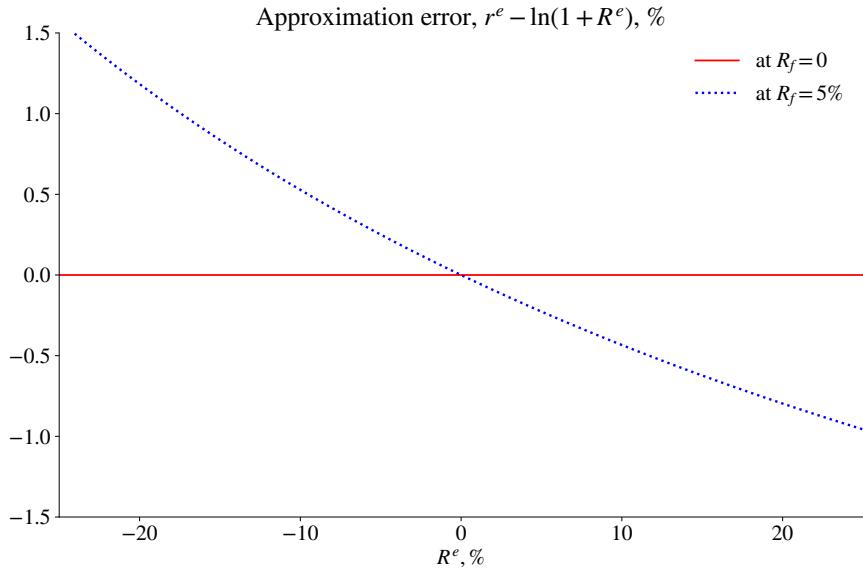


Figure 1.1: Approximation error from using $\ln(1 + R^e)$ instead of r^e

Example 1.8 (*Nominal and real prices*) If $V = 110$ is the nominal value of an asset and $\Gamma = 5$ is the nominal price of the consumption basket, then the real value is $V/\Gamma = 22$. This represents the number of consumption baskets required to match the asset's value.

The real return (corresponding to (1.1)) is

$$\begin{aligned}\tilde{R}_t &= \frac{\Gamma_{t-1}}{\Gamma_t} \frac{V_t}{V_{t-1}} - 1 \\ &= \frac{1 + R_t}{1 + \pi_t} - 1,\end{aligned}\tag{1.8}$$

where $\pi_t = \Gamma_t/\Gamma_{t-1} - 1$ is the inflation rate. We get a similar expression for the risk-free rate, so the excess real return (cf. (1.5)) is

$$\tilde{R}_t^e = \frac{R_t^e}{1 + \pi_t}.\tag{1.9}$$

Example 1.9 (*Real returns*) With $R_t = 0.08$ and $\pi_t = 0.05$, the real return is $1.08/1.05 - 1 \approx 0.029$. Also, with $R_t^e = 0.07$, the excess real return is $0.07/1.05 \approx 0.067$.

It is clear that the real excess return (1.9) is less affected by inflation than the real net return (1.8). (Actually, for log returns the real excess log return is unaffected by inflation.) The reason is straightforward: while inflation reduces the real value of the long position,

it also reduces the real value of the short position. In practice, many investors use the traditional excess return (R_t^e , not \tilde{R}_t^e) as a proxy for real excess returns.

1.1.4 Descriptive Statistics of Asset Returns

The properties of returns in a sample are often summarised by the mean, standard deviation, the Sharpe ratio (mean/std of excess returns) and the coefficients from a linear regression (see below).

Remark 1.10 *(On notation) Mean returns are denoted $E R$ or μ . (Subscripts to indicate the asset are used when needed.) An expression like $E x^2$ means the expected value of x^2 and $E xy$ is the expectation of the product xy . Variances are denoted σ^2 or $\text{Var}(R)$ and the standard deviations σ or $\text{Std}(R)$. Covariances are denoted σ_{ij} or $\text{Cov}(R_i, R_j)$. Clearly, σ_{ii} is the same as the variance.*

The *scaling of returns* (for instance, in percentages) can often cause confusion. Let R_{it} be the net return with mean μ , standard deviation σ and covariance with asset j σ_{ij} . When you work with percentage returns, $100R_{it}$, then

$$\begin{array}{ll} \text{mean:} & 100\mu \\ \text{variance:} & 100^2\sigma^2 \\ \text{standard deviation} & 100\sigma \\ \text{covariance with } 100R_{jt} & 100^2\sigma_{ij} \end{array} \quad (1.10)$$

100 R_{it} has the

Notice that the mean and standard deviation are scaled by 100, but the variance and covariance are scaled by 10,000. This can easily cause problems when trading off means and variances. However, it works well when comparing means and standard deviations (for instance, the Sharpe ratio is a mean divided by a standard deviation). Also, in a regression, $\tilde{R}_{it} = \alpha + \beta \tilde{R}_{jt} + \varepsilon_{it}$, the slope is unaffected, but the intercept is scaled by 100.

It is a common convention to *annualise return statistics* when reporting the results. If the return data is for a $1/k$ -year horizon (for instance, $k = 12$ for monthly data), then we typically annualise as

$$\begin{array}{ll} \text{mean:} & k\mu \\ \text{variance:} & k\sigma^2 \\ \text{standard deviation} & \sqrt{k}\sigma \\ \text{covariance with } R_j & k\sigma_{ij}. \end{array} \quad (1.11)$$

For daily data use $k = 252$ (the approximate number of trading days per year) and for weekly data $k = 52$. Also, the results from a linear regression are annualised by multiplying the intercept by k (since it is a mean), but not changing the slope coefficient (since it is a covariance divided by a variance). The convention in (1.11) is based on the idea that returns are almost iid (see below for details). It is probably advisable to annualise only at the very last stage of the computations.

Example 1.11 (*Annualisation*) *If the monthly average return is 0.67% and the monthly standard deviation is 2.89%, then the annualised values are 8% and 10%, respectively.*

The expected excess return, $E R_i^e$ or μ_i^e , is often called a *risk premium* since it measures the expected return of taking risk (of holding asset i) minus the return of a risk-free asset. The *Sharpe ratio* is

$$SR = \mu^e / \sigma, \quad (1.12)$$

where (μ^e, σ) indicate the mean and standard deviations of the excess returns. The SR can be interpreted as a reward/risk ratio. Typically, a high Sharpe ratio is considered favourable.

Example 1.12 (*Risk premium and Sharpe ratio*) *If $(\mu^e, \sigma) = (0.1, 0.5)$, then Sharpe ratio is 0.2.*

The so-called “market model” is a regression of an asset’s excess return on the excess return on the market index (R_{mt}^e)

$$R_t^e = \alpha + \beta R_{mt}^e + u_t. \quad (1.13)$$

A slope coefficient $\beta > 1$ indicates that the asset is strongly pro-cyclical (moves more than proportionally with the market), whereas $0 < \beta < 1$ indicates a weaker pro-cyclicality. This is sometimes referred to as “cyclical” and “defensive” assets, respectively. $\beta < 0$ indicates counter-cyclicality, but such assets are rare. The α is often interpreted as an abnormal excess return (see the chapters on CAPM and performance measures). See Table 1.1 for an example.

Remark 1.13 (*Motivation of the convention in (1.11)**) *Suppose we have semi-annual data. Notice that an annual return would be $P_t/P_{t-2} - 1 \approx R_t + R_{t-1}$. If returns are iid (in particular, the same mean and variance across time and also uncorrelated across time), then to a first approximation, the expected annual return is $E(R_t + R_{t-1}) = 2 E R_t$.*

	Small growth	Small value	Large growth	Large value	Equity market
mean (ann.)	7.31	11.13	10.30	9.69	9.17
std (ann.)	23.39	20.00	15.99	18.50	15.62
SR (ann.)	0.31	0.56	0.64	0.52	0.59
α (ann.)	-4.46	1.44	1.16	0.43	0.00
β	1.28	1.06	1.00	1.01	1.00

Table 1.1: Means and std of asset class returns, US, monthly excess returns (%), 1985:01-2024:12. The mean and α are annualised by 12, the standard deviation by $\sqrt{12}$, and the Sharpe ratio is the ratio of the annualised mean and standard deviation.

Similarly, the variance is $\text{Var}(R_t + R_{t-1}) = 2 \text{Var}(R_t)$, which implies $\text{Std}(R_t + R_{t-1}) = \sqrt{2} \text{Std}(R_t)$. Similarly, if $\text{Cov}(R_{it}, R_{j,t-1}) = 0$, then $\text{Cov}(R_{it} + R_{i,t-1}, R_{jt} + R_{j,t-1}) = 2\sigma_{ij}$.

1.1.5 Cumulating Returns

If an investment in period $t = 0$ equals V_0 , then its value in t is

$$V_t = V_0(1 + R_1)(1 + R_2) \dots (1 + R_t), \quad (1.14)$$

where all subscripts refer to time periods and R_τ is the return on your portfolio in period τ . This expression assumes that all dividends have been reinvested, making V_t a *total return index*. We can clearly write this on recursive form as

$$V_t = V_{t-1}(1 + R_t). \quad (1.15)$$

Empirical Example 1.14 Figure 1.2 shows the cumulated return of a U.S. equity market index.

Example 1.15 With net returns for three time periods $(R_1, R_2, R_3) = (0.2, -0.35, 0.25)$, we get portfolio values $(1.2, 0.78, 0.975)$ for period 1 – 3 (assuming $V_0 = 1$).

Remark 1.16 (Adjusted closing price) The adjusted closing price of an asset is an index calculated as (1.15) where R_t is the return (including dividends, splits, etc) of holding the asset from $t - 1$ to t . This means that it is a total return index. If you have such an index, then the returns can be calculated as $R_t = V_t / V_{t-1} - 1$, without having to handle dividend payments separately.

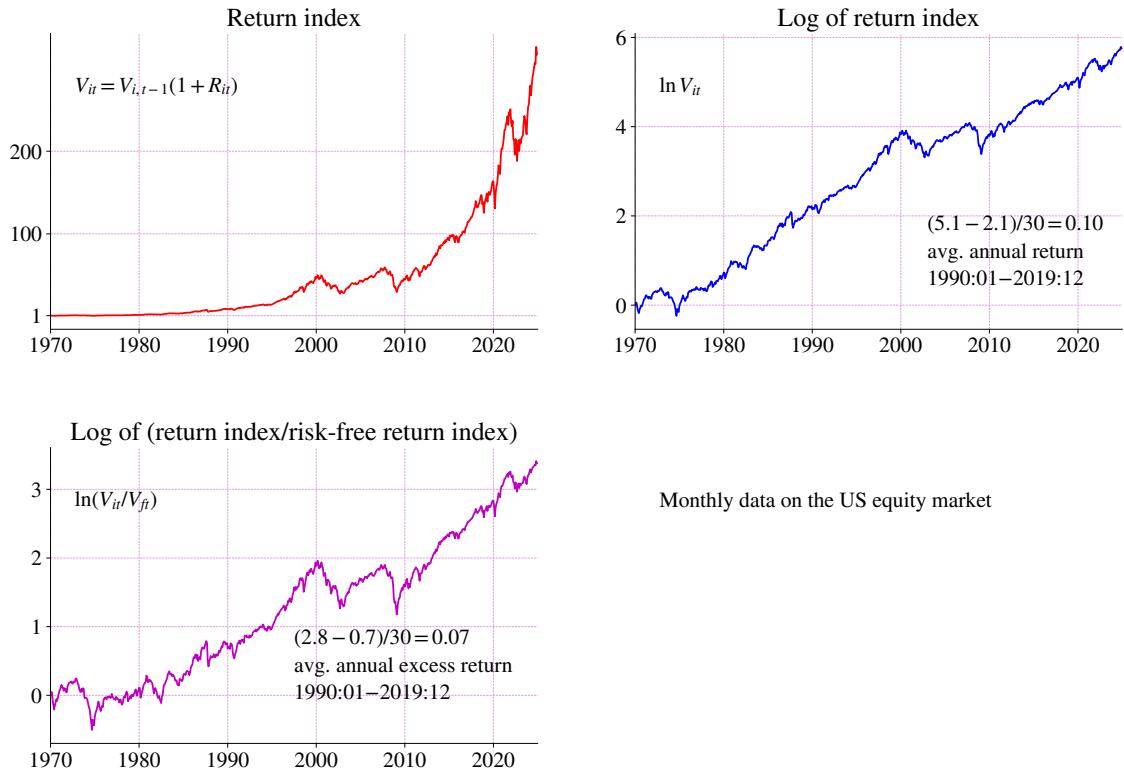


Figure 1.2: Cumulating returns

Unfortunately, excess returns cannot be cumulated directly. Instead, you need to cumulate the net return R_t and the risk-free return R_{ft} separately (as in (1.15)) and then form the difference

$$V_t^e = V_t - V_{ft}. \quad (1.16)$$

Sometimes the ratio V_t / V_{ft} is a preferred way of illustrating the performance of the two assets.

1.1.6 Cumulating Logarithmic Returns*

Similarly, the log value can be calculated as

$$\ln V_t = \ln V_0 + r_1 + r_2 \dots + r_t, \text{ so} \quad (1.17)$$

$$= \ln V_{t-1} + r_t. \quad (1.18)$$

You *can* cumulate excess log returns (because it is just summing). Since the initial

positions are equal ($V_0 = V_{f,0}$) we have

$$\ln(V_t / V_{ft}) = (r_1 + r_2 + \dots + r_t) - (r_{f1} + r_{f2} \dots + r_{ft}) \quad (1.19)$$

$$= r_1^e + r_2^e + \dots + r_t^e, \text{ so} \quad (1.20)$$

$$= \ln(V_{t-1} / V_{f,t-1}) + r_t^e, \quad (1.21)$$

starting from $\ln(V_0 / V_{f0}) = 0$. Notice that the exponential function of this gives the ratio V_t / V_{ft} (not the difference). Again, see Figure 1.2 for an illustration

1.2 Portfolio Returns

Remark 1.17 *(On notation) These notes use $\sum_{i=1}^n x_i$ to denote the sum $x_1 + \dots + x_n$. (In the running text, it might happen that this is sometimes written as just $\Sigma_i x_i$.) Note: Σ may also denote a variance-covariance matrix. The distinction should be clear from the context.*

1.2.1 Portfolio Return: Definition

Let R_i represent the return on asset i over a given time period (the time subscript is omitted for convenience). The return on a portfolio (R_p) with the portfolio weights w_1, w_2, \dots, w_n is

$$R_p = \sum_{i=1}^n w_i R_i, \text{ with } \sum_{i=1}^n w_i = 1. \quad (1.22)$$

Using vectors, this can also be written

$$R_p = w' R, \quad (1.23)$$

where w is an n -vector of weights and R an n -vector of asset returns.

Clearly, one of the assets in (1.22)–(1.23) could be risk-free with return R_f . However, in this case we will typically choose to consider n risky assets and the risk-free (in total, $n + 1$) and write the portfolio return as

$$R_p = v' R + (1 - \mathbf{1}' v) R_f \quad (1.24)$$

$$= v' R^e + R_f, \quad (1.25)$$

where v are the weights on the risky assets and $1 - \mathbf{1}' v$, that is, $1 - \sum_{i=1}^n v_i$, the weight on the risk-free asset. This automatically imposes the condition that the weights on *all* assets

sum to one.

Example 1.18 (*Portfolio return*) With the portfolio weights 0.8 and 0.2 for two assets and the returns 0.1 and 0.05 for the same assets, the portfolio has the return

$$R_p = 0.8 \times 0.10 + 0.2 \times 0.05 = 0.09,$$

that is, 9%.

Example 1.19 (*Number of assets and portfolio returns**) For asset 1 we have $P_{1,t-1} = 10$, $P_{1,t} = 11$ and for asset 2 we have $P_{2,t-1} = 8$, $P_{2,t} = 8.4$. Assume no dividends. Yesterday you bought 16 of asset 1 and 5 of asset 2: $16 \times 10 + 5 \times 8 = 200$. Today your portfolio is worth $16 \times 11 + 5 \times 8.4 = 218$, so $R_p = (218 - 200)/200 = 0.09$. This is the same as in Example 1.18 since the two returns are 0.1 ($11/10 - 1$) and 0.05 ($8.4/8 - 1$) respectively, and the portfolio weights are 0.8 ($16 \times 10 / 200$) and 0.2 ($5 \times 8 / 200$) respectively.

1.2.2 Portfolio Return with Short Positions

The portfolio weights in (1.22) should sum to unity ($\sum_{i=1}^n w_i = 1$), but some weights could potentially be negative: “*short*” positions. Notice that a short position pays off if the asset price decreases. Clearly, some investors have very strict limits on their positions. For instance, mutual funds can typically not shorten assets and not put more than 10% in a particular asset. In contrast, hedge funds have very few limits.

Remark 1.20 (*Short selling*) How can we short sell an asset? Borrow the asset (for a fee and typically against collateral) and sell it. If there are derivatives on the asset, then we do not need to borrow it: just issue a futures/option.

Example 1.21 (*Return on a short position*) Suppose you borrow an asset (for one month, at a fee of 0.5) and sell it for 100. One month later, you buy the asset on the market for 90. Your profit is thus $100 - 90 - 0.5 = 9.5$. Expressed in terms of the initial value of the asset, this is a return of 9.5%.

1.2.3 Zero-Cost Portfolios*

A zero-cost “portfolio” (also called an arbitrage portfolio) means that the investor shortens some assets (perhaps borrows) in order to invest in other (perhaps risky) assets. The return

on such a portfolio is not well defined (dividing by zero...), but we can define an excess return as follows. Split up the portfolio in a “long” portfolio and denote the weights by w_i^L , and a “short” portfolio with weights w_i^S . Clearly, $w_i^L \geq 0$ and $w_i^S \geq 0$ and when one of them is positive then the other is zero.

Example 1.22 (*Zero-cost portfolio*) Suppose you invest 40 in asset 1, 60 in asset 2 and -100 in asset 3. The total investment is zero. We then have $w^L = (0.4, 0.6, 0)$ and $w^S = (0, 0, 1)$.

Define the returns on the long and short portfolios as

$$R_p^L = \sum_{i=1}^n w_i^L R_i \quad (1.26)$$

$$R_p^S = \sum_{i=1}^n w_i^S R_i, \quad (1.27)$$

where all subscripts refer to different assets (and the subscripts for time are suppressed).

We can then consider an “excess return” of the total portfolio as

$$R_p^e = R_p^L - R_p^S = \sum_{i=1}^n (w_i^L - w_i^S) R_i. \quad (1.28)$$

Conversely, the traditional excess return of an asset (1.5) is the return of a zero cost portfolio: a long position in the asset and a short position in the risk-free asset.

Example 1.23 (*Excess return of a zero-cost portfolio*) If the returns of the assets in Example 1.22 are $R_1 = 10\%$, $R_2 = -1\%$ and $R_3 = 2\%$, then the excess return is

$$0.4 \times 0.1 + 0.6 \times (-0.01) - 0.02 = 0.014.$$

Remark 1.24 (*A broader definition of excess returns**) The definition of the excess return of a zero-cost portfolio discussed above uses portfolio weights that sum to unity ($\sum w_i^L = 1$ and $\sum w_i^S = 1$), which is often a natural choice. However, another convention is used in some cases: the “excess return” of a zero cost portfolio is just its payoff (profit).

1.2.4 Trading Costs

As a investor you typically pay a *commission* to the broker. In addition, the price depends on whether you are buying (high price, the ask price) or selling (low price, the bid price).

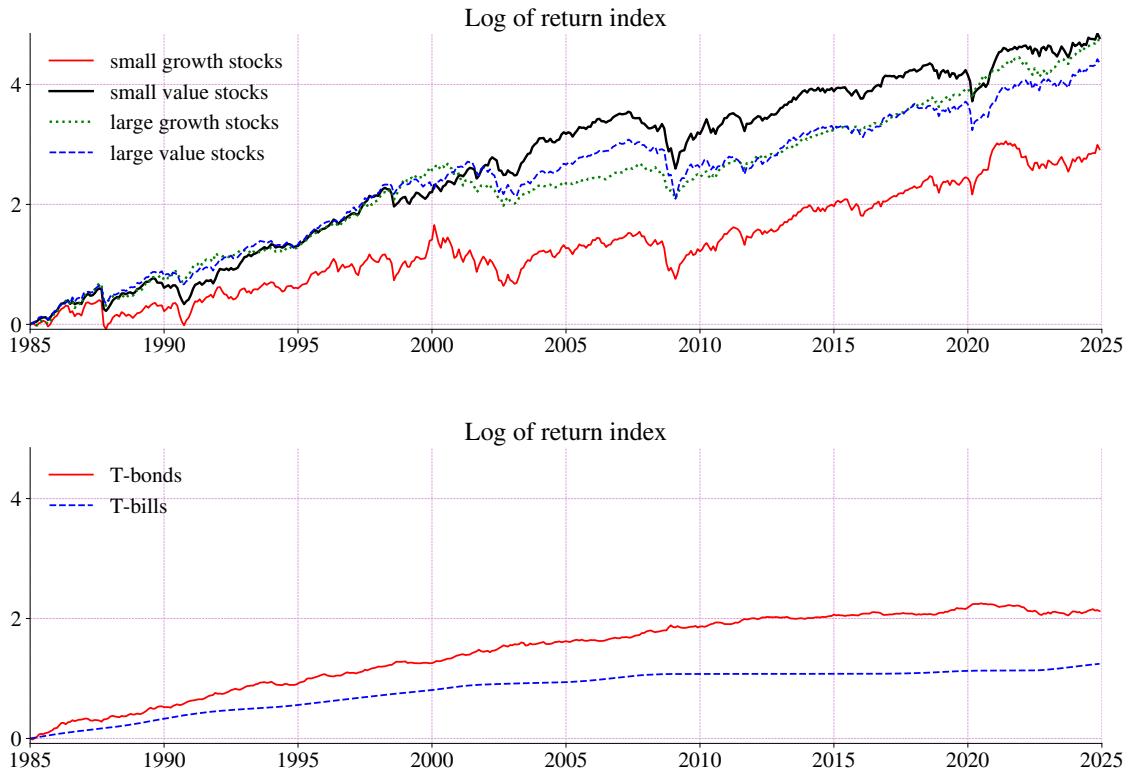


Figure 1.3: Performance of US equity and fixed income

Notice that portfolio weights shift as a result of returns (price changes). After returns have been realized, the new portfolio weight on asset i is

$$w_{it}(1 + R_{it})/(1 + R_{pt}), \quad (1.29)$$

where w_{it} was the initial weight. If $w_{i,t+1}$ is the desired weight going forward, the trading need for the portfolio is the absolute value of the difference to (1.29)

$$\sum_{i=1}^n |w_{i,t+1} - w_{it}(1 + R_{it})/(1 + R_{pt})|. \quad (1.30)$$

Example 1.25 (Trading need) For Example 1.18, (1.29) gives $0.8 \times 1.1/1.09 \approx 0.807$ and $0.2 \times 1.05/1.09 \approx 0.193$, respectively for the two assets. If the aim is to keep the weights fixed, then we need to sell off (buy) approximately 0.007 of asset 1 (2), so the total trading need is approximately 0.014.

Remark 1.26 (*Trading costs**) Suppose bid and ask prices are:

	<i>Definition</i>	<i>Example</i>
Ask price	<i>lowest price at which someone will sell</i>	90.05
Bid price	<i>highest price at which someone will buy</i>	90.00
Bid-ask spread		0.05

If you want to buy immediately: you submit a market buy order (*buy at best available price*) and you need to pay ask price (90.05). Instead, if you want to sell immediately, you submit a market sell order and get the bid price (90.00). A round-trip (first buy, then sell) costs $90.05 - 90.00 = 0.05$ (the bid-ask spread). Alternatively, you can (at least on some markets) submit a limit buy order at a higher bid price (eg. 90.01) or a limit sell order at a lower ask price (eg. 90.04). With some luck someone hits that order.

1.3 Asset Classes

Many investors and asset managers choose to focus on asset classes, rather than on individual assets. This approach helps average out idiosyncratic (for instance, firm specific) noise and focuses attention to the macroeconomic perspective.

Empirical Example 1.27 Table 1.2 illustrates the return distributions for different U.S. asset classes. There are distinct differences between small and large firms and between growth and value firms. However, the most pronounced difference is between equity and bonds (the latter have much less volatility and often lower returns). Figures 1.3 –1.4 illustrate the dynamics behind the figures for the entire sample in Table 1.2. Table 1.3 gives the annual ranking of the asset classes (for a shorter sample). Much of portfolio management is about trying to time these changes. The changes of the ranking—and in the returns—highlight both the opportunities (if you time it right) and risks (if you don't) with such an approach.

1.4 Markets, Instruments and Some Key Terms

The initial issuance of an asset (for instance, an IPO) takes place at the *primary market* while the subsequent trading takes place on the *secondary market*. Trading in the secondary market can be done on an *exchange* (NYSE, Tokyo, EuroNext, Nasdaq, London, Shanghai, HK, CBOE, CME, etc), an *electronic platform* (EBS, Reuters), or *over the counter* (OTC).

	Small growth	Small value	Large growth	Large value	Bonds	T-bills
mean	0.87	1.19	1.12	1.07	0.45	0.26
std	6.74	5.76	4.62	5.34	1.40	0.21
min	-32.48	-28.09	-23.22	-27.23	-4.39	0.00
max	28.09	19.54	14.47	18.17	5.31	0.79
market corr	0.86	0.82	0.97	0.85	0.00	0.03
beta	1.28	1.05	1.00	1.01	0.00	0.00

Table 1.2: Descriptive statistics of asset classes, US, monthly returns (%), 1985:01-2024:12. The beta is the slope coefficient from regressing the asset on the market return.

Different asset classes are typically traded on distinct exchanges or platforms. This motivates terms like the “money market”, “bond market”, “currency market”, “stock market”, and “derivative markets”.

Alternative asset classes (for instance, hedge funds, infrastructure, private equity) have gained interest over the last decade, especially among institutional long-run investors (wealthy individuals, endowments, some pension funds).

Remark 1.28 (*Useful terms*) *The following stock market terms are useful*

- Market capitalization: *value of all shares*
- Float: *number of not closely held shares*
- Volume: *number of traded shares*
- Short interest: *number of shortened shares*
- Consensus estimate: *the average forecast (of eg. earnings) across analysts*
- ROE: *net income/book value of equity*
- ROI: *(net income + interest rates)/book value of (equity + debt)*

1.8 Appendix – Matrix Algebra*

This appendix introduces fundamental concepts of matrix algebra.

The discussion will use the vectors and matrices

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

	6th	5th	4th	3rd	2nd	1st
2005	SG 0	B 3	TB 3	LG 4	SV 10	LV 14
2006	B 3	TB 5	SG 9	LG 11	LV 22	SV 22
2007	SV -14	LV -2	TB 5	SG 6	B 9	LG 13
2008	SG -41	LV -39	LG -34	SV -34	TB 2	B 14
2009	B -4	TB 0	LV 18	SV 30	LG 31	SG 37
2010	TB 0	B 6	LV 7	LG 15	SV 27	SG 29
2011	LV -11	SV -8	SG -6	TB 0	LG 4	B 10
2012	TB 0	B 2	SG 15	LG 15	SV 21	LV 29
2013	B -3	TB 0	LG 33	LV 40	SV 43	SG 45
2014	TB 0	SV 4	SG 5	B 5	LV 12	LG 14
2015	SV -10	LV -8	SG -3	TB 0	B 1	LG 4
2016	TB 0	B 1	SG 8	LG 9	LV 26	SV 37
2017	TB 1	B 2	SV 9	LV 18	SG 25	LG 29
2018	LV -15	SV -13	SG -8	LG 0	B 1	TB 2
2019	TB 2	B 7	SV 15	LV 28	SG 30	LG 34
2020	LV -3	TB 0	SV 4	B 8	LG 36	SG 57
2021	B -2	TB 0	SG 3	LG 25	LV 37	SV 42
2022	SG -28	LG -26	B -12	SV -6	TB 1	LV 4
2023	B 4	TB 5	SV 13	LV 14	SG 16	LG 38
2024	B 1	TB 5	SV 9	SG 18	LV 21	LG 30

Table 1.3: Ranking and return (in %) of asset classes, US. SG: small growth firms, SV: small value, LG: large growth, LV: large value, B: T-bonds, TB: T-bills.

where the elements (for instance, x_1 and A_{12}) are numbers. In most cases, we think of a vector as a matrix with one column (sometimes referred to as a column vector). In contrast, a row vector is a matrix with one row.

Example 1.29 (Vector and matrix)

$$x = \begin{bmatrix} 10 \\ 11 \end{bmatrix} \text{ and } A = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}.$$

1.8.1 Matrix and Scalar Addition and Multiplication

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

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

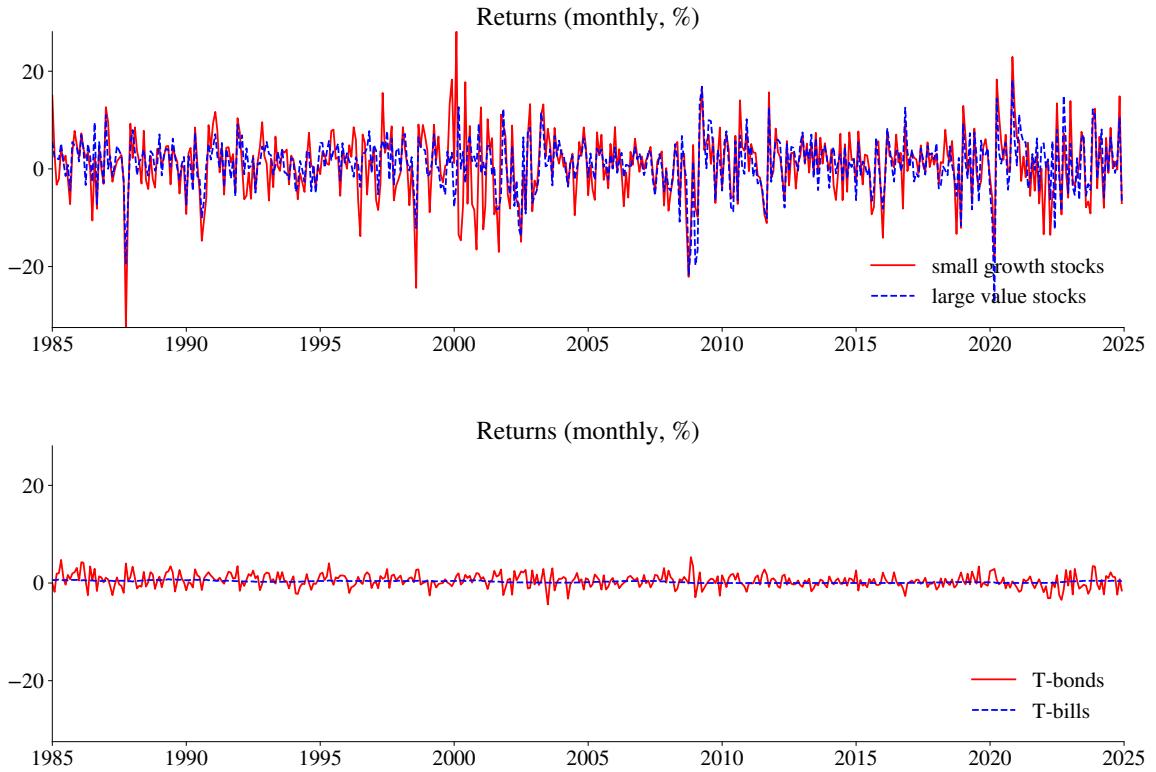


Figure 1.4: Performance of US equity and fixed income

Example 1.30 (*Matrix \times scalar*)

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

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

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

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

Example 1.31 (*Matrix ± scalar*)

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

1.8.2 Adding and Multiplying: Two Matrices

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

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

Example 1.32 (*Matrix addition and subtraction*)

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

Matrix *multiplication* requires the two matrices to be conformable: with AB where A is $m \times n$, B must be $n \times p$. Element ij of the result (which is $m \times p$) is the multiplication of the i th row of the first matrix with the j th column of the second matrix

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

As a special case, multiplying a matrix A with a vector z gives a new vector

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

Example 1.33 (*Matrix multiplication*)

$$\begin{aligned}\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & -2 \end{bmatrix} &= \begin{bmatrix} 10 & -4 \\ 15 & -2 \end{bmatrix} \\ \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} &= \begin{bmatrix} 17 \\ 26 \end{bmatrix}\end{aligned}$$

1.8.3 Transpose

Transposing a column vector gives a row vector. Similarly, transposing a matrix is like flipping it around the main diagonal so the former columns become the new rows

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

Example 1.34 (*Matrix transpose*)

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

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

1.8.4 Inner and Outer Products, Quadratic Forms

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

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

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

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

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

Example 1.35 (*Inner and outer products*)

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

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

If z is a vector and A a square matrix, then the product $z'Az$ is a quadratic form (a scalar)

$$z'Az = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}' \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = z_1A_{11}z_1 + z_1A_{12}z_2 + z_2A_{21}z_1 + z_2A_{22}z_2.$$

Example 1.36 (*Quadratic form*)

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

1.8.5 Kronecker Product

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

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

Example 1.37 (*Kronecker product*)

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \otimes \begin{bmatrix} 10 & 11 \end{bmatrix} = \begin{bmatrix} 10 & 11 & 30 & 33 \\ 20 & 22 & 40 & 44 \end{bmatrix}.$$

1.8.6 Matrix Inverse

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

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

where I is the *identity matrix* (ones along the diagonal, and zeros elsewhere). The matrix inverse is useful for solving systems of linear equations, $y = Ax$ as $x = A^{-1}y$. Notice that not every square matrix is invertible, in particular not if some rows (or columns) are linear combinations of the other rows (columns).

For a 2×2 matrix we have

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

Example 1.38 (*Matrix inverse*)

$$\begin{aligned} \begin{bmatrix} -0.8 & 0.6 \\ 0.6 & -0.2 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ so} \\ \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}^{-1} &= \begin{bmatrix} -0.8 & 0.6 \\ 0.6 & -0.2 \end{bmatrix}. \end{aligned}$$

1.8.7 Solving Systems of Linear Equations

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

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

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

Example 1.39 (*Solving a system of linear equations*)

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

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

Using this in the first equation verifies that we indeed get the right result.

1.8.8 Derivatives of Matrix Expressions

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

Example 1.40 (*Derivative of an inner product*) With $n = 2$

$$x'z = x_1z_1 + x_2z_2, \text{ so } \partial(x'z)/\partial x = \begin{bmatrix} \partial(z_1x_1 + z_2x_2)/\partial x_1 \\ \partial(z_1x_1 + z_2x_2)/\partial x_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

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

Example 1.41 (*Derivative of a symmetric quadratic form*) With $n = 2$, the symmetric quadratic form is

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

The derivatives with respect to x_1 and x_2 are

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

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

Chapter 2

The Basics of Portfolio Choice

There are two key elements to portfolio choice: (1) how to mix the risky assets with a risk-free asset (leverage) to handle the overall risk level; and (2) how to mix various (risky) assets to average out volatility (diversification). This chapter introduces each of these topics. Later chapters will put them together in a unified framework.

2.1 Expected Portfolio Return and Variance

This technical section summarizes how beliefs about expected returns (μ) of the investable assets and their variance-covariance matrix (Σ) can be combined with portfolio weights to calculate the implied beliefs about the portfolio returns. These beliefs should be interpreted as representing those of the investor, conditional on the information available at the time of the investment. In later chapters, we will extend this approach to find optimal weights.

Remark 2.1 (*Expected value and variance of a linear combination*) Recall that if w_1 and w_2 are two constants, while the returns R_1 and R_2 are random variables, then

$$\begin{aligned} \mathbb{E}(w_1 R_1 + w_2 R_2) &= w_1 \mu_1 + w_2 \mu_2, \text{ and} \\ \text{Var}(w_1 R_1 + w_2 R_2) &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \sigma_{12}, \end{aligned}$$

where $\mu_i = \mathbb{E} R_i$, $\sigma_{ij} = \text{Cov}(R_i, R_j)$, and $\sigma_i^2 = \text{Var}(R_i)$.

The expected return on the portfolio is (time subscripts are suppressed)

$$\begin{aligned} \mathbb{E} R_p &= \sum_{i=1}^n w_i \mu_i & (2.1) \\ &= w' \mu, & (2.2) \end{aligned}$$

where w is the n -vector of portfolio weights and μ is a corresponding vector of expected asset returns.

The variance of a portfolio return is

$$\text{Var}(R_p) = \sum_{i=1}^n w_i^2 \sigma_i^2 + \sum_{i=1}^n \sum_{j=1, j \neq i}^n w_i w_j \sigma_{ij} \quad (2.3)$$

$$= w' \Sigma w, \quad (2.4)$$

where Σ is the $n \times n$ variance-covariance matrix of the returns.

Remark 2.2 ($n = 2$) With two assets, $E R_p = w_1 \mu_1 + w_2 \mu_2$, $\text{Var}(R_p) = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \sigma_{12}$ and $\text{Cov}(R_q, R_p) = v_1 w_1 \sigma_1^2 + v_2 w_2 \sigma_2^2 + (v_1 w_2 + v_2 w_1) \sigma_{12}$.

Example 2.3 (Expected value and variance of portfolio return) Let the portfolio weights be $w = [0.8, 0.2]$. Assume the following the expected values and covariance matrix for the returns: $\mu = \begin{bmatrix} 9 \\ 6 \end{bmatrix} / 100$ and $\Sigma = \begin{bmatrix} 256 & 96 \\ 96 & 144 \end{bmatrix} / 100^2$. This gives

$$E R_p = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix} \begin{bmatrix} 9 \\ 6 \end{bmatrix} \frac{1}{100} = 0.084,$$

$$\text{Var}(R_p) = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix} \begin{bmatrix} 256 & 96 \\ 96 & 144 \end{bmatrix} \frac{1}{100^2} \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \approx 0.020, \text{ and}$$

$$\text{Std}(R_p) \approx 0.142.$$

More details and examples are found in the statistics appendix.

2.2 Leverage

2.2.1 A Portfolio of a Single Risky Asset and a Risk-free Asset

Suppose you can only invest in a risky asset (with return R) and a risk-free (with return R_f). The risky asset could represent the (equity) market portfolio. To observe the effect of the portfolio choice on the mean and the volatility, notice that

$$R_p = vR + (1 - v)R_f, \text{ so} \quad (2.5)$$

$$E R_p = v\mu + (1 - v)R_f \text{ and} \quad (2.6)$$

$$\text{Std}(R_p) = |v|\sigma, \quad (2.7)$$

where we use $(\mu$ and $\sigma)$ as short hand notation for the mean and standard deviation of the risky asset.

The expected value follows from $E R_f = R_f$ as the risk-free rate is known. Similarly, the standard deviation follows from $\text{Var}(R_p) = v^2 \sigma^2$, since $\text{Var}(R_f) = 0$ (the risk-free rate over the investment horizon is known when the portfolio is formed) and hence also the covariance is zero. If we use an interest rate to represent the risk-free rate, then we should typically use a maturity that corresponds to the investment horizon. Often a floating overnight rate is used instead, but that is (strictly speaking) not risk-free for investment horizons of more than one day. Still, the uncertainty might be so small that it can be used as an approximation.

How much to put in the risky asset is a matter of *leverage*, and v is often called the *leverage ratio*. This equals the investment in risky assets divided by our total capital.

Example 2.4 (*Leveraged portfolios*) *Portfolio weights for three different portfolios*

	<i>Portfolio A</i>	<i>Portfolio B</i>	<i>Portfolio C</i>	<i>Portfolio D</i>
v (<i>in risky assets</i>)	0.5	1	2	-1
$1 - v$ (<i>in risk-free</i>)	0.5	0	-1	2
<i>Sum</i>	1	1	1	1

Portfolio A: *your capital is 200, invest 100 in risky assets and 100 in risk-free*; Portfolio B: *your capital is 200, invest 200 in risky assets and 0 in risk-free*; Portfolio C: *your capital is 200, invest 400 in risky assets and -200 in risk-free (borrow 200 = short position in risk-free)*. Portfolio D: *short-sell the risky asset for 100 and put 200 in the risk-free*.

Remark 2.5 (*Assuming that R and R_f do not depend on v*) *These notes assume that the portfolio choice (here v) does not affect the returns. This means that we assume that the investor is small compared to the overall market. It also means that we effectively assume that lending ($1 - v > 0$) and borrowing ($1 - v < 0$) can be done at the same rate. This is a reasonable approximation for a large financial institution and simplifies the analysis considerably.*

The mean and the standard deviation in (2.6)–(2.7) are both scaled by the leverage ratio (v). Notice that taking on leverage (borrowing to invest in the risky asset) typically is a way to increase the expected return of the portfolio, but at the cost of increasing the risk.

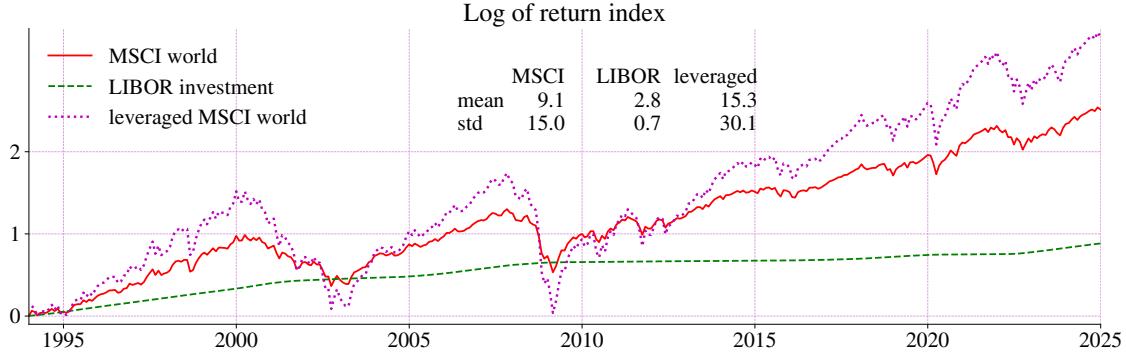


Figure 2.1: The effect of leverage on the portfolio performance

Empirical Example 2.6 *Figure 2.1 shows the effect on the log cumulated excess returns from holding a leveraged position in equity. The LIBOR rate (London Interbank Offered Rate) is, of course, not entirely without variation in this figure, thus, the result in (2.7) applies only approximately. However, for each separate 1-month investment horizon, the LIBOR rate is known in advance and thus risk-free. More recently, risk-free rates are often proxied by floating overnight rates.*

Example 2.7 With $(\mu, \sigma) = (9.5\%, 8\%)$ and $R_f = 0.03$, we get (in %)

	Portfolio A	Portfolio B	Portfolio C
Mean	6.25	9.5	16
Std	4	8	16

As long as the leverage ratio is positive ($v > 0$), we can combine (2.6)–(2.7) to get a relation between portfolio mean and standard deviation as

$$E R_p = R_f + SR \times \text{Std}(R_p), \quad (2.8)$$

where the slope is $SR = \mu^e / \sigma$ (the Sharpe ratio of the risky asset). This shows that the average portfolio return is linearly related to its standard deviation. (For $v < 0$, the relation is also linear, but with the slope $-SR$.) Figure 2.2 illustrates.

2.3 Diversification

This section demonstrates that the portfolio variance can be reduced by forming a portfolio by mixing (a) assets that are only weakly correlated and (b) many assets. These

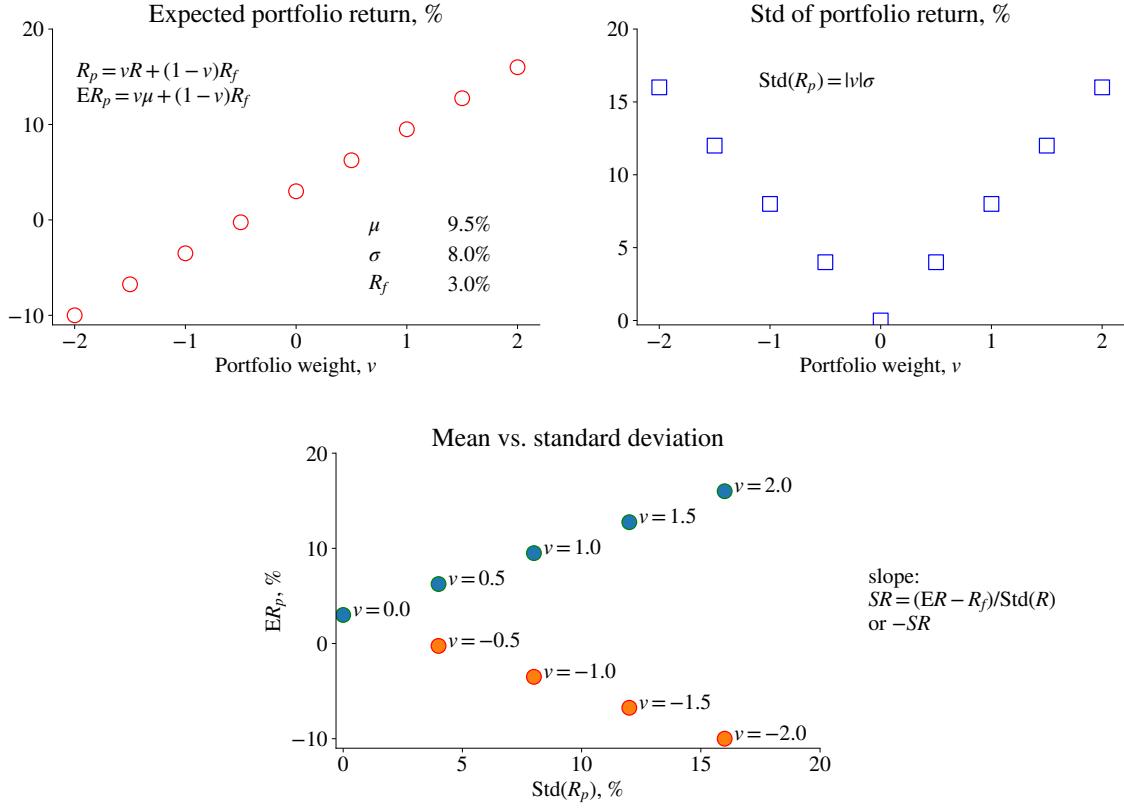


Figure 2.2: The effect of leverage on the mean and volatility of the portfolio return

diversification benefits can often be achieved without hurting the expected returns.

Recall that the variance of a portfolio return is

$$\text{Var}(R_p) = w' \Sigma w, \quad (2.9)$$

where w is the vector of portfolio weights and Σ the variance-covariance matrix. For instance, with two assets we have

$$\text{Var}(R_p) = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \sigma_{12}, \quad (2.10)$$

where w_i is the portfolio weight on asset i , σ_i^2 is the variance of asset i and σ_{ij} is the covariance of assets i and j .

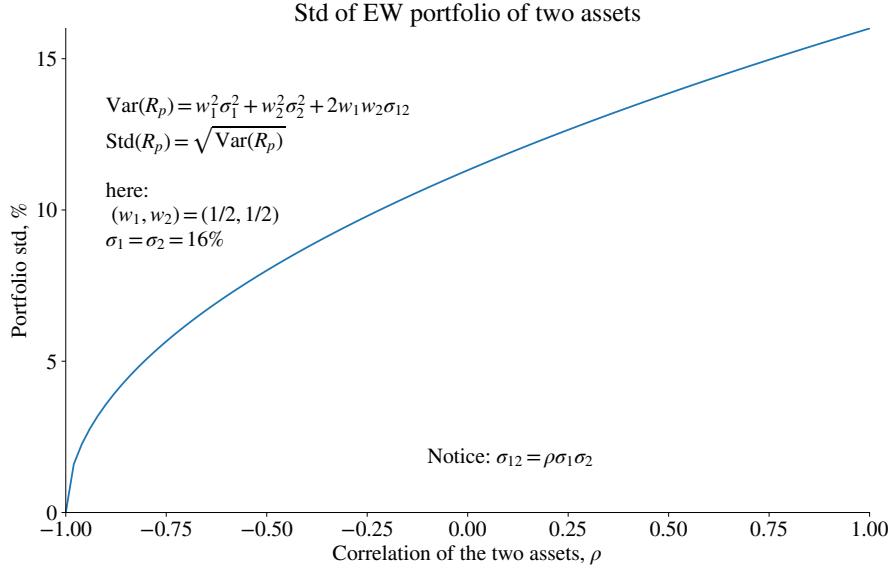


Figure 2.3: Effect of correlation on the diversification benefits

2.3.1 Diversification: The Correlations

As a simple example, consider an *equally weighted (EW) portfolio* of two risky assets (use $w_1 = w_2 = 1/2$ in (2.10)). Denote the correlation by ρ and write as (since $\sigma_{12} = \rho\sigma_1\sigma_2$)

$$\begin{aligned}\text{Var}(R_p) &= \sigma_1^2/4 + \sigma_2^2/4 + \rho\sigma_1\sigma_2/2 \\ &= \sigma^2(1 + \rho)/2 \text{ if } \sigma_1 = \sigma_2 = \sigma,\end{aligned}\tag{2.11}$$

where the second equality assumes that both assets have the same standard deviation.

If the assets are uncorrelated ($\rho = 0$), then the variance of this portfolio is half that of the assets—which demonstrates the importance of diversification. This effect is even stronger when the correlation is negative: with $\rho = -1$ the portfolio variance is actually zero, which we call *hedging*. In contrast, with a high correlation, the benefit from diversification is smaller (and zero when the correlation is perfect, $\rho = 1$). See Figure 2.3 for an illustration.

Example 2.8 (Diversification) If $\sigma = 16\%$ (so $\sigma^2 = 256/100^2$) and $\rho = 0.5$, then (2.11) gives $\text{Var}(R_p) = 192/100^2$ and thus $\text{Std}(R_p) \approx 13.9\%$.

Empirical Example 2.9 Table 2.1 provides an empirical example of the correlations between major asset classes.

	Small growth	Small value	Large growth	Large value	Bonds	T-bills
Small growth	1.00	0.86	0.80	0.69	-0.09	-0.03
Small value	0.86	1.00	0.71	0.85	-0.10	-0.03
Large growth	0.80	0.71	1.00	0.76	0.05	0.04
Large value	0.69	0.85	0.76	1.00	-0.06	0.04
Bonds	-0.09	-0.10	0.05	-0.06	1.00	0.20
T-bills	-0.03	-0.03	0.04	0.04	0.20	1.00

Table 2.1: Correlations of asset class returns, US, monthly returns, 1985:01-2024:12

2.3.2 Diversification: The Number of Assets

In order to see the importance of mixing many assets in the portfolio, we will consider equally weighted portfolios of n assets ($w_i = 1/n$), to focus on the basic idea. Clearly, there are other (not equally weighted) portfolios with even lower variance.

The variance of an equally weighted ($w_i = 1/n$ so $w_i^2 = 1/n^2$) portfolio is

$$\text{Var}(R_p) = (\bar{\sigma}^2 - \bar{\sigma}_{ij})/n + \bar{\sigma}_{ij}, \quad (2.12)$$

where $\bar{\sigma}^2$ is the average variance (average across the n assets) and $\bar{\sigma}_{ij}$ is the average covariance of two returns. Both can be treated as constants if we pick assets randomly. In case the assets are uncorrelated, (2.12) shows that the portfolio variance goes to zero as the number of assets (included in the portfolio) goes to infinity. More realistically, $\bar{\sigma}_{ij}$ is positive. When the portfolio includes many assets, then the average covariance dominates. In the limit (as n goes to infinity), only this non-diversifiable risk matters, $\bar{\sigma}_{ij}$. See Elton, Gruber, Brown, and Goetzmann (2014) 4 for a more detailed discussion.

Example 2.10 (*Variance of portfolio return*) With $\bar{\sigma}^2 = 256/100^2$ and $\bar{\sigma}_{ij} = 128/100^2$, we get a portfolio variance of $(256, 192, 170.7)/100^2$ for $n = (1, 2, 3)$, and thus portfolio standard deviations of (16%, 13.9%, 13.1%).

Empirical Example 2.11 Figure 2.4 shows an empirical example of what diversification implies. Clearly, the covariances start to dominate as the number of assets in the portfolio increases—and the portfolio variance goes towards the average covariance. Figure 2.5 suggests that the diversification benefits are not constant across time.

Proof of (2.12). Note that $\text{Var}(R_p) = (\mathbf{1}/n)' \Sigma (\mathbf{1}/n)$, where $\mathbf{1}$ is a vector of ones. This is just summing the elements in Σ and dividing by n^2 . In this sum, there are n

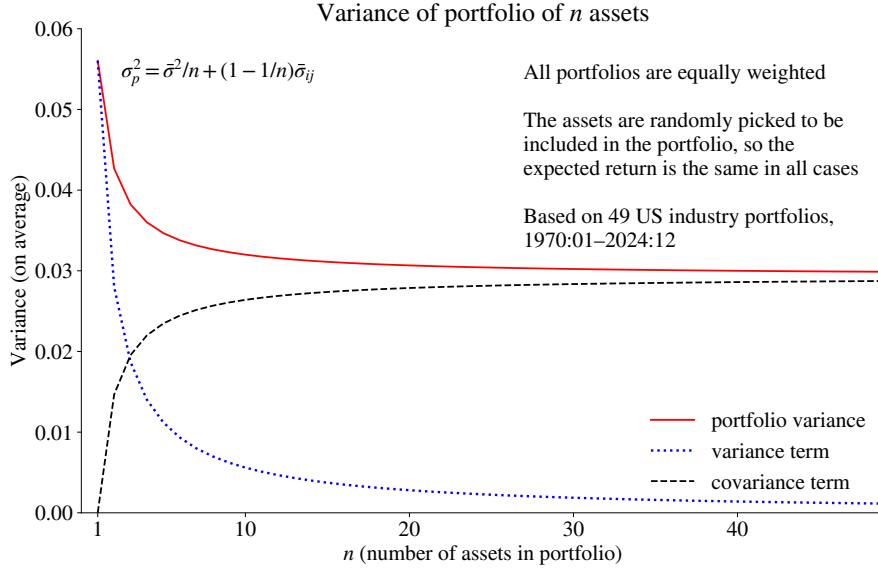


Figure 2.4: Effect of diversification

variances and $n(n - 1)$ covariances. We can thus write the variance as

$$\begin{aligned}\text{Var}(R_p) &= \frac{1}{n} \sum_{i=1}^n \frac{\sigma_i^2}{n} + \frac{n-1}{n} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \frac{\sigma_{ij}}{n(n-1)} \\ &= \bar{\sigma}^2/n + \bar{\sigma}_{ij}(n-1)/n,\end{aligned}$$

which can be rearranged as (2.12). \square

Remark 2.12 (*On negative covariances in (2.12)**) Formally, it can be shown that $\bar{\sigma}_{ij}$ must be non-negative as $n \rightarrow \infty$. It is simply not possible to construct a very large number of random variables that are, on average, negatively correlated with each other. In (2.12) this manifests itself in that $\bar{\sigma}_{ij} < 0$ would give a negative portfolio variance as n increases.

2.4 Covariances Do Matter

This section will (once again) illustrate the importance of covariances for the portfolio variance, but from another perspective. We relax the assumption of equal weights, but consider only small changes to an existing portfolio.

Suppose we are initially invested in a portfolio p (with portfolio weights of the risky assets in the vector v risky assets and $1 - v'\mathbf{1}$ in the risk-free). We now consider a small increase (δ) of the portfolio weight of asset i financed by borrowing at the risk-free rate.

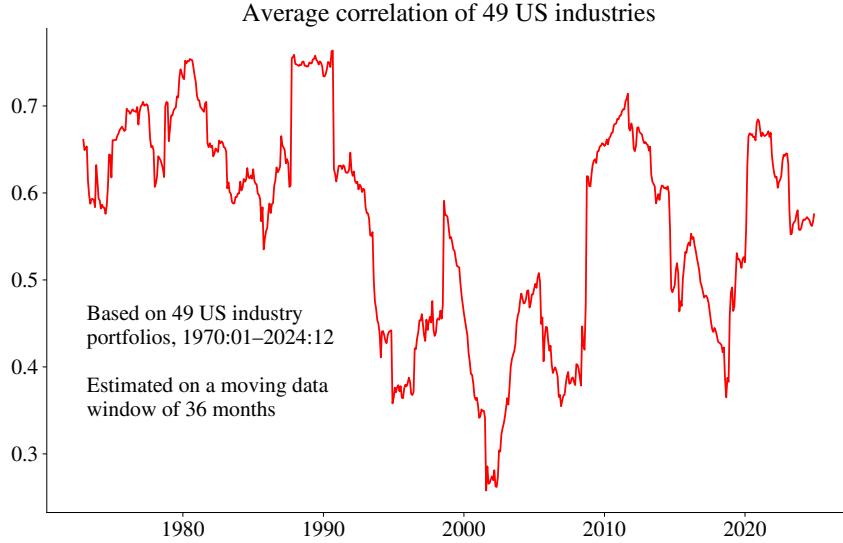


Figure 2.5: Time-varying correlations

The portfolio return of the new portfolio (q) would then be

$$R_q = R_p + \delta R_i^e, \quad (2.13)$$

The *incremental return*, δR_i^e , is just δ times the excess return on the asset i . This is straightforward since we have increased the exposure to asset i and financed it with borrowing at the risk-free rate.

The portfolio variance is

$$\sigma_q^2 = \sigma_p^2 + \underbrace{\delta^2 \sigma_i^2 + 2\delta \sigma_{ip}}_{\text{incremental variance}}, \quad (2.14)$$

where σ_{ip} is the covariance of our portfolio p with asset i .

The *incremental variance* is $\delta^2 \sigma_i^2 + 2\delta \sigma_{ip}$, so it depends on the variance of asset i and on how it correlates with our current portfolio p . For small values of δ (say, $\delta = 5\%$) the *covariance effect might dominate* (since δ^2 decreases very quickly). Conversely, adding a small amount of an uncorrelated asset ($\sigma_{ip} = 0$) to your portfolio does not change the portfolio variance much at all. See Figure 2.6 for an illustration.

Example 2.13 (of (2.14)) The easiest case is when σ_p and σ_i both equal 1, so σ_{ip} equals the correlation ρ . Then, the incremental variance is $\delta^2 + 2\delta\rho$. For $\delta = 0.05$ we have 0.0025 + 0.08 when $\rho = 0.8$ so the covariance effect is 32 times larger than the effect of

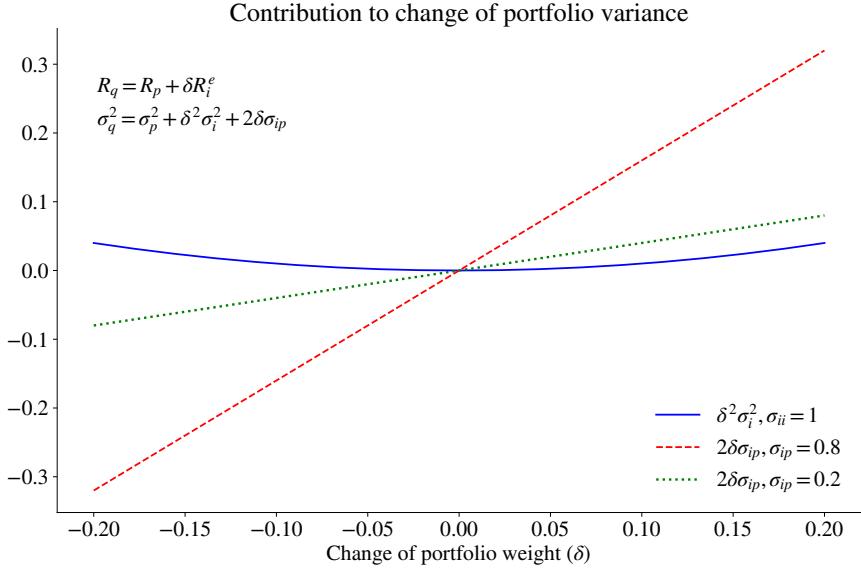


Figure 2.6: The effect of a portfolio change on the variance

$\delta^2 \sigma_i^2$. Conversely, with $\rho = 0$ the covariance effect is zero. See also Figure 2.6.

2.5 Appendix – Statistics*

This appendix first summarizes some mathematical statistics required for the financial models discussed in the text. Towards the end, it briefly addresses topics in estimation and testing.

2.5.1 The Distribution of a Random Variable

The distribution of a random variable x represents the probabilities of its possible values. See Figure 2.7 for illustrations of the (discrete) distribution of a binomial variable and of several different (continuous) normal distributions, often denoted $N(\mu, \sigma^2)$ to indicate the mean and variance.

The probability that $x \leq B$ is given by the *cumulative distribution function*, $\text{cdf}(B)$. For instance, if x has a $N(0, 1)$ distribution, then $\Pr(x \leq -1.645) = 0.05$ and $\Pr(x \leq 0) = 0.5$. See Figure 2.8 for an illustration.

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

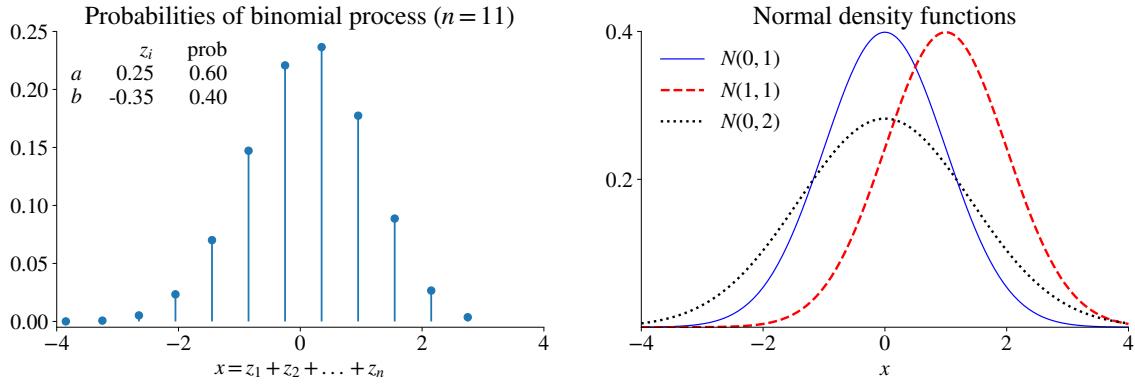


Figure 2.7: Density functions for a binomial and several normal distributions

2.5.2 Expected Value and Variance

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

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

for a discrete and continuous random variable, respectively. For the former, π_s denotes the probability of outcome x_s , and for the latter $f(x)$ represents the probability density function (pdf). The probabilities must sum to unity; therefore $\sum_{s=1}^S \pi_s = 1$ and $\int f(x) dx = 1$. Again, see Figure 2.7. The expected value is sometimes denoted μ .

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

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

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

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

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

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

$$\begin{aligned} E(a + bx) &= a + b E x \\ \text{Var}(a + bx) &= b^2 \text{Var}(x) \text{ or } \text{Std}(a + bx) = |b| \text{Std}(x). \end{aligned}$$

Again, consider $E g(x)$ and suppose x depends on a choice variable v , for instance,

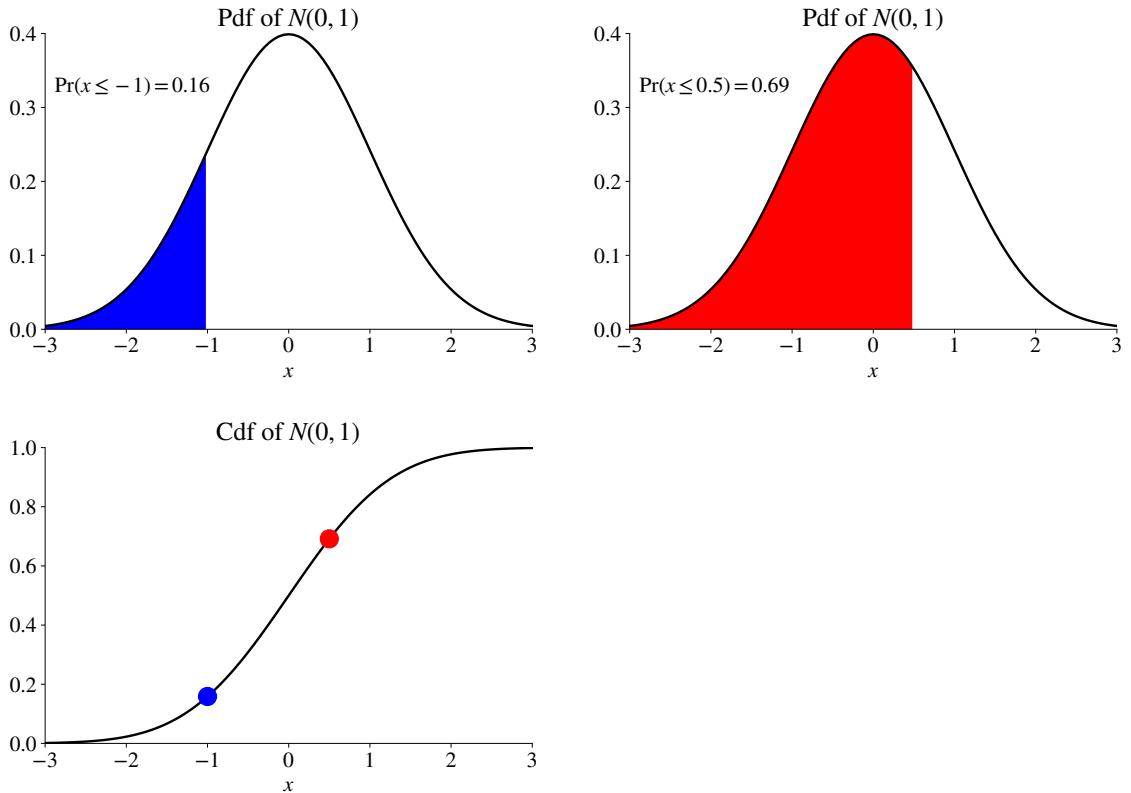


Figure 2.8: Pdf and cdf of $N(0, 1)$

when x is the return of a portfolio of two assets, $vR_1 + (1 - v)R_2$. The derivative of $E g(x)$ is then the expected value of the derivative, so we can interchange the order of E and the derivative

$$\frac{d E g(x)}{dv} = \sum_{s=1}^S \pi_s \frac{dg(x_s)}{dx} \frac{dx_s}{dv} = E \frac{dg(x)}{dv}.$$

A similar expression holds for a continuous distribution.

2.5.3 Expected Value and the Variance of a Vector

There are straightforward extensions to vectors of random variables. For instance, if $x = [x_1, x_2]$ is a vector of the two random variables (returns?) x_1 and x_2 (the subscripts here indicate different variables, not time periods), then the mean of x is a vector of the

means of the two individual returns

$$\mathbb{E}x = \begin{bmatrix} \mathbb{E}x_1 \\ \mathbb{E}x_2 \end{bmatrix}.$$

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

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

Clearly, the variance-covariance matrix is symmetric (the two covariances are the same). The *correlation* of x_1 and x_2 is $\rho_{12} = \text{Cov}(x_1, x_2)/[\text{Std}(x_1) \text{Std}(x_2)]$. See Figure 2.9 for an example of a bivariate distribution.

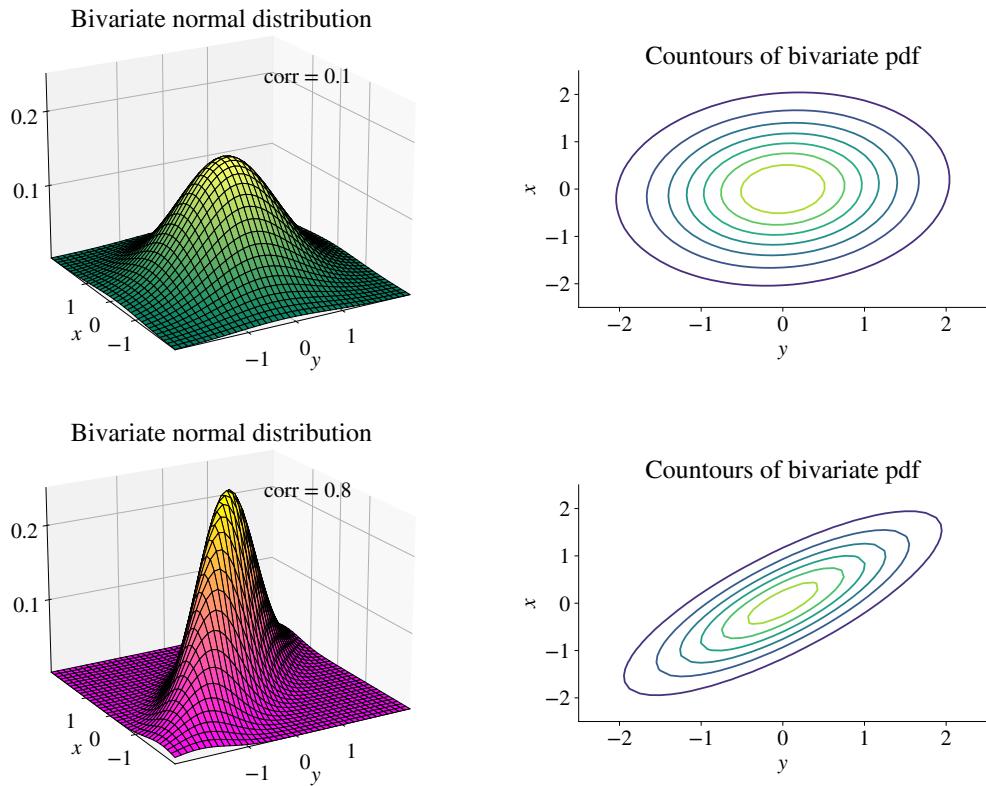


Figure 2.9: Density functions of bivariate normal distributions

2.5.4 Expected Value and the Variance of a Linear Combination

Consider a linear combination of the random variables x_1, \dots, x_n

$$y = \sum_{i=1}^n w_i x_i = w' x.$$

For instance, x could be a vector of portfolio returns and w a vector of portfolio weights.

The expected value and the variance are

$$\begin{aligned} E y &= w' \mu \\ \text{Var}(y) &= w' \Sigma w, \end{aligned}$$

where μ is a vector of average returns and Σ is the $n \times n$ variance-covariance matrix of x .

Also, consider another linear combination, $z = v' x$. Then, the covariance

$$\text{Cov}(z, y) = v' \Sigma w.$$

This could, for instance, be two different portfolios.

Remark 2.14 (*Details on the matrix form*) With two assets, we have the following:

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix},$$

where we use σ_{ii} to indicate σ_i^2 (this helps reading the matrices).

$$\begin{aligned} E y &= w' \mu \\ &= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &= w_1 \mu_1 + w_2 \mu_2. \end{aligned}$$

$$\begin{aligned} \text{Var}(y) &= w' \Sigma w \\ &= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\ &= \begin{bmatrix} w_1 \sigma_{11} + w_2 \sigma_{12} & w_1 \sigma_{12} + w_2 \sigma_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\ &= w_1^2 \sigma_{11} + 2w_1 w_2 \sigma_{12} + w_2^2 \sigma_{22}. \end{aligned}$$

2.5.5 Conditional Moments

Portfolio choice is based on expected future returns, variance and covariances. In general, these represent the beliefs of the investor at the time of investment. Clearly, this means that they may change over time and differ from the properties of historical data. Also, they are to be considered *conditional* in the sense that they refer to the current information/situation—and may therefore differ from *unconditional* moments.

As an example, suppose a random variable (return?) follows an AR(1) process

$$x_{t+1} = (1 - \rho)\mu + \rho x_t + u_{t+1},$$

where u_{t+1} is an iid term (innovation). In this case, the *conditional* expectation and variance are

$$\begin{aligned} E_t x_{t+1} &= (1 - \rho)\mu + \rho x_t, \text{ and} \\ \text{Var}_t(x_{t+1}) &= \text{Var}(u_{t+1}). \end{aligned}$$

This differ from the *unconditional*/long-run values which do not take into consideration the current state and are

$$\begin{aligned} E x_{t+1} &= \mu, \text{ and} \\ \text{Var}(x_{t+1}) &= \text{Var}(u_{t+1})/(1 - \rho^2). \end{aligned}$$

Note that there is no difference between conditional and unconditional moments when x is *iid* (independently and identically distributed), which here means $\rho = 0$. Notice that iid implies, among other things, that x is unpredictable and that the variance is constant over time.

2.5.6 Linear Regressions

Consider the linear regression model

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

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

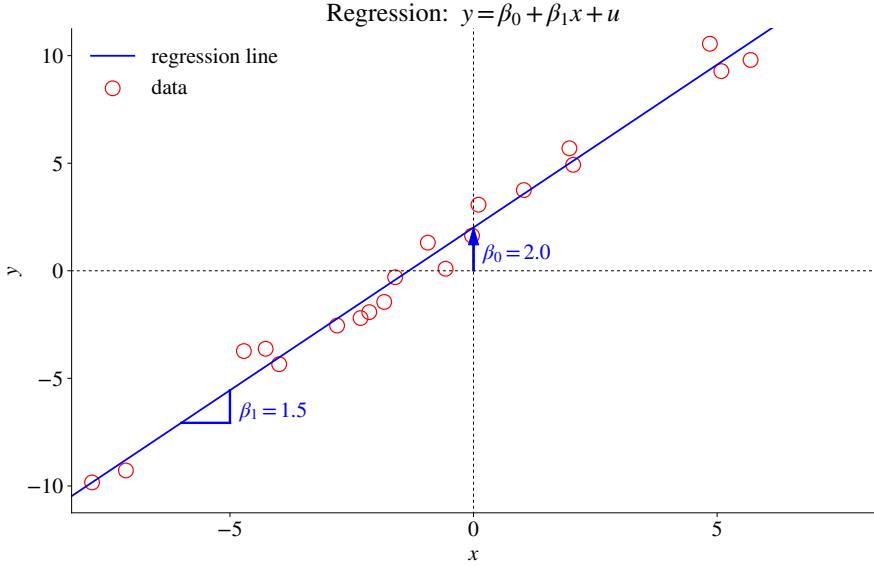


Figure 2.10: Example of OLS

Least squares minimizes the sum of the squared fitted residuals and gives

$$\hat{\beta} = S_{xx}^{-1} \sum_{t=1}^T x_t y_t, \text{ where } S_{xx} = \sum_{t=1}^T x_t x_t'.$$

Clearly, S_{xx} is an $k \times k$ matrix (and is often calculated as $X'X$ if row t of X contains x_t').

If the residuals are iid, then in large samples, we can approximate the distribution of $\hat{\beta}$ as

$$\hat{\beta} \sim N(\beta, S_{xx}^{-1} \sigma^2),$$

where β are the true values and $\sigma^2 = \text{Var}(u_t)$ denotes the variance of the residuals. (In contrast, with autocorrelated residuals or time-varying variance of the residuals, then we have to apply Newey-West's or White's method for approximating the variance-covariance matrix.) Based on this distribution, it is straightforward to test if a single coefficient equals a particular value by a t -test.

2.5.7 t-tests

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

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

To see this, notice that $x - \mu$ has a mean of zero and that x/σ has a standard deviation of unity. A t -distribution is sometimes used instead, since σ has to be estimated. However, with 30 or more data points, the t -distribution and the $N(0, 1)$ are almost indistinguishable.

Chapter 3

The Mean-Variance Frontier

3.1 The Mean-Variance Frontier of Risky Assets

The mean-variance frontier (MVF, see [Markowitz \(1952\)](#)) is based on the idea that the investor seeks high average portfolio returns but dislikes portfolio return variance.

To find the mean-variance frontier, we first have to specify the n -vector of average returns of the investable assets (μ) and their variance-covariance matrix (Σ). As in earlier chapters, the means and the variance-covariance matrix should be interpreted as representing the investor's beliefs, conditional on the information available at the time of the investment.

		$\mu, \%$			Σ, bp		
					A	B	C
A		11.5	166	34	58		
B		9.5	34	64	4		
C		6.0	58	4	100		

Table 3.1: Characteristics of the three assets in some examples. Notice that $\mu, \%$ is the expected return in % (that is, $\times 100$) and Σ, bp is the covariance matrix in basis points (that is, $\times 100^2$).

Example 3.1 (*Mean and Std of a portfolio*) [Table 3.1](#) illustrates a case with three investable assets (A, B and C). The mean returns are given in percentages; thus, 6% should be read as 0.06. In contrast, the variance-covariance matrix is given in terms of basis points (bp, where $1\text{bp} = 1/10000$); thus, 64bp. should be read as 0.0064. The square root of a variance is the standard deviation, so $\sqrt{0.0064} = 0.08$, that is, 8%.

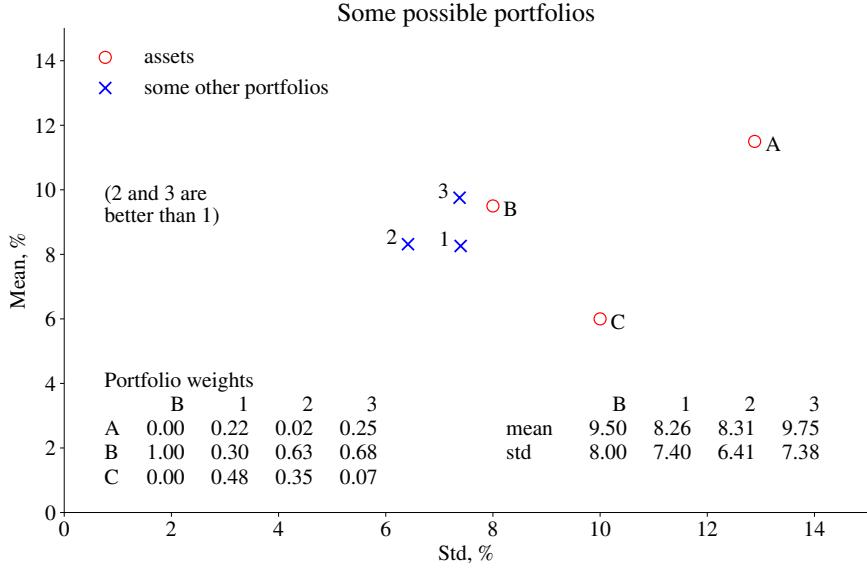


Figure 3.1: Mean vs standard deviation. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

Figure 3.1 illustrates the location of the investable assets (A , B and C) from Table 3.1, as well as some portfolios (1, 2, 3) of them. The figure has the *standard deviation* on the horizontal axis and the *expected return* on the vertical axis. It is reasonable to think that portfolio 3 is better than B (lower volatility and higher expected returns) and also that portfolios 2 and 3 are better than portfolio 1 (lower volatility and higher expected returns, respectively). The mean-variance frontier extends this logic by considering all possible portfolios based on the same investable assets.

To calculate a point on the MVF, we have to find the portfolio that minimizes the portfolio variance, $\text{Var}(R_p)$, for a given expected return, μ^* . The problem is thus

$$\begin{aligned} \min_{w_i} \text{Var}(R_p) \text{ subject to} \\ \mathbb{E} R_p = \mu^* \text{ and } \sum_{i=1}^n w_i = 1. \end{aligned} \tag{3.1}$$

The solution can be found with numerical methods or linear algebra, as shown below.

Remark 3.2 (*Portfolio average and variance*) Let μ be the $n \times 1$ vector of average returns of all n investable assets, Σ the $n \times n$ covariance matrix of the returns and w the $n \times 1$ vector of portfolio weights. The portfolio mean and variance can then be calculated as $\mathbb{E} R_p = w' \mu$ and $\text{Var}(R_p) = w' \Sigma w$.

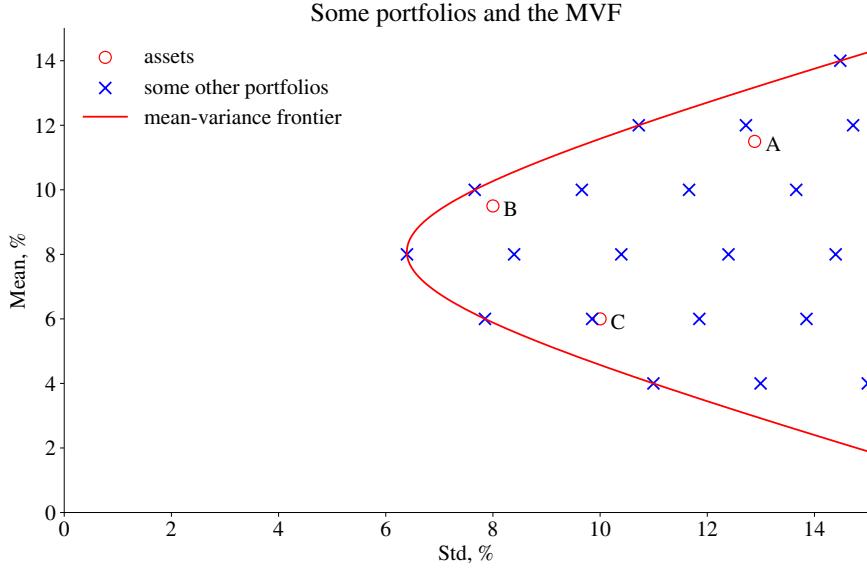


Figure 3.2: Some portfolios and the mean-variance frontier. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

The whole mean-variance frontier is generated by solving this problem for different values of the expected return, μ^* . See Figure 3.2 for an example and comparison with some other portfolios with the same expected return. The *efficient frontier* (EF) is the upper leg of the curve. Reasonably, a portfolio on the lower leg is dominated by one on the upper leg at the same volatility (since it has a higher expected return). Notice that there are no portfolios (based on the given investable assets and their assumed properties μ and Σ) above or to the left of the efficient frontier.

Remark 3.3 (*How many different portfolios are there with $E R_p = \mu^*$?*) With two assets, we require $w\mu_1 + (1 - w)\mu_2 = \mu^*$ and there is only one choice of w that satisfies this (assuming $\mu_1 \neq \mu_2$). Instead with three assets, we require $w_1\mu_1 + w_2\mu_2 + (1 - w_1 - w_2)\mu_3 = \mu^*$ which can hold for a continuum of (w_1, w_2) values.

3.1.1 The Mean Variance Frontier with Portfolio Restrictions

There are sometimes *additional restrictions*, for instance, of no short sales

$$\text{no short sales: } w_i \geq 0, \quad (3.2)$$

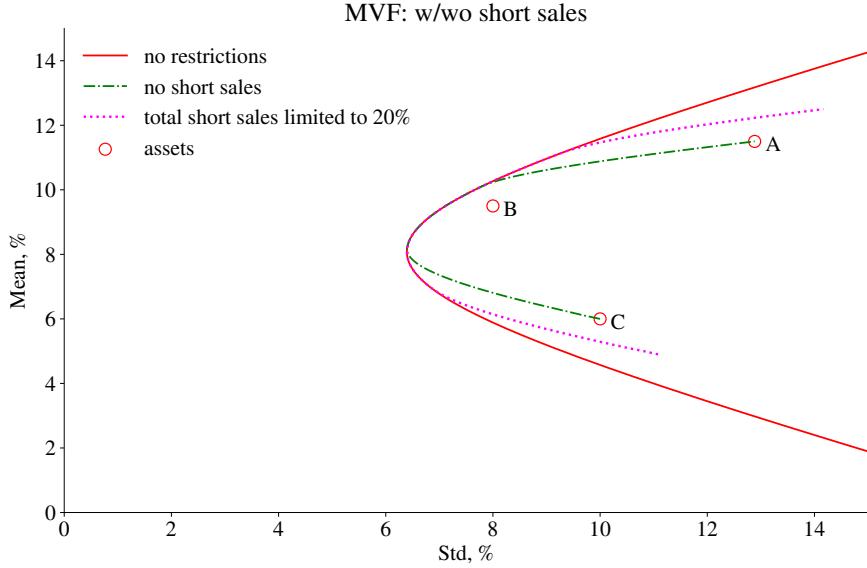


Figure 3.3: Mean-variance frontiers with restrictions. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

In other cases, there are both lower and upper bounds on the weights

$$L_i \leq w_i \leq U_i. \quad (3.3)$$

For instance, mutual funds often have to obey $L_i = 0$ and $U_i = 0.1$.

Funds may also impose restrictions on themselves; for instance, they may allow limited short sales

$$\text{limited total short sales: } \sum_{i=1}^n \min(w_i, 0) \geq Q. \quad (3.4)$$

With such restrictions we typically have to apply some explicit numerical minimization algorithm to find portfolio weights. Algorithms that solve quadratic problems are best suited. See Figures 3.3 –3.4 for an example.

3.1.2 The Mean Variance Frontier with Two Risky Assets

In the case of only two investable assets, the mean-variance frontier can be calculated by simply calculating the mean and variance

$$E R_p = w\mu_1 + (1-w)\mu_2 \quad (3.5)$$

$$\text{Var}(R_p) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\sigma_{12}. \quad (3.6)$$

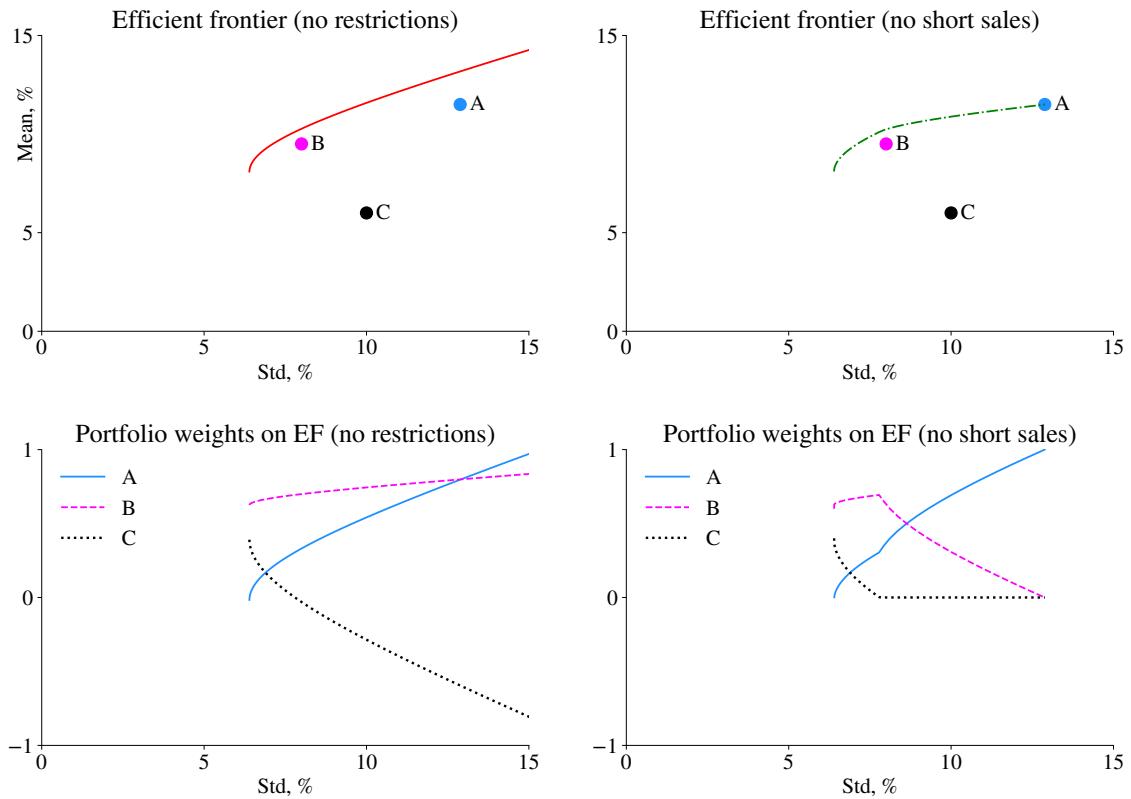


Figure 3.4: Portfolio weights on the efficient frontier. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

at a set of different portfolio weights, for instance, $w = (-1, -0.5, 0, 0.5, 1)$. The reason is that with only two assets, all portfolios of them are on the mean-variance frontier (cf. Remark 3.3). For that reason no explicit minimization is needed. See Figure 3.5 for an example.

3.1.3 The Shape of the Mean-Variance Frontier of Risky Assets

Consider what happens when we *add assets to the investment opportunity set*. The old mean-variance frontier is, of course, still obtainable: we can always put zero weights on the new assets. In most cases, we can do better than that: the mean-variance frontier is shifted to the left (lower volatility at the same expected return). See Figure 3.5 for an example. In this example the new asset is not very attractive (low average returns, high volatility), but it may be useful in a portfolio (diversification, or for shortening).

With intermediate correlations ($-1 < \rho < 1$), the mean-variance frontier is a hyper-

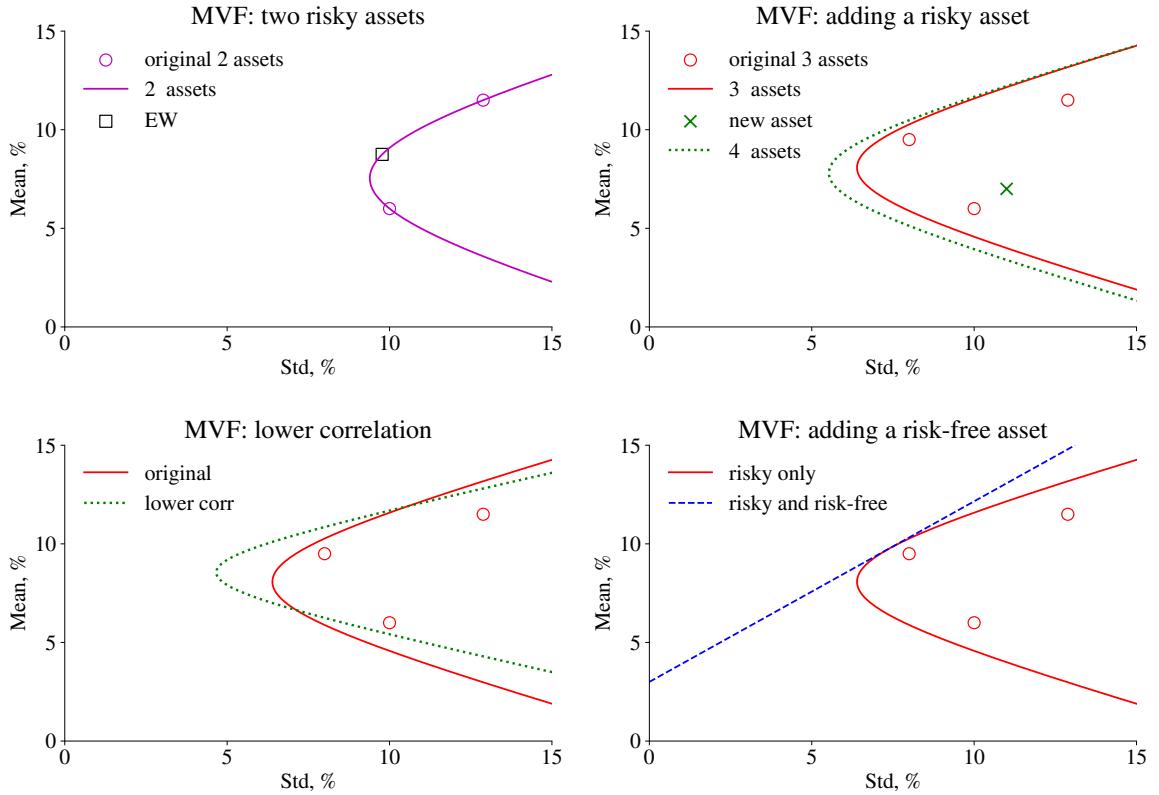


Figure 3.5: The shape of the mean-variance frontiers. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

bola (see Figure 3.5). Notice that the mean–volatility trade-off improves as the correlation decreases: a lower correlation means that we get a lower portfolio standard deviation at the same expected return—at least for the efficient frontier (above the bend).

When we allow investment also in a risk-free asset (to be discussed in detail in a separate section), then the MVF becomes a straight line (again, see Figure 3.5).

Empirical Example 3.4 *Figure 3.6 shows the MVF implied by the sample means and variance-covariance matrix of 10 U.S. industry portfolios. It is therefore an ex post construction, which may (or not) be close the beliefs investors held during the sample period.*

3.1.4 Calculating the Mean-Variance Frontier of Risky Assets

When there are no restrictions on the portfolio weights, then both numerical optimization or some simple matrix algebra can solve the optimization problem. The section demonstrates

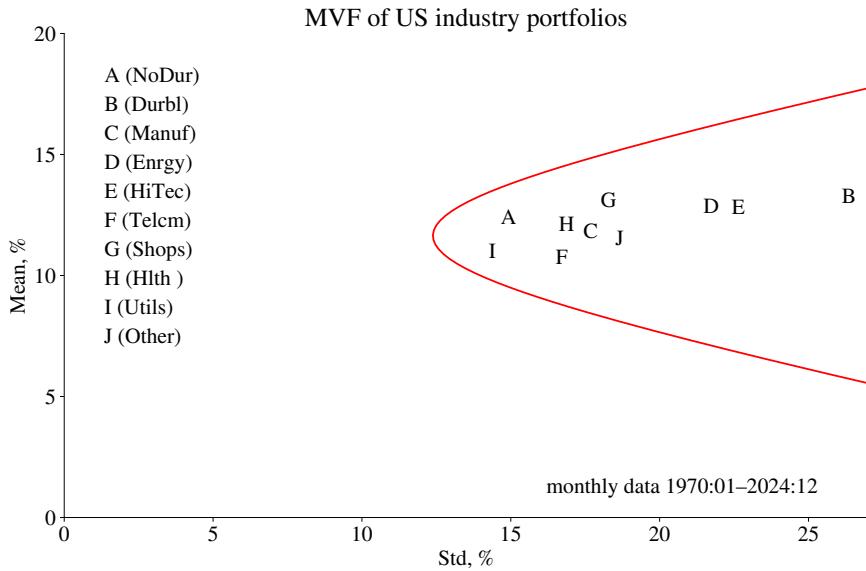


Figure 3.6: Mean-variance frontier from US industry portfolios

the second approach.

The minimization problem (3.1) can be written

$$\begin{aligned} \min_w w' \Sigma w \text{ subject to} \\ w' \mu = \mu^* \text{ and } w' \mathbf{1} = 1, \end{aligned} \tag{3.7}$$

where $\mathbf{1}$ is a vector of n ones (as many as there are assets). Again, μ and Σ summarise the beliefs of the investor, conditional on the information available at the time of the investment.

Remark 3.5 (*First order condition for optimising a differentiable function*). We want to find the value of b in the interval $b_{low} \leq b \leq b_{high}$, which makes the value of the differentiable function $f(b)$ as small (or large) as possible. The answer is b_{low} , b_{high} , or a value of b where $df(b)/db = 0$.

The first order conditions are

$$\begin{bmatrix} \Sigma & \mu & \mathbf{1} \\ \mu' & 0 & 0 \\ \mathbf{1}' & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ \lambda \\ \delta \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mu^* \\ 1 \end{bmatrix}, \tag{3.8}$$

where $\mathbf{0}$ is a vector of n zeros. Solve for the vector (w, λ, δ) and extract the w vector.

Using the solution in $\sqrt{w' \Sigma w}$ gives the standard deviation of a portfolio with expected return μ^* (which should equal $w' \mu$). We can trace out the entire mean-variance frontier, by repeating this calculations for different values of the required return μ^* and then connecting the dots. In the std×mean space, the efficient frontier (the upper part) is *concave*. See, for instance, Figure 3.2.

Proof of (3.8). We set up this as a Lagrangian problem

$$L = (w_1^2 \sigma_{11} + w_2^2 \sigma_{22} + 2w_1 w_2 \sigma_{12})/2 + \lambda(w_1 \mu_1 + w_2 \mu_2 - \mu^*) + \delta(w_1 + w_2 - 1).$$

Dividing the variance by 2 does not affect the solution. (Variances are denoted σ_{ii} in order to facilitate comparison with the matrix expressions.) The first order conditions with respect to the portfolio weights are

$$\text{for } w_1 : w_1 \sigma_{11} + w_2 \sigma_{12} + \lambda \mu_1 + \delta = 0,$$

$$\text{for } w_2 : w_1 \sigma_{12} + w_2 \sigma_{22} + \lambda \mu_2 + \delta = 0.$$

Similarly, the first order conditions with respect to the Lagrange multipliers are

$$\text{for } \lambda : w_1 \mu_1 + w_2 \mu_2 = \mu^*,$$

$$\text{for } \delta : w_1 + w_2 = 1.$$

In matrix notation these first order conditions are

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \mu_1 & 1 \\ \sigma_{12} & \sigma_{22} & \mu_2 & 1 \\ \mu_1 & \mu_2 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \lambda \\ \delta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mu^* \\ 1 \end{bmatrix},$$

which is (3.8). \square

3.1.5 Further Properties of the MVF of Risky Assets*

Remark 3.6 states that the weights for a portfolio on the MVF of risky assets (at a given required return μ^*) can be solved as

$$w = \Sigma^{-1}(\mu \tilde{\lambda} + \mathbf{1} \tilde{\delta}), \quad (3.9)$$

where $(\tilde{\lambda}, \tilde{\delta})$ depend on (μ, Σ, μ^*) . This provides a closed form solution and can also be used to show that $\text{Var}(R_p)$ is a U-shaped parabola, as a function of μ^* . (See below for a proof.)

Remark 3.6 (*Alternative expression for the portfolio weights*) Define the scalars a, b and c as $a = \mu' \Sigma^{-1} \mu$, $b = \mu' \Sigma^{-1} \mathbf{1}$, and $c = \mathbf{1}' \Sigma^{-1} \mathbf{1}$. Then, calculate the scalars (for a

given required return μ^*)

$$\tilde{\lambda} = \frac{c\mu^* - b}{ac - b^2} \text{ and } \tilde{\delta} = \frac{a - b\mu^*}{ac - b^2}$$

to get (3.9). To show this, solve (3.8) and rearrange.

Proof that the MVF is a parabola*. From (3.9), the portfolio variance is

$$\text{Var}(R_p) = w' \Sigma w = (w' \mu \tilde{\lambda} + w' \mathbf{1} \tilde{\delta}) = \mu^* \tilde{\lambda} + \tilde{\delta},$$

where we use the facts that $w' \mu = \mu^*$ and $w' \mathbf{1} = 1$. Use the definitions of $(\tilde{\lambda}, \tilde{\delta})$ in Remark 3.6, complete the square and simplify to get

$$\text{Var}(R_p) = \frac{c(\mu^* - b/c)^2}{ac - b^2} + \frac{1}{c},$$

where (a, b, c) depend on (μ, Σ) , not on μ^* . This is a U -shaped parabola when μ^* is on the horizontal axis and $\text{Var}(R_p)$ on the vertical. The minimum is $1/c$, which is the variance of the global minimum variance portfolio (defined below). See Pennacchi (2008) 2 for a detailed discussion. \square

Another way to construct the MVF of risky assets is to retrace it by *combining any two portfolios on the frontier*. This is sometimes referred to as the “two-fund theorem”, although that should not be confused with the two-fund separation theorem discussed in later chapters. For instance, we can use

$$w_\kappa = \kappa w_g + (1 - \kappa) w_T, \text{ where} \quad (3.10)$$

$$w_g = \Sigma^{-1} \mathbf{1} / \mathbf{1}' \Sigma^{-1} \mathbf{1} \text{ and}$$

$$w_T = \Sigma^{-1} \mu^e / \mathbf{1}' \Sigma^{-1} \mu^e.$$

The first line defines a portfolio in terms of two portfolios (w_g and w_T) that are known to be on the MVF. The first (w_g) is the global minimum variance portfolio (lowest possible variance) and the second (w_T) is the tangency portfolio (to be discussed later on), but we could have used other portfolios. (See below for a proof.)

Proof that w_κ in (3.10) is on the MVF*. Notice that (3.9) can be rewritten as $w = \omega + \phi \mu^*$, where (ω, ϕ) depend on (μ, Σ) , but not on μ^* . (The expressions are $\omega = \Sigma^{-1}(\mathbf{1}a - \mu b)/(ac - b^2)$ and $\phi = \Sigma^{-1}(\mu c - \mathbf{1}b)/(ac - b^2)$.) It follows that a portfolio (p) with weights $w_p = \omega + \phi \mu_p^*$ must be somewhere on the MVF. In particular, consider a combination of two MVF portfolios (w_g and w_T , say), that is, $w_p = \kappa w_g + (1 - \kappa) w_T$. Notice that $w_p = \kappa(\omega + \phi \mu_g^*) + (1 - \kappa)(\omega + \phi \mu_T^*)$, where (μ_g^*, μ_T^*) are the expected returns on the two respective MVF portfolios. Since

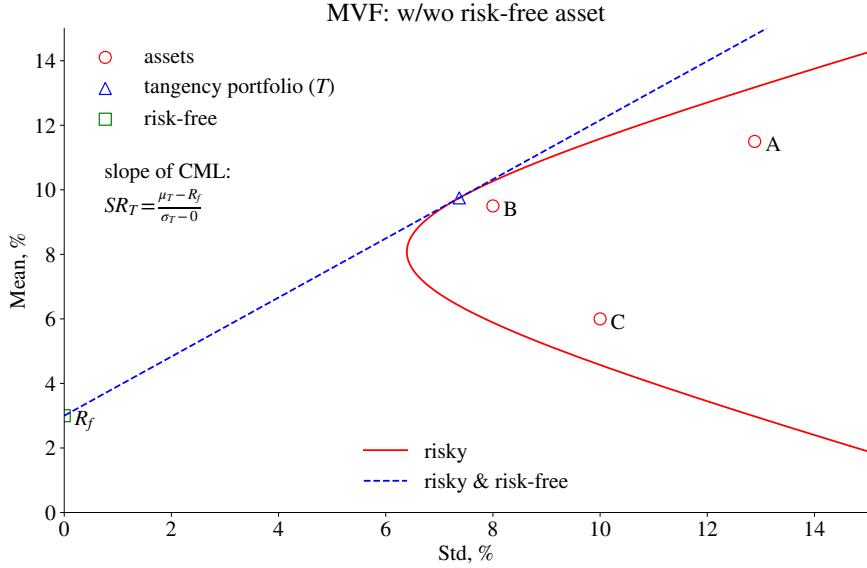


Figure 3.7: Mean-variance frontiers, w/wo risk-free asset. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

$\mu_p^* = \kappa\mu_g^* + (1 - \kappa)\mu_T^*$, the weights for portfolio p can be written as $w_p = \omega + \phi\mu_p^*$. Vary μ_p^* (or equivalently, κ) to trace out the frontier. \square

3.2 The Mean-Variance Frontier of Risk-Free and Risky Assets

We now add a risk-free asset with return R_f and notice that the restriction that $E R_p = \mu^*$ can be written as

$$w'\mu + (1 - w'\mathbf{1})R_f = w'\mu^e + R_f = \mu^*, \quad (3.11)$$

where μ^e the vector of mean excess returns ($\mu - R_f$). Here we use w to denote the vector of portfolio weights on the risky assets only, with $1 - w'\mathbf{1}$ (that is, $1 - \sum_{i=1}^n w_i$) as the weight on the risk-free asset. This means that the requirement that all portfolio weights sum to 1 is automatically satisfied.

The minimization problem (3.1) can now be written

$$\begin{aligned} \min_w w'\Sigma w \text{ subject to} \\ w'\mu^e + R_f = \mu^*. \end{aligned} \quad (3.12)$$

When there are no additional constraints, then we can find an explicit solution. In other cases we need to apply numerical optimization. The weights of the risky assets for

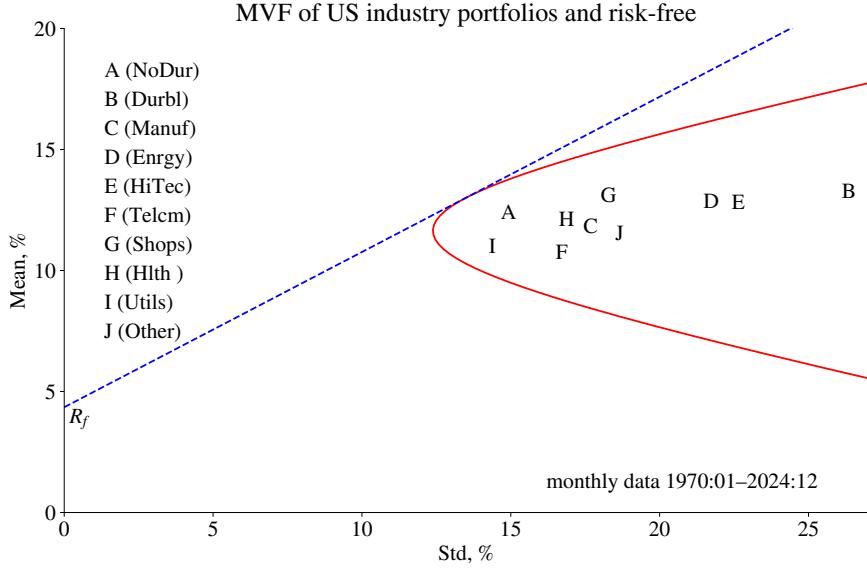


Figure 3.8: Mean-variance frontier from US industry indices

a portfolio on the mean-variance frontier, at a given required return μ^* , are

$$w = \frac{\mu^* - R_f}{\mu^{e'} \Sigma^{-1} \mu^e} \Sigma^{-1} \mu^e. \quad (3.13)$$

As mentioned before, the weight on the risk-free asset is $1 - w' \mathbf{1}$. (See below for a proof.)

Repeating the calculation for different expected return, μ^* , allows us to trace out the entire mean-variance frontier. In the std×mean space, the efficient frontier (the upper part) is just a *line*, called the *Capital Market Line* (CML). See Figure 3.7 for an illustration and Figure 3.8 for an empirical example showing an (ex post) mean-variance frontier from a sample of U.S. data.

Remark 3.7 (*Alternative way to calculate w) The proof of (3.13) shows that we calculate w by solving the following system of equations

$$\begin{bmatrix} \Sigma & \mu^e \\ \mu^{e'} & 0 \end{bmatrix} \begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mu^* - R_f \end{bmatrix}.$$

Remark 3.8 (Minimizing the standard deviation) It can be shown that the solution (3.13) also solves the problem $\min \text{Std}(R_p)$ st $E R_p = \mu^*$ and $\sum_{i=1}^n w_i = 1$.

Remark 3.9 (MVF with different lending and borrowing rates*) Figure 3.9 illustrates the MVF when the borrowing rate is higher than the lending rate. The frontier has three

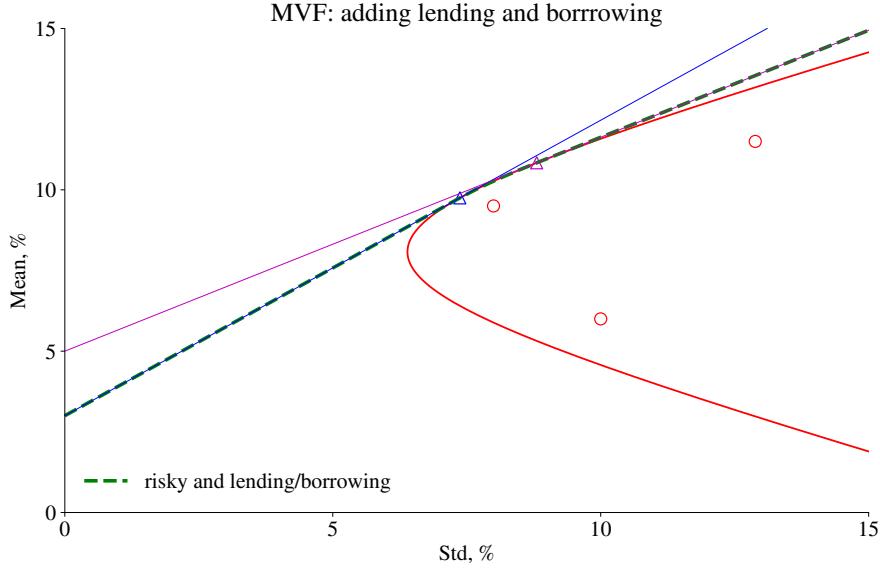


Figure 3.9: Mean-variance frontier with different lending and borrowing rates. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

segments: (1) a straight CLM between the lending rate and the tangency portfolio defined by that rate; (3) a different straight CLM defined by and starting at the tangency portfolio calculated from the borrowing rate; and (2) a segment (arc) in the middle where the investment is in risky assets only. As the lending and borrowing rates gets closer, this converges to the earlier (single) CLM and a single tangency portfolio.

Proof of (3.13). Define the Lagrangian problem

$$L = (w_1^2 \sigma_{11} + w_2^2 \sigma_{22} + 2w_1 w_2 \sigma_{12})/2 + \lambda(w_1 \mu_1^e + w_2 \mu_2^e + R_f - \mu^*).$$

(Variances are denoted σ_{ii} in order to facilitate comparison with the matrix expressions.)

The first order condition with respect to the portfolio weights are

$$\begin{aligned} \text{for } w_1 : w_1 \sigma_{11} + w_2 \sigma_{12} + \lambda \mu_1^e &= 0, \\ \text{for } w_2 : w_1 \sigma_{12} + w_2 \sigma_{22} + \lambda \mu_2^e &= 0. \end{aligned}$$

Similarly, the first order condition with respect to the Lagrange multiplier is

$$w_1 \mu_1^e + w_2 \mu_2^e + R_f = \mu^*.$$

Combine as

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \mu_1^e \\ \sigma_{12} & \sigma_{22} & \mu_2^e \\ \mu_1^e & \mu_2^e & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mu^* - R_f \end{bmatrix},$$

which can be written

$$\begin{bmatrix} \Sigma & \mu^e \\ \mu^{e'} & 0 \end{bmatrix} \begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mu^* - R_f \end{bmatrix}.$$

To simplify further, notice that the first set of equations is $\Sigma w = -\lambda \mu^e$, which can be (partially) solved as $w = -\Sigma^{-1} \lambda \mu^e$. The second set of equations is $\mu^{e'} w = \mu^* - R_f$. Use the (partial solution) of w to write this as $-\mu^{e'} \Sigma^{-1} \lambda \mu^e = \mu^* - R_f$, which can be solved for as $\lambda = -(\mu^* - R_f)/\mu^{e'} \Sigma^{-1} \mu^e$. Finally, using this in the partial solution of w gives (3.13). \square

3.3 The Tangency Portfolio

The mean-variance frontier for risky assets only and the frontier for risky assets plus the risk-free asset are tangent at one point—called the *tangency portfolio*: see Figure 3.7. In this case the portfolio weights from (3.8) and (3.13) coincide. Therefore, the portfolio weights of the risky assets (3.13) must sum to unity (so the weight on the risk-free asset is zero) at this value of the required return, μ^* . This gives the portfolio weights of the tangency portfolio

$$w_T = \frac{\Sigma^{-1} \mu^e}{\mathbf{1}' \Sigma^{-1} \mu^e}. \quad (3.14)$$

Proof of (3.14). Put the sum of the portfolio weights in (3.13) equal to one and solve for the μ^* value where that holds. Use in (3.13). \square

Notice that Capital Market Line (CML) starts at the location $(\sigma, \mu) = (0, R_f)$ and goes through the point (μ_T, σ_T) where the latter are the mean and standard deviation of the tangency portfolio. It is then clear that the slope of the CML, $(\mu_T - R_f)/(\sigma_T - 0)$, represents the *Sharpe ratio of the tangency portfolio*. The line is thus

$$\mathbb{E} R_{opt} = R_f + \sigma_{opt} SR_T. \quad (3.15)$$

Interestingly, the tangency portfolio has the *highest Sharpe ratio of any portfolio* that can be created from the investable assets. See Figure 3.10.

It follows that every portfolio on the CML is a combination of the tangency portfolio and the risk-free asset

$$R_{opt} = v R_T + (1 - v) R_f \quad (3.16)$$

where R_T is the return on the tangency portfolio. Again, see Figure 3.10.

Remark 3.10 (*Maximising the Sharpe ratio directly.) Maximizing $v' \mu^e / \sqrt{v' \Sigma v}$ gives

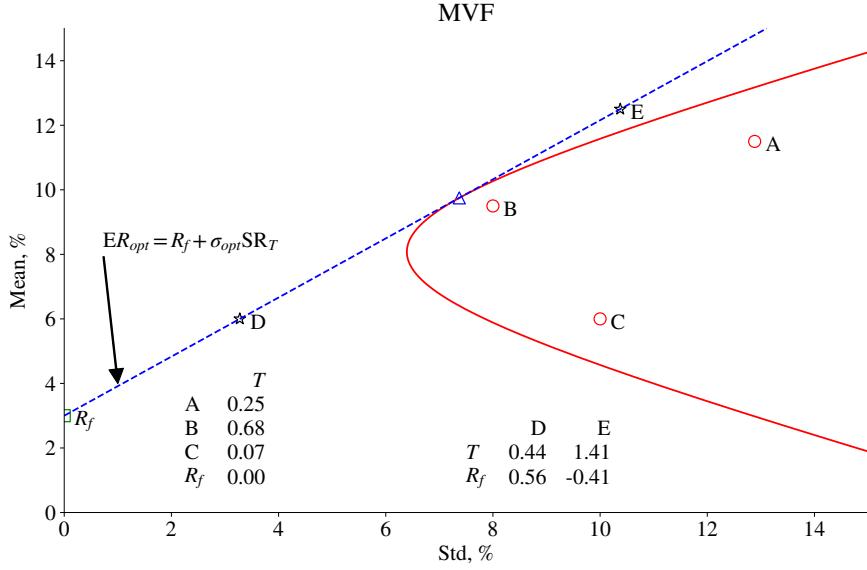


Figure 3.10: Mean-variance frontiers, creating portfolios by combining the tangency portfolio and the risk-free. The properties of the investable assets (A, B, and C) are shown in Table 3.1.

the following n first order conditions

$$\mu^e = \frac{v' \mu^e}{v' \Sigma v} \Sigma v.$$

Setting v equal to the tangency portfolio in (3.14) satisfy those first order conditions. (It helps to notice that $w_T' \mu^e / w_T' \Sigma w_T = \mathbf{1}' \Sigma^{-1} \mu^e$.) To be precise, any proportional scaling of the tangency portfolio ($v = \delta w_T$ where $\delta \neq 0$ is a scalar) will satisfy those first order conditions. This means any point on the capital market line. To find a unique solution, we therefore have to impose at least one restriction, for instance, that the portfolio weights v on the risky assets sum to 1.

Remark 3.11 (Properties of tangency portfolio*) The expected excess return and the variance of the tangency portfolio are $\mu_T^e = \mu^e' \Sigma^{-1} \mu^e / \mathbf{1}' \Sigma^{-1} \mu^e$ and $\text{Var}(R_T^e) = \mu^e' \Sigma^{-1} \mu^e / (\mathbf{1}' \Sigma^{-1} \mu^e)^2$. It follows that $\mu_T^e / \text{Var}(R_T^e) = \mathbf{1}' \Sigma^{-1} \mu^e$ and that the squared Sharpe ratio is $(\mu_T^e)^2 / \text{Var}(R_T^e) = \mu^e' \Sigma^{-1} \mu^e$.

3.3.1 Examples of Tangency Portfolios*

Consider the simple case with two risky assets which are uncorrelated ($\sigma_{12} = 0$). The tangency portfolio (3.14) is then

$$\begin{bmatrix} w_{T,1} \\ w_{T,2} \end{bmatrix} = \frac{1}{\sigma_2^2 \mu_1^e + \sigma_1^2 \mu_2^e} \begin{bmatrix} \sigma_2^2 \mu_1^e \\ \sigma_1^2 \mu_2^e \end{bmatrix}. \quad (3.17)$$

This shows that if both excess returns are positive, then (i) the weight on asset i increases when μ_i^e increases and when σ_{ii} decreases; (ii) both weights are positive. (You may notice that scaling the mean returns and/or the variance-covariance matrix does not matter.)

Example 3.12 (*Tangency portfolio, numerical*) When $(\mu_1^e, \mu_2^e) = (8\%, 5\%)$, the correlation is zero, and $(\sigma_1^2, \sigma_2^2) = (256 \text{ bp}, 144 \text{ bp})$, then (3.17) gives

$$\begin{bmatrix} w_{T,1} \\ w_{T,2} \end{bmatrix} = \begin{bmatrix} 0.47 \\ 0.53 \end{bmatrix}.$$

When μ_1^e increases from 8% to 12%, then we get

$$\begin{bmatrix} w_{T,1} \\ w_{T,2} \end{bmatrix} = \begin{bmatrix} 0.57 \\ 0.43 \end{bmatrix}.$$

Now, consider another simple case, where both variances are the same, but the correlation is non-zero ($\sigma_1 = \sigma_2 = 1$ as a normalization, $\sigma_{12} = \rho$). Then (3.17) becomes

$$\begin{bmatrix} w_{T,1} \\ w_{T,2} \end{bmatrix} = \frac{1}{(\mu_1^e + \mu_2^e)(1 - \rho)} \begin{bmatrix} \mu_1^e - \rho \mu_2^e \\ \mu_2^e - \rho \mu_1^e \end{bmatrix}. \quad (3.18)$$

Results: (i) both weights are positive if the returns are negatively correlated ($\rho < 0$) and both excess returns are positive; (ii) $w_{T,2} < 0$ if $\rho > 0$ and μ_1^e is considerably higher than μ_2^e (so $\mu_2^e < \rho \mu_1^e$). The intuition for the first result is that a negative correlation means that the assets “hedge” each other (even better than diversification), so the investor would like to hold both of them to reduce the overall risk. (Unfortunately, most assets tend to be positively correlated.) The intuition for the second result is that a positive correlation reduces the gain from holding both assets (they don’t hedge each other, and there is relatively little diversification to be gained if the correlation is high). On top of this, asset 1 gives a higher expected return, so it is optimal to sell asset 2 short (essentially a risky “loan” which allows the investor to buy more of asset 1).

Example 3.13 (*Tangency portfolio, numerical*) In the case of (3.18) with $(\mu_1^e, \mu_2^e) = (8\%, 5\%)$, and $\rho = -0.8$ we get

$$\begin{bmatrix} w_{T,1} \\ w_{T,2} \end{bmatrix} = \begin{bmatrix} 0.51 \\ 0.49 \end{bmatrix}.$$

If, instead, $\rho = 0.8$, then

$$\begin{bmatrix} w_{T,1} \\ w_{T,2} \end{bmatrix} = \begin{bmatrix} 1.54 \\ -0.54 \end{bmatrix}.$$

3.4 Appendix – Calculus*

The following derivatives (with respect to x) are often used in this text

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

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

Example 3.14 (*Derivative of power function*) $3x^2 + 7x$ has the derivative $6x + 7$ which is -5 at $x = -2$ and 13 at $x = 1$.

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

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

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

Example 3.15 (*Chain rule*) Let $g(u) = u^2$ and $u = f(x) = 2 - 3x$, so we are considering

the composite function $(2 - 3x)^2$. We then get

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

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

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

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

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

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

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

Example 3.16 (Gradient) For the function $f(x) = (x_1 - 2)^2 + (4x_2 + 3)^2$, the gradient is

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

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

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

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

Example 3.17 (Hessian) Using the same function as in Example 3.16, we get

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

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

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

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

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

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

See Figure 3.11 for an illustration.

Example 3.18 (Taylor approximation) Let $f(x) = \ln x$ and consider $(a, b) = (1, 1.2)$, so $f(1.2) \approx 0 + 1 \times 0.2 = 0.2$, when the true value is approximately 0.18. Instead, with $b = 2$, we get the approximation 1 and the true value around 0.69, so the error is considerable.

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

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

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

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

but where the c vector might differ across the various functions. Again, see Figure 3.11 for an illustration.

Example 3.19 (Mean-value theorem) Let $f(x) = \ln x$ and consider $(a, b) = (1, 2)$. With $c \approx 1.443$, we have $0.693 \approx 0 + \frac{1}{1.443}(2 - 1)$.

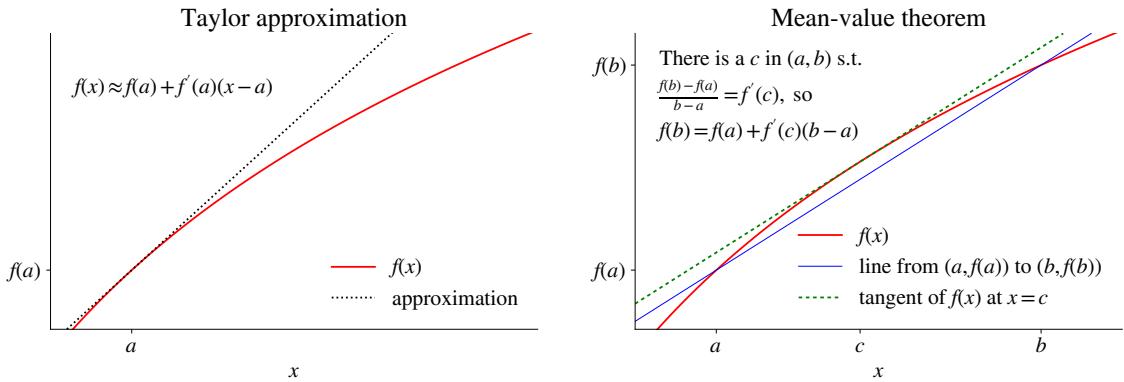


Figure 3.11: Illustration of a Taylor approximation and the mean-value theorem.

3.5 Appendix – Optimization*

Remark 3.20 (*First order condition for optimising a differentiable function*). We want to find the value of b in the interval $b_{low} \leq b \leq b_{high}$, which makes the value of the differentiable function $f(b)$ as small as possible (a minimization problem). The answer is b_{low} , b_{high} , or a value of b where $df(b)/db = 0$. The latter is a necessary and sufficient condition for an unconstrained problem where $f(b)$ is convex. (If the function is twice differentiable, then convexity means that $f''(b) \geq 0$.) A maximization problem, except that we rather want $f(b)$ to be concave ($f''(b) \leq 0$).

Suppose we want to *minimize* the loss function

$$L = (4y + 3)^2$$

then we have to find the value of y that satisfy the *first order condition*

$$0 = dL/dy = 8(4y + 3),$$

which requires $y = -3/4$. Notice that a *maximization problem* has the same type of first order conditions.

Instead, consider a loss function that depends on both x and y

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

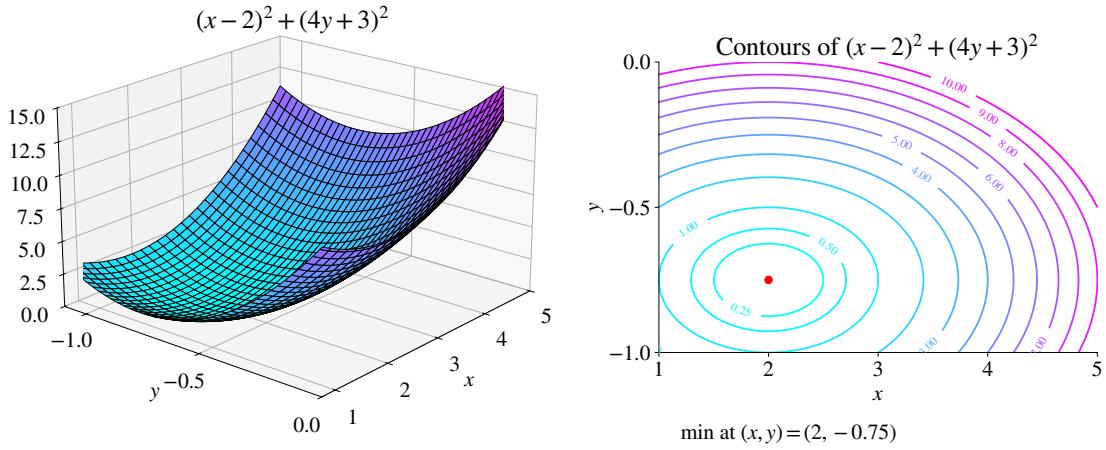


Figure 3.12: Minimization problem

In this case, the first order conditions are

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

which clearly requires $x = 2$ and $y = -3/4$. In this particular case, the first order condition with respect to x does not depend on y , but that is not a general property. See Figure 3.12 for the surface of the loss function and the contours. Also, in this case, there is a unique solution—but in more complicated problems, the first order conditions could be satisfied at different values of x and y .

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

$$x + 2y = 3,$$

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

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

with first order condition

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

which requires $y = -1/2$, which by implies $x = 4$. (We could equally well have substituted for y). This is also the unique solution. See Figure 3.13. This is an easy way

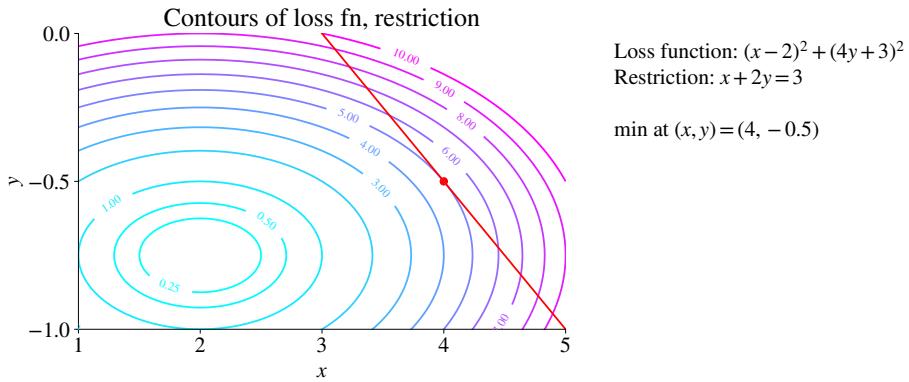


Figure 3.13: Minimization problem with restriction

to eliminate an equality restriction.

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

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

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

The first order conditions are now

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

These are three equations in three unknowns (x, y, λ) which can be solved as $(x, y, \lambda) = (4, -1/2, -4)$.

Remark 3.21 *The three equations are linear, so we could rewrite them on matrix form as*

$$\begin{bmatrix} 2 & 0 & 1 \\ 0 & 32 & 2 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ \lambda \end{bmatrix} = \begin{bmatrix} 4 \\ -24 \\ 3 \end{bmatrix},$$

which is easily solved.

Chapter 4

The Inputs to MV Calculations

Mean-variance (MV) analysis and portfolio choice depend on assumptions regarding average returns and the variance-covariance matrix. This means that the moments (means, variances and covariances) are *conditional* on the information available at the time of portfolio formation, depend on the *investment horizon* and that they may *change over time*.

This chapter discusses how estimates on historical data can help in forming such assumptions, although judgemental adjustments are likely to be made.

The *conditional* nature of the moments distinguish them different from traditional sample estimates. To illustrate, consider the definition

$$R_{t+1} = E_t R_{t+1} + \varepsilon_{t+1}, \quad (4.1)$$

where R_{t+1} is the return over the investment horizon, $E_t R_{t+1}$ the forecast based on information when then portfolio is formed in period t , and ε_{t+1} is the unforecasted part of the return (news, surprise).

A traditional sample estimate of the variance measures the historical variance in R_t . In contrast, the portfolio formation is based on the variance of the forecast error, ε_t . For most assets, returns are difficult to forecast; hence, the difference between the two measures of variance is small. For instance, if the forecasting model has a coefficient of determination (“ R^2 ”) of 0.05, which seems to be close to the upper limits of most return forecasting models, then the variance of ε_t is 0.95 times the variance of R_t . In this situation, the sample variance of R_t might be a good approximation. In contrast, for the risk-free rate, the conditional variance is zero, while sample variance is not (albeit small).

It is also important to consider *time-variation* in the moments. In particular, variances and covariances have considerable (predictable) movements, which motivates using some kind of time-series method for estimation. In addition, *sample estimates can be noisy*,

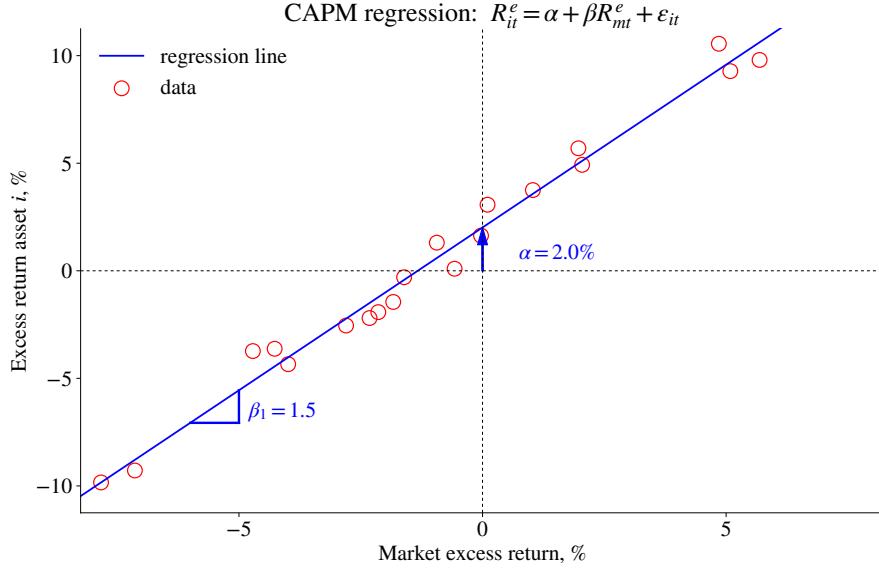


Figure 4.1: CAPM regression

especially when the sample is small, which may motivate forming a compromise between the sample estimates and a priori information (“shrinkage”).

4.1 The Market Model: Betas

The beta (slope coefficient) from the *market model* is often used to describe the cyclicalities of an asset. It is useful as a statistical description of the returns. The regression is

$$R_{it}^e = \alpha_i + \beta_i R_{mt}^e + \varepsilon_{it}, \text{ where} \quad (4.2)$$

$$\mathbb{E} \varepsilon_{it} = 0, \text{ Cov}(\varepsilon_{it}, R_{mt}^e) = 0.$$

Here, R_{it}^e is the excess return on asset i in period t , while R_{mt}^e is the market excess return in the same period. The regression is done on time series (R_{it}^e and R_{mt}^e for $t = 1, 2, \dots, T$). As usual, the regression slope is $\beta_i = \sigma_{im}/\sigma_m^2$. This regression may use the excess returns as indicated above, or the net returns. The two assumptions (the residual has a zero mean and is uncorrelated with the regressor) are standard in regression analysis. See Figures 4.1 for an illustration.

Empirical Example 4.1 See Figures 4.2–4.3 for results on U.S. industry portfolios. See also Table 4.1 for some alternative assets. In general, we need to move away from equities

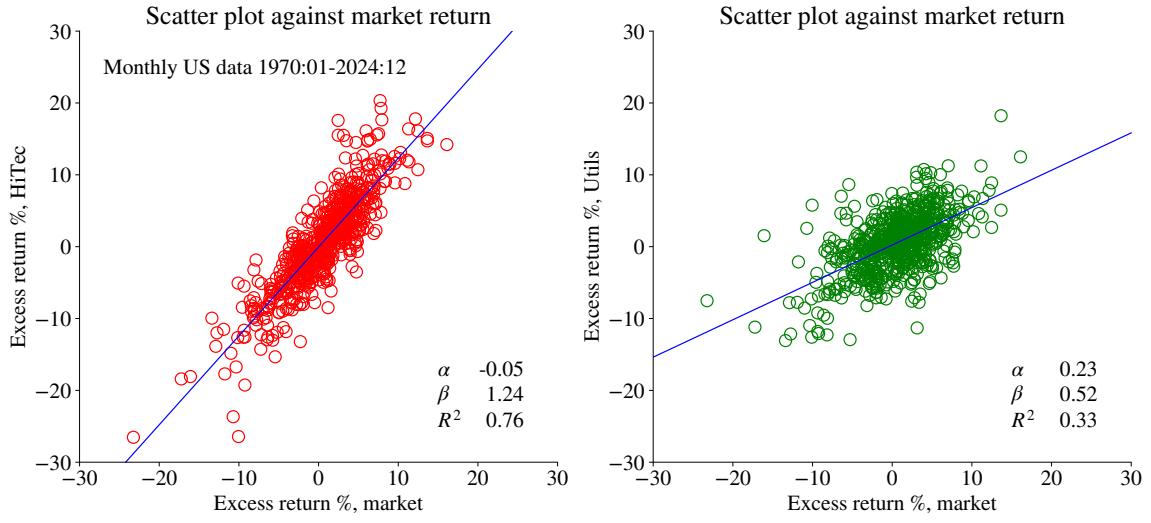


Figure 4.2: Scatter plot against market return

to get β values close to zero (or negative).

The beta of a portfolio is the portfolio of betas. That is, for a portfolio of assets, the beta is

$$\beta_p = \sum_{i=1}^n w_i \beta_i. \quad (4.3)$$

(This follows from the fact that $\text{Cov}(w_i R_i + w_j R_j, R_m) = w_i \sigma_{im} + w_j \sigma_{jm}$. Dividing by σ_m^2 gives the betas.)

We will later discuss how the market model can help in estimating the variance-covariance matrix of the assets. It is also clear that the β values can be useful in *portfolio formation*. For instance, suppose we want to combine assets 1 and 2 in such a way that our overall position is market neutral ($\beta_p = 0$). This can be done by choosing the portfolio weights $(w_1, w_2) = (1, -\beta_1/\beta_2)$, combined with a position in the risk-free rate. Also, if the investor wants a portfolio with a particular beta (β_q), then this can be achieved by investing β_q in the market portfolio and $1 - \beta_q$ in the risk-free asset.

The result in (4.3) also shows that the (value weighted) β of all assets must equal 1. (This follows from the fact that the value weighted portfolio of all assets equals the market portfolio—and regressing the market on itself must give a slope of 1.)

Remark 4.2 (*Market indices I*) A market index I_t is calculated as

$$I_t = (1 + R_{mt})I_{t-1}, \text{ where } R_{mt} = \sum_{i=1}^n w_{it} R_{it},$$

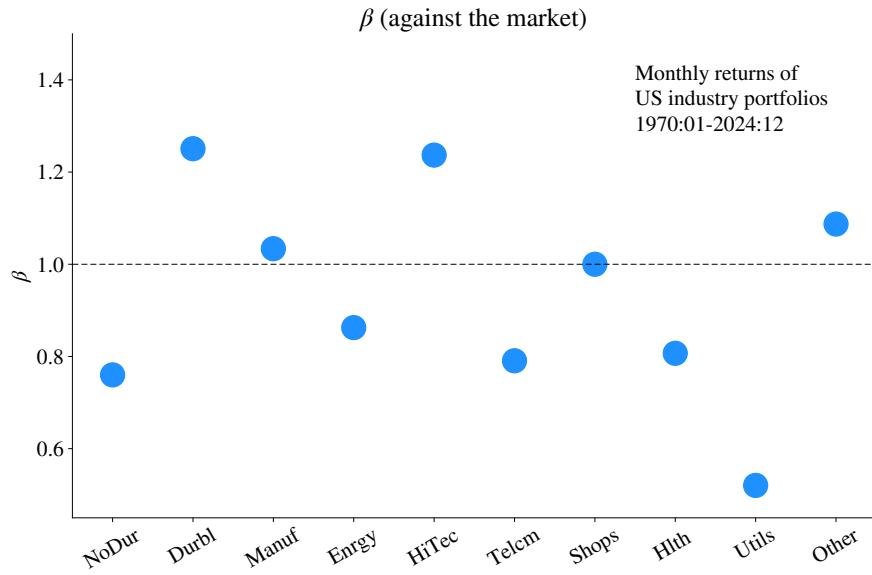


Figure 4.3: β s of US industry portfolios

where i denotes the n different components/assets (for instance, stocks) of the index. This is a capital weighted return index if (a) R_{it} is the net return on holding asset i between $t - 1$ and t ; and (b) w_{it} is the market capitalization of asset i relative to the total market capitalization of all n assets, measured at the end of period $t - 1$. Most of the important indices are of this sort. Instead, if R_{it} only includes the capital gain of holding asset i , then the index is a price index. In other cases, the weights may reflect the market capitalization of the floats (those shares that are actively traded). In yet other cases the weights are the same across the assets (an equally weighted index).

Remark 4.3 (Market indices II*) Dow Jones Industrial Average and Nikkei 225 have very special weights. In practice, these two indices are just the average prices of all (30 or 225) stocks in the index. This means that the portfolio weights are proportional to the stock price.

Remark 4.4 (Market indices III*) More recently, a large number of alternative indices have been introduced, for instance of (a) “sustainable” companies (DJSI); (b) fundamentally weighted indices (weights based on sales, earnings or dividends); (c) 1/volatility based indices; (d) performance based indices (large weights on recent winners).

	β
MSCI world	0.93
CT hedge funds	0.26
Global govt bonds	0.07
Gold	0.02
Oil	0.26

Table 4.1: β against the U.S. equity market, monthly returns, 1993:12-2024:12.

4.1.1 Estimating Historical Beta: OLS and Other Approaches

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

$$\beta = \eta \hat{\beta}_{OLS} + (1 - \eta)1. \quad (4.4)$$

This could be motivated by empirical findings (Blume (1975)) or by a Bayesian principle (see Greene (2018) 16). In the latter case, η would be higher if the sample is long and when the fit is good.

To capture time-variation in betas, we could either estimate on a moving data window or apply an exponentially weighted moving average estimate (EWMA). The latter is a weighted OLS where an observation s periods ago gets the weight λ^s where λ is close to one (for instance, 0.95).

Empirical Example 4.5 See Table 4.2 for an evaluation of several methods: most are of the form (4.4), but an EWMA approach is also considered. A negative number indicate that the method is better than OLS.

	OLS adj	OLS adj	OLS adj	1	EWMA
	$0.67\hat{b} + 0.33$	$0.5\hat{b} + 0.5$	$0.33\hat{b} + 0.67$	1	$0.5\hat{b} + 0.5$
error in β	-6.1	-5.9	-2.9	8.7	-10.2

Table 4.2: Absolute forecast errors of future betas, as a percentage difference to OLS: the average $|\text{next 2 year } \beta - \text{predicted } \beta|$ compared to the results from OLS. A negative number is better performance than OLS. The models are estimated on moving 10-year windows and EWMA uses $\lambda = 0.95$. 25 FF portfolios, monthly data for 1970:01-2024:12.

4.1.2 Fundamental Betas

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

It is often found (see Rosenberg and Guy (1976) for an early paper and Damodaran (2012) for a more recent text) that betas are related to fundamental variables as follows (with signs in parentheses indicating the effect on the beta): dividend payout (-), asset growth (+), leverage (+), liquidity (-), asset size (-), earning variability (+), earnings Beta (slope in earnings regressed on economy wide earnings) (+). Such relations can be used to make an educated guess about the beta of an asset without historical data on the returns—but with data on (at least some) of these fundamental variables.

4.2 Estimation of the Covariance Matrix of the Asset Returns

There are several issues with estimating variance-covariance matrices: (1) the number of parameters increase very quickly as the number of assets increases $(n(n + 1)/2)$ with n assets, for instance 5,050 for 100 assets); (2) there may not be relevant historical data; (3) historical estimates have proven somewhat unreliable for future periods due to small sample issues and time-variation of the parameters.

Remark 4.6 (*Fama-French portfolios*) *The 25 FFF portfolios (used in the examples below) are calculated by annual rebalancing (June/July). The US stock market is divided into 5×5 portfolios as follows. First, split up the stock market into 5 groups based on the book value/market value: put the lowest 20% in the first group, the next 20% in the second group etc. Second, split up the stock market into 5 groups based on size: put the smallest 20% in the first group etc. Then, form portfolios based on the intersections of these groups (also called double sorting). For instance, in Table 4.3 the portfolio in row 2, column 3 (portfolio 8) belong to the 20%-40% largest firms and the 40%-60% firms with the highest book value/market value.*

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

Table 4.3: Numbering of the FF portfolios.

4.3 Covariance Matrix with Time-Varying Parameters

To handle the time-variation, we could apply a simple EWMA estimator (cf. the Risk-Metrics approach of JP Morgan (1996))

$$\mu_{it} = \lambda \mu_{i,t-1} + (1 - \lambda)x_{i,t-1}, \quad (4.5)$$

$$\sigma_{ij,t} = \lambda \sigma_{ij,t-1} + (1 - \lambda)(x_{i,t-1} - \mu_{i,t-1})(x_{j,t-1} - \mu_{j,t-1}), \quad (4.6)$$

with $0 \leq \lambda \leq 1$ and where x_{it} is element i of the vector x_t . The first equation provides a time-varying estimate of the mean and the second equation of the covariance between x_{it} and x_{jt} (set $i = j$) for the variance. In many application on daily data a value of $\lambda \approx 0.94$ is common. For monthly data, often slightly lower values.

4.4 Covariance Matrix with Average Correlations

A commonly used method to address the instability of the variance-covariance matrix, caused by excessive parameters or insufficient data, is to replace the historical correlation with an average historical correlation. To do that, estimate ρ_{ij} on historical data, but use the average estimate $\bar{\rho}$ as the “forecast” of all correlations and calculate adjusted covariances as

$$\sigma_{ij} = \bar{\rho}\sigma_i\sigma_j \text{ for } i \neq j. \quad (4.7)$$

(The variances are not adjusted.) Notice that $\bar{\rho}$ is the average of the $n(n - 1)/2$ elements below (or above) the main diagonal of the correlation matrix.

4.5 Covariance Matrix from a Single-Index Model

The single-index model is another way to reduce the number of parameters that we need to estimate in order to construct the covariance matrix of assets. The model assumes that the co-movement between assets is due to a single common influence (here denoted R_{mt}). This means that we add one assumption to (4.2)

$$\text{Cov}(\varepsilon_{it}, \varepsilon_{jt}) = 0, \quad (4.8)$$

which says that the residuals for different assets are uncorrelated. This means that all comovements of two assets (R_i and R_j , say) are due to movements in the common “index” (R_{mt}).

If (4.2) and (4.8) are true, then the covariance of assets i and j is

$$\sigma_{ij} = \beta_i \beta_j \sigma_{mm} \text{ for } i \neq j, \quad (4.9)$$

where σ_{mm} is the variance of R_m . For instance, when there is too little (relevant) historical data to estimate reliable covariances, then we could first assess the betas of the assets (for instance, based on firm characteristics) and then draw conclusions about the covariances. The variances are often estimated directly from the sample. See Elton, Gruber, Brown, and Goetzmann (2014) 7–8 for more details on index models.

Example 4.7 (*Two assets*) Let $[\beta_1, \beta_2] = [0.9, 1.1]$, $[\text{Var}(\varepsilon_{1t}), \text{Var}(\varepsilon_{2t})] = [100, 25]/100^2$, and $\sigma_{mm} = 225/100^2$. Then

$$\text{Cov}(R_t) \approx \begin{bmatrix} 282.25 & 222.75 \\ 222.75 & 297.25 \end{bmatrix} / 100^2.$$

Proof of (4.9). By using (4.2) and (4.8) and recalling that $\text{Cov}(R_m, \varepsilon_i) = 0$ direct calculations give that the covariance of assets i and j ($i \neq j$) is (recalling also that $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$)

$$\begin{aligned} \sigma_{ij} &= \text{Cov}(R_i, R_j) \\ &= \text{Cov}(\alpha_i + \beta_i R_m + \varepsilon_i, \alpha_j + \beta_j R_m + \varepsilon_j) \\ &= \beta_i \beta_j \sigma_{mm} + 0. \end{aligned}$$

□

4.6 Covariance Matrix from a Multi-Index Model

The multi-index model is an extension of the single-index model

$$R_{it}^e = a_i + b_i' I_t + \varepsilon_{it}, \text{ where} \quad (4.10)$$

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

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

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

$$\sigma_{ij} = b_i' \Omega b_j \text{ for } i \neq j. \quad (4.11)$$

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

4.7 Covariance Matrix From A Shrinkage Estimator

The historical sample covariance matrix, S , can exhibit significant noise in small samples. One way of handling that is to “shrink” the sample covariance matrix towards a target, F , as

$$\Sigma = \delta F + (1 - \delta)S, \text{ where } 0 \leq \delta \leq 1. \quad (4.12)$$

Ledoit and Wolf (2003) suggest an F matrix from the single index model and Ledoit and Wolf (2004) instead suggest an F matrix which implies the same correlations of all assets. See the previous sections for how to construct such F matrices. In both cases, the diagonal elements of F are the same as in S . The articles develop algorithms for calculating an approximately optimal value of δ , which tend to be large in small samples and with crude targets.

Empirical Example 4.8 *Table 4.4 suggest that with the 25 FF assets δ is small for long samples, but may be non-trivial for shorter samples. With more assets, the δ value is likely to be larger. An alternative approach to choose δ is to investigate the performance on earlier samples.*

	Full sample	10-year samples
constant correlation	0.11	0.64
single index model	0.02	0.11

Table 4.4: Shrinkage parameter δ in Ledoit and Wolf's (2003,2004) covariance estimator $\delta F + (1 - \delta)S$. The target matrix F is either a constant correlation covariance matrix or the covariance matrix from a single index model. 25 FF portfolios, monthly data for 1970:01–2024:12. The result for the 10-year samples is the average across moving 10-year data windows.

4.8 An Evaluation of Different Approaches

This section presents a simple assessment the different methods. However, it does not make any general claims as the conclusions depend on (a) which assets; (b) sample period and time horizons; and (c) various modelling parameters.

Empirical Example 4.9 *Table 4.5 present results for the 25 FF portfolios. Several models are estimated on moving 10-year data windows (except the exponentially weighted moving average, EWMA, method which follows (4.5)–(4.6)). We focus on the correlations, since most methods share the same approach for the standard deviations. For each method and period, the implied correlations are calculated and then compared with the actual (realised) correlations for two years after the estimation window. Then, the data window is moved one month and the procedure is repeated. The table show the average (across time and assets) of absolute forecast error of the correlation.*

average corr	1-factor model	3-factor model	shrink to avg corr	shrink to 1-factor	EWMA $\lambda = 0.95$
12.4	40.5	2.4	6.5	1.2	-5.0

Table 4.5: Average absolute forecast errors of future correlations, as a percentage difference to the sample correlation: average $|\text{correlation next 2 years} - \text{predicted correlation}|$ compared to the sample correlation. A negative number is better performance than the sample correlation. All models (except EWMA) are estimated on moving 10-year windows. The shrinkage approach reports results from covariance matrix $= \delta F + (1 - \delta)S$ with δ optimally chosen as in Ledoit and Wolf (2003,2004). 25 FF portfolios, monthly data for 1970:01–2024:12.

4.9 Estimating Expected Returns

A later chapter will discuss return predictability at some length. For now it suffices to say that even the best prediction models have limited performance, often with coefficient of determination (“ R^2 ”) below 5%. In particular, it turns out that it is hard to beat the historical average return as a predictor.

Chapter 5

Portfolio Choice

5.1 Portfolio Choice with MV Preferences

This chapter discusses optimal portfolio choice when the investor has mean-variance (MV) preferences and can invest in both risky assets and a risk-free asset.

As in earlier chapters, the beliefs about the returns (in particular, their average returns μ and variance-covariance matrix Σ) are taken for granted. The analysis is focused on optimal portfolio choice, given those beliefs about the investable assets.

The investor chooses the portfolio weights to maximize expected utility

$$E U(R_p) = E R_p - \frac{k}{2} \text{Var}(R_p), \text{ where} \quad (5.1)$$

$$R_p = v' R + (1 - w' \mathbf{1}) R_f = v' R^e + R_f. \quad (5.2)$$

The k parameter indicates the degree of risk aversion. (Dividing k by 2 is made for convenience: it makes the equation for the optimal portfolio choice look a bit less involved.) The portfolio return in (5.2) assumes investment in risky assets (vector of portfolio weights v) and a risk-free asset (portfolio weight $1 - w' \mathbf{1}$). Notice that this expression automatically imposes the restriction that all portfolio weights sum to one. As before, the expectation and the variance summarise the beliefs of the investor, conditional on the information available at the time of the investment.

The optimisation problem is illustrated in Figure 5.1. This figure shows utility contours which are combinations of $E R_p$ and $\text{Std}(R_p)$ so that expected utility, $E U(R_p)$, in (5.1) is constant. Contours further to the upper left have higher expected utility and are thus preferred. In contrast, the capital market line (CML) shows what is possible to achieve: points on or below the line.

The optimization problem in (5.1) is to move to a point *as far to the upper left as*

	$\mu, \%$	Σ, bp		
		A	B	C
A	11.5	166	34	58
B	9.5	34	64	4
C	6.0	58	4	100

Table 5.1: Characteristics of the three assets in some examples. Notice that $\mu, \%$ is the expected return in % (that is, $\times 100$) and Σ, bp is the covariance matrix in basis points (that is, $\times 100^2$).

possible. Clearly, this will be a point on the CML, so the optimal portfolio is a mix of the risk-free and the tangency portfolio, which will later be referred to as the “two-fund separation theorem”.

Also, see Figure 5.2 for an illustration of how the risk aversion determines *which point on the CML* that is optimal: lower risk aversion (k) will lead the investor to accept more risk in exchange for a higher average return.

Remark 5.1 (*Utility contours**) Let u be a fixed level of expected utility and rewrite (5.1) as $u + \frac{k}{2} \text{Std}(R_p)^2 = E R_p$. For a given value of $\text{Std}(R_p)$ this gives the required $E R_p$ needed to get expected utility u .

5.2 A Single Risky Asset and a Risk-Free Asset

Suppose initially that there is a single investable risky asset. The investor then maximizes (5.1) but where (5.2) simplifies to

$$R_p = vR^e + R_f. \quad (5.3)$$

Remark 5.2 (*Real or nominal returns*) The objective function (5.1) makes more sense if the returns are real, that is, nominal returns minus inflation. During periods (or over horizons) when inflation is fairly stable, the practical difference is small. However, it might matter in other periods. In particular, the risk profile of an asset will then depend on how its return covaries with inflation. For instance, equity returns are often considered to be better hedges against inflation than bond returns.

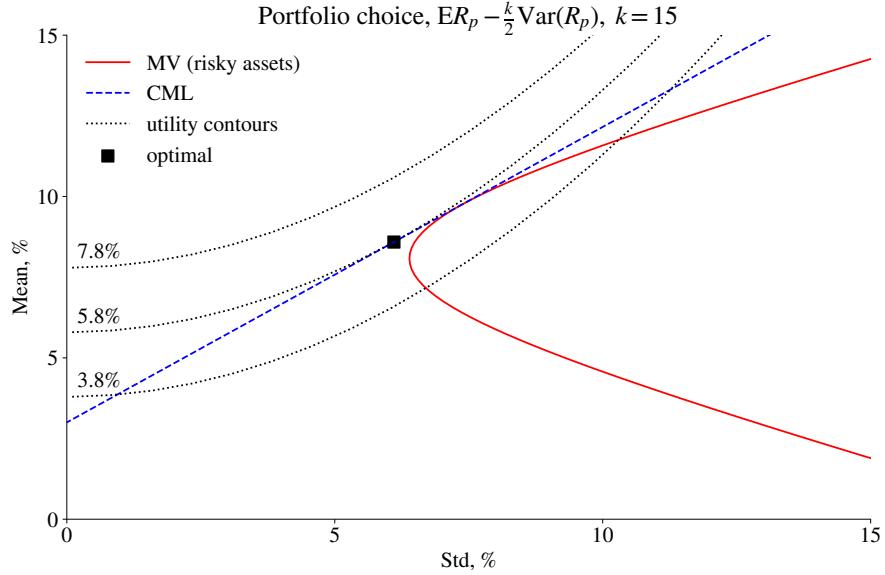


Figure 5.1: Iso-utility curves, mean-variance utility. The calculations use the properties of the assets in Table 5.1.

Use the budget constraint in the objective function to get

$$\begin{aligned} \mathbb{E} U(R_p) &= \mathbb{E}(vR^e + R_f) - \frac{k}{2} \text{Var}(vR^e + R_f) \\ &= v\mu^e + R_f - \frac{k}{2}v^2\sigma^2, \end{aligned} \quad (5.4)$$

where (μ^e, σ^2) denote the investor's beliefs about the mean excess return and variance of the risky asset. In the second equation we use the fact that R_f is known.

The first order condition for an optimum ($d \mathbb{E} U(R_p)/dv = 0$) is

$$\mu^e - k v \sigma^2 = 0, \quad (5.5)$$

which trades off how a marginal increase of v gives a higher expected return but also volatility.

Solve for the optimal portfolio weight of the risky asset as

$$v = \frac{1}{k} \frac{\mu^e}{\sigma^2}. \quad (5.6)$$

The weight on the risky asset is increasing in the expected excess return of the risky asset, but decreasing in the risk aversion and variance. See Figure 5.3, which also illustrates that

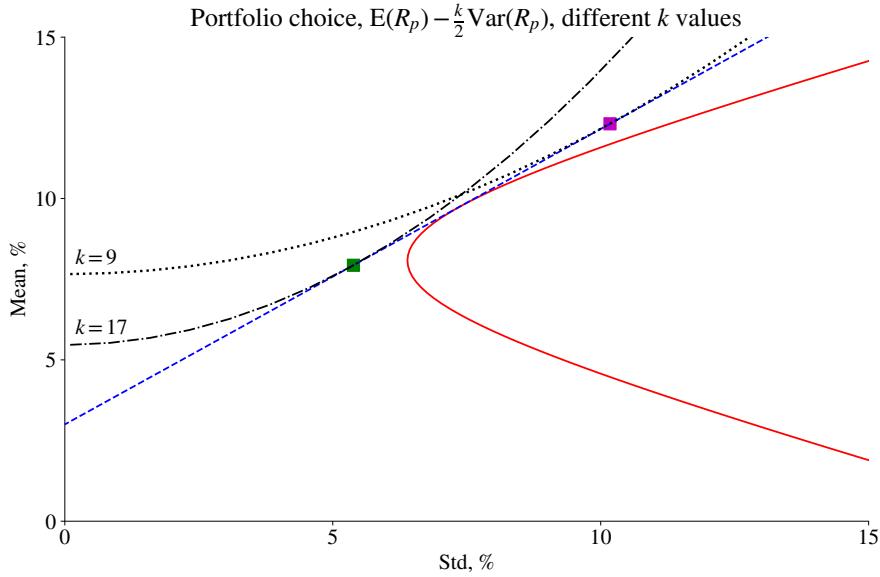


Figure 5.2: Iso-utility curves, mean-variance utility (different risk aversions). The calculations use the properties of the assets in Table 5.1.

the objective function is concave, meaning that the first order condition is both necessary and sufficient. (From (5.5) we also see that the 2nd-order derivative is $-k\sigma^2$, which is negative.)

Example 5.3 (Portfolio choice) If $\mu^e = 6.5\%$, $\sigma_i = 8\%$ and $k = 25$, then $v \approx 0.41$. Instead, with $k = 10$, $v \approx 1.02$.

Remark 5.4 (*Why not use $E R_p - k \text{Std}(R_p)$?) Because it may not have a finite optimum as the objective function is not strictly concave. To see this, consider changing (5.4) to $v\mu^e - k\sqrt{v^2\sigma^2}$ and suppose $v \geq 0$ is optimal. The objective function is then $v(\mu^e - k\sigma)$ where $v = \infty$ is optimal if $\mu^e > k\sigma$. The problem is that both the average returns and the standard deviation are linear in v . Instead, if we were to maximize $v\mu^e - k(v^2\sigma^2)^{0.51}$ (notice the 0.51 instead of 0.5), then the problem is well behaved.

5.3 Several Risky Assets and a Risk-Free Asset

We now consider the case with n investable risky assets and a risk-free asset. Combining (5.1) and (5.2) gives

$$E U(R_p) = v'\mu^e + R_f - \frac{k}{2}v'\Sigma v, \quad (5.7)$$

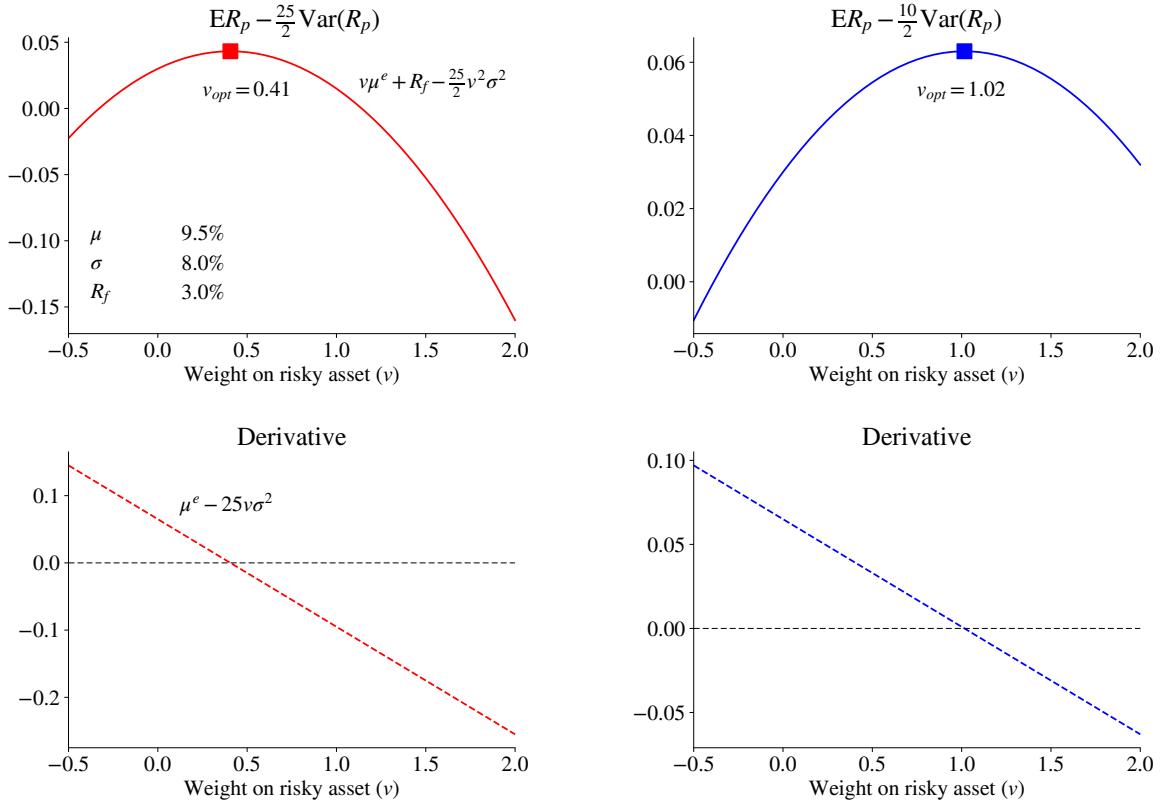


Figure 5.3: Portfolio choice, a single risky and a risk-free asset

where μ^e the n -vector of average excess returns and Σ is the $n \times n$ covariance matrix of the returns. As before, these moments represent the beliefs of the investor, conditional on the information available at the time of the investment.

The first order conditions (for the vector v) are that the partial derivatives with respect to v are zero

$$\mu^e - k\Sigma v = \mathbf{0}, \quad (5.8)$$

which can be solved as

$$v = \Sigma^{-1}\mu^e/k. \quad (5.9)$$

Notice that the weight on the risk-free asset is $1 - v'\mathbf{1}$, where $\mathbf{1}$ is a (column) vector of ones. We will later provide an interpretation of the first order conditions.

Remark 5.5 For two assets, (5.9) can be written

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \frac{1}{k} \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} \mu_1^e \\ \mu_2^e \end{bmatrix},$$

	$\mu, \%$	Σ, bp	
		A	B
A	11.5	166	34
B	9.5	34	64

Table 5.2: Characteristics of the two assets in some examples. Notice that $\mu, \%$ is the expected return in % (that is, $\times 100$) and Σ, bp is the covariance matrix in basis points (that is, $\times 100^2$). The risk-free rate is 3%.

where we use σ_{ii} to indicate the variance of asset i , since this facilitates the comparison with the matrix expressions. Notice that the denominator $\sigma_{11}\sigma_{22} - \sigma_{12}^2$ is positive since the correlation $\sigma_{12}/(\sigma_1\sigma_2)$ is between -1 and 1 . This means that

$$v_i > 0 \text{ if } SR_i > \rho SR_j,$$

where ρ is the correlation. This shows that an asset should be held in positive amounts if its Sharpe ratio exceeds the correlation times the Sharpe ratio of the other asset.

Example 5.6 ((5.9) with two assets) Let $\Sigma = \begin{bmatrix} 166 & 34 \\ 34 & 64 \end{bmatrix} / 100^2$, $\mu^e = \begin{bmatrix} 5.5 \\ 3.5 \end{bmatrix} / 100$ and $k = 9$. Then

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \approx \begin{bmatrix} 67.6 & -35.9 \\ -35.9 & 175.3 \end{bmatrix} \begin{bmatrix} 0.055 \\ 0.035 \end{bmatrix} \frac{1}{9} \approx \begin{bmatrix} 0.27 \\ 0.46 \end{bmatrix}.$$

The weight on the risk-free is (approximately) $1 - 0.27 - 0.46 = 0.27$. See also Figure 5.4.

Remark 5.7 (Covariance of portfolios) Σv is the vector of covariances of each asset return, R , with the return on the portfolio $v'R$. Also, $\partial v' \Sigma v / \partial v = 2\Sigma v$, is the marginal contribution of each of the assets to the variance of the portfolio.

Remark 5.7 says that Σv is the n -vector of covariances of each investable asset with the optimal portfolio. This means that row i of the first order conditions (5.8) can be written

$$\mu_i^e - k \operatorname{Cov}(R_i, R_v) = 0, \quad (5.10)$$

The first term is the marginal increase in the portfolio excess return from increasing the weight on asset i slightly—and financing it by borrowing. The second term is the risk aversion k times the marginal increase in portfolio variance (divided by 2).

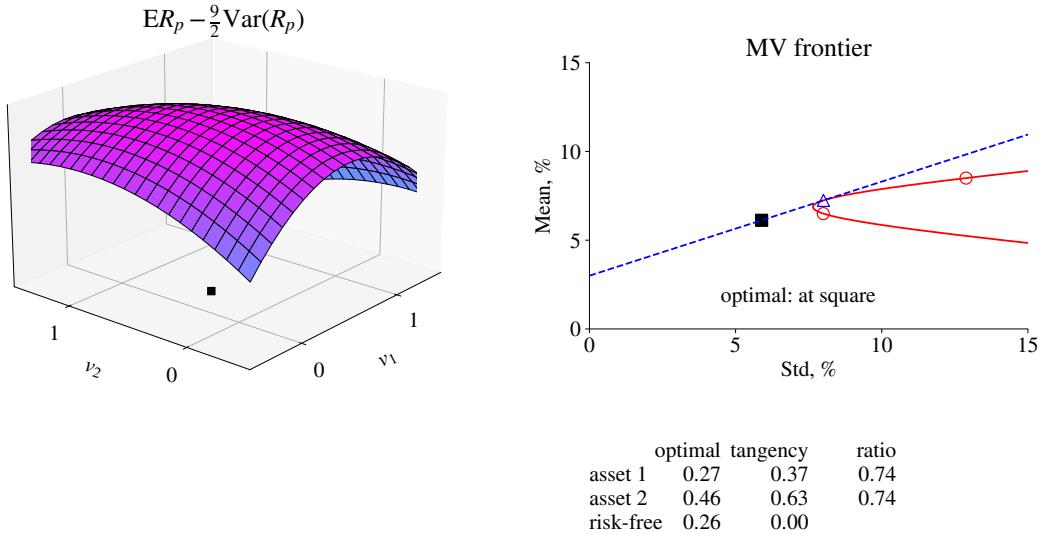


Figure 5.4: Choice of portfolios weights. The calculations use the properties of the assets in Table 5.2.

At the optimum, the two terms (the “benefit” and the “cost” of increasing the position in asset i) are equal. See Figure 5.5 for an illustration. Off the optimum, one is larger than the other—so it is beneficial to change the portfolio. Increasing v_i does not change the first term of (5.10) which is constant at μ_i^e but it will change the second term. The reason for the latter is that a higher v_i value will make the portfolio return more similar to R_i and thus increase the covariance.

Proof of (5.9). With $n = 2$ the portfolio return can be written $R_p = v_1 R_1^e + v_2 R_2^e + R_f$, so the objective is

$$E U(R_p) = v_1 \mu_1^e + v_2 \mu_2^e + R_f - \frac{k}{2} (v_1^2 \sigma_{11} + v_2^2 \sigma_{22} + 2v_1 v_2 \sigma_{12}),$$

where σ_{ii} denotes the variance of asset i and σ_{12} the covariance of asset 1 and 2. The first order conditions are

$$\begin{aligned} 0 &= \partial E U(R_p)/\partial v_1 = \mu_1^e - \frac{k}{2} (2v_1 \sigma_{11} + 2v_2 \sigma_{12}) \\ 0 &= \partial E U(R_p)/\partial v_2 = \mu_2^e - \frac{k}{2} (2v_2 \sigma_{22} + 2v_1 \sigma_{12}), \text{ or} \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} \mu_1^e \\ \mu_2^e \end{bmatrix} - k \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}. \end{aligned}$$

This is the same as (5.8). \square

Remark 5.8 (*Several risky assets, no risk-free**) Maximizing the Lagrangian $v'\mu -$

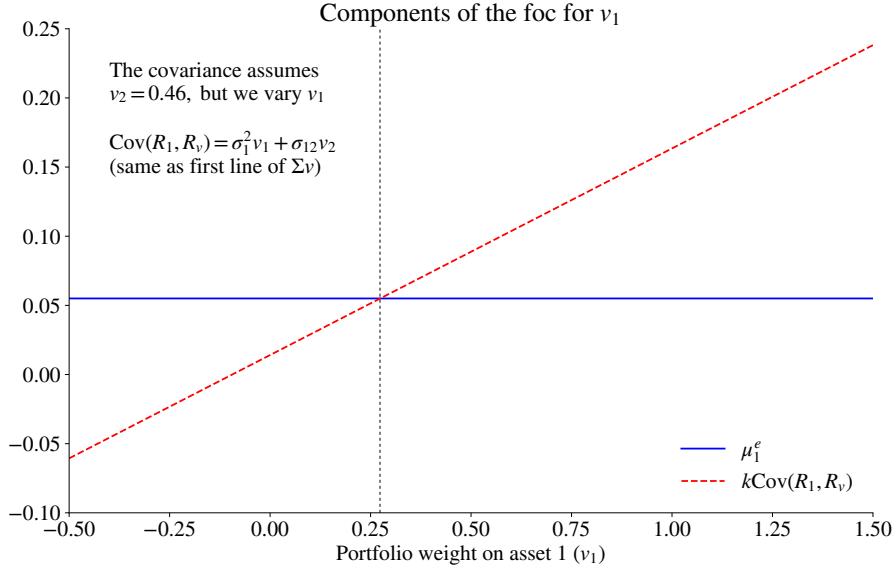


Figure 5.5: Choice of portfolios weights, first order condition. The calculations use the properties of the assets in Table 5.2.

$\frac{k}{2}v'\Sigma v + \theta(1 - v'\mathbf{1})$ where θ is a Lagrange multiplier (for the constraint that the portfolio weights sum to one) gives the first order conditions (wrt. v) $\mu - k\Sigma v - \theta\mathbf{1} = 0$. Rewrite and use the restriction ($\mathbf{1}'v=1$) to write on matrix form

$$\begin{bmatrix} k\Sigma & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix} \begin{bmatrix} v \\ \theta \end{bmatrix} = \begin{bmatrix} \mu \\ 1 \end{bmatrix}.$$

Solve for $[v; \theta]$.

5.4 MV Preferences Gives a Portfolio on the MV Frontier

It is evident that the optimal portfolio (5.9) is a scaling up/down of the *tangency portfolio* (see previous chapters)

$$w_T = \frac{\Sigma^{-1}\mu^e}{\mathbf{1}'\Sigma^{-1}\mu^e}. \quad (5.11)$$

This confirms the result previously discussed in Section 5.1 and illustrated in Figure 5.1.

To be precise, the optimal portfolio can be written

$$v = cw_T, \text{ where } c = \mathbf{1}'\Sigma^{-1}\mu^e/k \quad (5.12)$$

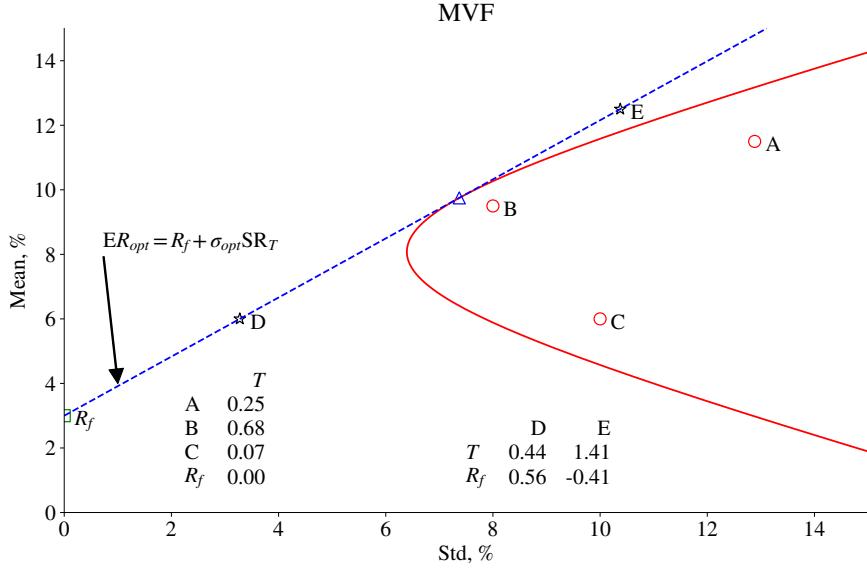


Figure 5.6: Mean-variance frontiers. The properties of the investable assets (A, B, and C) are shown in Table 5.1.

and $1 - c$ in the risk-free asset. The return on any optimal portfolio is thus a mix of the risk-free and tangency portfolio returns

$$\begin{aligned} R_{opt} &= v' R^e + R_f \\ &= c R_T^e + R_f. \end{aligned} \quad (5.13)$$

This means that the optimal portfolio is on the CML and can be constructed by combining the tangency portfolio and the risk-free asset. This result is often called the “two-fund separation theorem”. This has important practical consequences, since it suggests that only two “funds” (one mimicking the tangency portfolio, the other being a risk-free asset) are needed to form optimal portfolios. See Figure 5.6 for an illustration.

Equation (5.13) also shows that the beta of the optimal portfolio is

$$\beta_{opt} = c. \quad (5.14)$$

(This follows directly from the fact that regressing $c R_T^e + R_f$ on R_T^e must give a slope of c , since R_f is constant.) Equation (5.12) then shows that risk averse investors (high k) will choose portfolios with low β and vice versa.

Example 5.9 (*Portfolio choice to get a desired β*) To construct a portfolio with $\beta = 1.2$

against the tangency portfolio, invest $c = 1.2$ in the tangency portfolio and -0.2 in the risk-free.

Remark 5.10 (*The mathematics of why $\max E R_p - k \text{Var}(R_p)/2$ gives a MV portfolio**)
The efficient set solves the problem $\max E R_p$ subject to $\text{Var}(R_p) \leq q$ (where we vary q to trace out the efficient set). Notice that maximizing $E R_p - k \text{Var}(R_p)/2$ can be thought of as the Lagrangian formulation of the efficient set problem.

5.7 Appendix – Numerical Optimization Routines*

5.7.1 Unconstrained Minimization

Consider the loss function

$$f(\theta) = (x - 2)^2 + (4y + 3)^2, \quad (5.15)$$

where $\theta = (x, y)$ contains the two choice variables. Since this loss function is particularly simple—quadratic and also separable in x and y —the solution below is straightforward (the minimum is at $(x, y) = (2, -3/4)$). However, the methods presented can also be used with more complicated loss functions.

A numerical minimization routine searches through different values of θ , typically starting from an initial guess, to find the values that makes $f(\theta)$ as small as possible. 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 stabilise. The starting guess is often important, so be sure to use reasonable values. See Figure 5.7 for an example.

Some algorithms use derivatives of the loss function (which may have to be coded by the user or are calculated numerically), while others do not (“derivative free”). 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 $-g(\theta)$.

5.7.2 Bounds on Variables

Many numerical optimization packages have options for setting bounds on the solution (“box minimization”). As an alternative, we could transform the variables and then apply

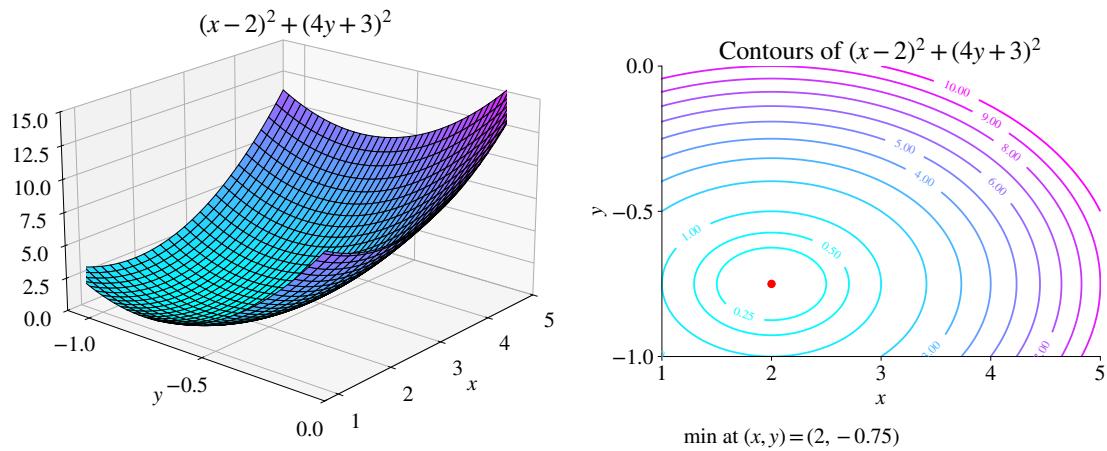


Figure 5.7: Numerical optimization, no restrictions

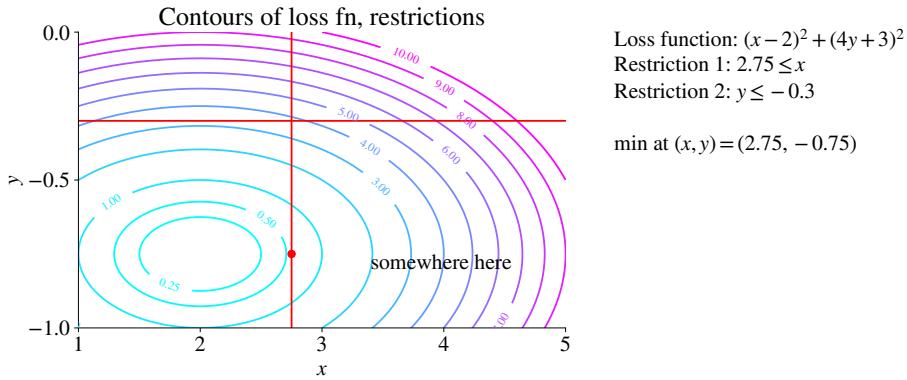


Figure 5.8: Numerical optimization with bounds on the solution

an algorithm for unconstrained optimisation. The latter is briefly discussed below.

A simple way to handle a lower bound, such as $a \leq x$, is to let the routine optimize, without any restrictions, with respect to a transformed variable, $\tilde{x} = \ln(x - a)$. Within the loss function—and also after having obtained the minimizer—the variable can be transformed back using $x = \exp(\tilde{x}) + a$.

Instead, with an upper bound, $x \leq b$, we optimize over $\tilde{x} = \ln(b - x)$ and transform back using $x = b - \exp(\tilde{x})$.

Suppose we use the same loss function (5.15) as before, but also impose the bounds

$$2.75 \leq x \text{ and } y \leq -0.3. \quad (5.16)$$

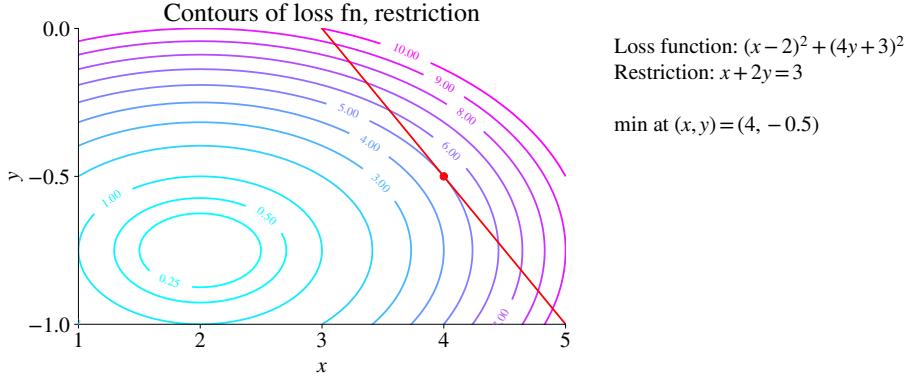


Figure 5.9: Numerical optimization with an equality restriction

The solution is $(x, y) = (2.75, -3/4)$, so only one of the bounds is really binding. See Figure 5.8 for an illustration.

Remark 5.11 With both lower and upper bounds $a \leq x \leq b$, we instead work with the (unbounded) $v = \text{logit}(\frac{x-a}{b-a})$, where the logit function and its inverse are defined as $\text{logit}(u) = \ln(\frac{u}{1-u})$ and $\text{logit}^{-1}(v) = \frac{1}{1+\exp(-v)}$. (The inverse is also called the logistic function.) We can transform back using $x = a + (b - a)\text{logit}^{-1}(v)$

5.7.3 Equality Constraints

Suppose you want an *equality constraint* on the minimization problem, say

$$h_1(\theta) = x + 2y - 3 = 0. \quad (5.17)$$

One way to handle this is to use the constraint to rewrite the loss function (in this case, we could use $x = 3 - 2y$ to replace x in (5.15)). If this is tricky, then we try to find a routine that can handle equality constraints. The short discussion below outlines how these routines work (and also suggests how we could construct such a routine ourselves).

A simple approach is to apply a penalty for deviations from the constraint, thereby modifying the overall loss function to

$$f(\theta) + \lambda \sum_{i=1}^p h_i(\theta)^2, \quad (5.18)$$

where $h_i(\theta)$ is the i th equality constraint. In our example (5.17), there is only one restriction ($p = 1$). See Figure 5.9 for an example. (The solution should be $(x, y) = (4, -1/2)$.)

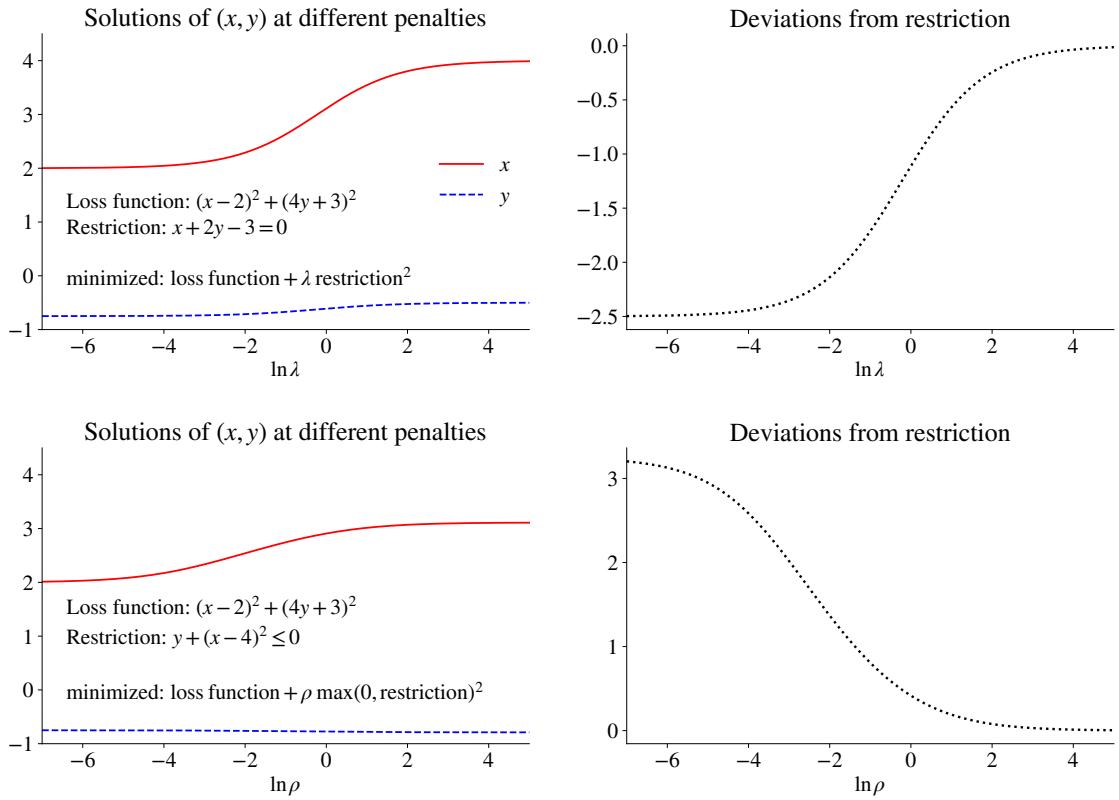


Figure 5.10: Numerical optimizations with penalty on the restriction

Start by setting $\lambda = 0$ and find the optimal value of θ , and call it θ_1 . This is clearly the unconstrained solution. Then, increase λ and redo the optimization (using θ_1 as the starting guess) to get the optimal value θ_2 . Now, increase λ further and redo the optimization (using θ_2 as the starting guess). Keep doing this (with 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 general, it seems as if initially making small increases and later larger ones works well in many cases. See Figure 5.10 for an example. (Clearly, there are more systematic ways to pick the sequence of penalties.)

5.7.4 Inequality Constraints

Instead, we now want to minimize (5.15) under the *inequality constraint* $y \leq -(x-4)^2$.

It is convenient to rewrite all inequality constraints on a common form, and we here

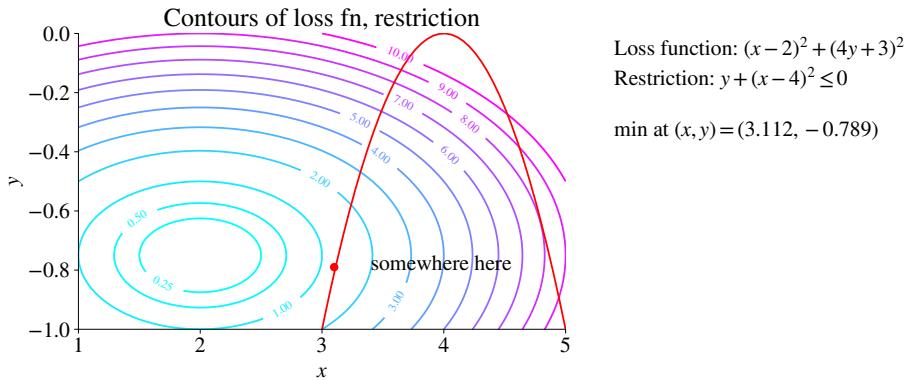


Figure 5.11: Numerical optimization with inequality restriction

choose to write them all on ≤ 0 form, which gives

$$g_1(\theta) = y + (x - 4)^2 \leq 0. \quad (5.19)$$

Now, we minimise the overall loss function

$$f(\theta) + \rho \sum_{j=1}^q \max[0, g_j(\theta)]^2, \quad (5.20)$$

where $g_j(\theta)$ is the j th inequality constraint (there is only one in our example). Notice that 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 a higher ρ , etc. See Figure 5.11 for an example. (The solution should be close to $(x, y) = (3.1, -0.79)$.) See Figure 5.10 for an iterative approach with a larger and larger penalty.

Finally, we can combine equality and inequality constraints as

$$f(\theta) + \lambda \sum_{i=1}^p h_i(\theta)^2 + \rho \sum_{j=1}^q \max[0, g_j(\theta)]^2. \quad (5.21)$$

Further Reading

See Brandimarte (2006), Stan manual (<http://mc-stan.org/users/documentation/>), Kochenderfer and Wheeler (2019)

Chapter 6

CAPM

6.1 Beta Representation of Expected Returns

6.1.1 Beta Representation: Definition

The beta representation (and eventually also CAPM as developed by Sharpe (1964), Lintner (1965) and Mossin (1966)) follows from the analysis of portfolio choice based on mean-variance preferences.

From an earlier chapter, we notice two things. First, the *first order conditions* for optimal portfolio choice are the n equations in

$$\mathbb{E} R^e = k \Sigma v. \quad (6.1)$$

Recall that Σv is a vector of covariances of each asset with the v -portfolio. Second, the optimal portfolio weights (v) are *proportional to the tangency portfolio*, w_T ,

$$v = c w_T, \quad (6.2)$$

where c is a scalar. Notice that these expressens depend on beliefs about the average returns (μ) and their variance-covariance matrix (Σ) of the investable assets. All the results in this section are therefore dependent on those beliefs.

Combining these two observations shows that the first order conditions for optimal portfolio choice can be written

$$\mathbb{E} R^e = kc \Sigma w_T, \text{ or} \quad (6.3)$$

$$\mathbb{E} R_i^e = kc \sigma_{iT}, \text{ for } i = 1, \dots, n, \quad (6.4)$$

where k is the risk aversion and σ_{iT} is shorthand notation for the covariance of R_i and

R_T , which is the same as element i of Σw_T . I use the notation $E R_i^e$ for the left hand side (rather than the equivalent μ_i^e) to suggest that this is a result, not “data.”

We can express this as a *beta representation*. Let μ_T^e be the expected excess return on the tangency portfolio and rewrite (6.4) as

$$E R_i^e = \beta_i \mu_T^e \text{ where } \beta_i = \sigma_{iT} / \sigma_T^2, \text{ or} \quad (6.5)$$

$$E R_i = R_f + \beta_i \mu_T^e. \quad (6.6)$$

(See below for a proof). Plotting $E R_i^e$ or $E R_i$ against β_i gives the *security market line*, see Figure 6.1.

It is important to acknowledge that this expression does not say anything about *causality*: it just shows how expected returns and betas relate to each other according to the first order conditions for optimal portfolio choice.

Proof of (6.5). Premultiply both sides of (6.3) by w'_T to get $w'_T \mu^e = k c w'_T \Sigma w_T$ which is the same as $\mu_T^e = k c \sigma_T^2$. Solve for $k c$ and use in (6.4). The end of the section presents an alternative proof. \square

The β_i is clearly the slope coefficient in a (time series) OLS regression

$$R_{it}^e = \alpha_i + \beta_i R_{Tt}^e + \varepsilon_{it}, \quad (6.7)$$

where R_{it}^e is the excess return on asset i in period t .

Remark 6.1 (*Calculating β_i from the covariance matrix**) *The traditional way of estimating β_i is to run a regression. However, if we know the variance-covariance matrix Σ of the investable assets, then we can also use the fact that $\beta_i = \sigma_{iT} / \sigma_T^2$ where $\sigma_{iT} = w'_i \Sigma w_T$. (Here, w_i is trivial: 1 in position i and 0 elsewhere.) Using the asset price characteristics in Table (6.1), together with the weights of the tangency portfolio gives the β values in Figure 6.1.*

Example 6.2 (β_i vs $E R_i$) With $R_f = 3\%$ and $\mu_T = 9.75\%$ (so $\mu_T^e = 6.75\%$) we get

β_i	$E R_i$	Comment
0.44	6.0%	low β , low avg. return
1.5	13.12%	high β , high avg. return
1	9.75%	same risk as market
0	3%	no risk
-0.5	-0.38%	the opposite of risk

	$\mu, \%$	Σ, bp		
		A	B	C
A	11.5	166	34	58
B	9.5	34	64	4
C	6.0	58	4	100

Table 6.1: Characteristics of the three assets in some examples. Notice that $\mu, \%$ is the expected return in % (that is, $\times 100$) and Σ, bp is the covariance matrix in basis points (that is, $\times 100^2$).

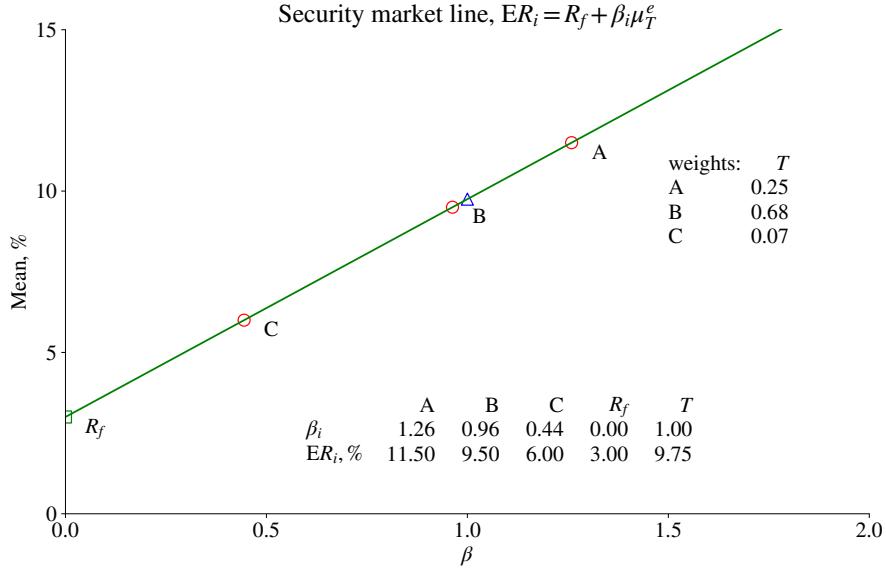


Figure 6.1: Security market line. The properties of the investable assets (A, B, and C) are shown in Table 6.1.

Proof of (6.5), alternative. The foc are $\Sigma v^z = E R^e / k^z$, where k^z is the risk aversion and v^z is the vector of portfolio weights of investor z . Beliefs are assumed to be shared by all investors. Assume all investors have the same capital (or that wealth is uncorrelated with risk aversion) and average the foc across the Z investors, $\sum_{z=1}^Z \Sigma v^z / Z$, to get $\Sigma v_m = E R^e / k^*$, where v_m is the market (average) portfolio and where $k^* = 1 / \sum_{z=1}^Z 1 / (k^z Z)$ defines an aggregate risk aversion k^* . Row i of this expression is $\sigma_{im} = E R_i^e / k^*$. For the v_m portfolio, this gives $\sigma_m^2 = E R_m^e / k^*$. Combine the last two expressions as $(\sigma_{im} / \sigma_m^2) E R_m^e = E R_i^e$. \square*

Remark 6.3 (*Expected return of the tangency portfolio**) *It follows directly that $E R_T^e = k \sigma_T^2$. (Multiply both sides of (6.1) by w_T .) This says that the risk premium on the tangency*

portfolio increases if the risk aversion or variance does.

6.1.2 Betas of Portfolios

Recall that the beta of any portfolio (not just the optimal one) is the weighted average (portfolio) of the betas of its components. That is, the portfolio with return

$$R_p = w' R^e + R_f \text{ has the beta} \quad (6.8)$$

$$\beta_p = w' \beta. \quad (6.9)$$

(This follows directly from $\beta_p = \text{Cov}(\sum_{i=1}^n w_i R_i, R_T)/\sigma_T^2 = \sum_{i=1}^n w_i \beta_i$.)

Example 6.4 Let $(\beta_1, \beta_2) = (1.2, 0.8)$. The portfolio return $R_p = 0.6R_1 + 0.4R_2$ has the beta $\beta_p = 0.6 \times 1.2 + 0.4 \times 0.8 = 1.04$.

In particular, consider the portfolios on the capital market line (CML): $R_{opt} = vR_T^e + R_f$, where v is the weight on the tangency portfolio. Using the result in (6.9) and noticing that the tangency portfolio has $\beta_T = 1$ gives that $\beta = v$ for any portfolio on the CML. This implies that it is straightforward to create a portfolio with any desired β : just invest β in the tangency portfolio and $1 - \beta$ in the risk-free.

Example 6.5 (Creating a portfolio with $\beta_p = 0.44$) We can create a portfolio with $\beta = 0.44$ by investing 0.44 in the tangency portfolio and 0.56 in the risk-free.

6.1.3 Beta Representation and the Capital Market Line

The beta representation (6.5) means that two assets with the same betas should have the same expected returns—even if they have very different volatilities.

To be precise, consider the regression (6.7) which has the usual property that the residual is uncorrelated with the regressor. We can therefore write the variance of R_i as

$$\sigma_i^2 = \beta_i^2 \sigma_T^2 + \sigma_\varepsilon^2. \quad (6.10)$$

This says that the variance of R_i has two components: *systematic risk* (the comovement of R_i with R_T , $\beta_i^2 \sigma_T^2$) and *idiosyncratic noise* (the variance of ε_i , σ_ε^2). In particular, *MV efficient portfolios have only systematic risk* ($\sigma_\varepsilon^2 = 0$) since they are formed from the tangency portfolio and risk-free ($R_{opt} = vR_T + (1 - v)R_f$). All other portfolios with the same β are to the right in the MV figure: see Figure 6.2 for an illustration.

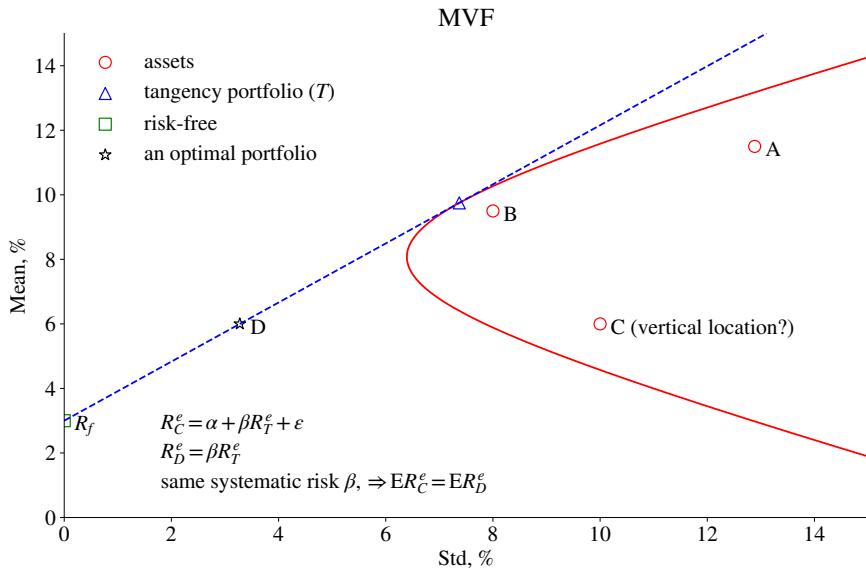


Figure 6.2: Mean-variance frontier and expected returns. The properties of the investable assets (A, B, and C) are shown in Table 6.1.

Example 6.6 In Figure 6.2, we want to understand the mean return (vertical location) of asset C (taking its volatility and β as given). We notice that C has the same systematic risk as the efficient portfolio D. According to the beta representation, C must then have the same average return as D.

6.1.4 The Tangency Portfolio is the Market Portfolio

To determine the equilibrium asset prices (and therefore expected returns) we have to equate demand (the mean variance portfolios) with supply, which we assume is exogenous. Since we assume a fixed and exogenous supply (say, 2000 shares of asset 1 and 407 shares of asset 2,...), prices, and therefore returns, are driven by demand, at least in the short run.

Suppose all investors have the same beliefs about the asset returns (same expected returns and covariance matrix). They will then all mix the same tangency portfolio with the risk-free—but possibly in different proportions due to different risk aversions.

In equilibrium, net supply of the risk-free assets is zero (lending = borrowing), so *the average investor must hold the tangency portfolio and no risk-free assets*. Therefore, *the tangency portfolio must be the market portfolio*, so we can replace R_T with R_m in all

expressions above. Therefore, CAPM says

$$\mathbb{E} R_i = R_f + \beta_i \mu_m^e \text{ where } \beta_i = \sigma_{im}/\sigma_m^2. \quad (6.11)$$

As discussed before, this expression is just a characterisation of the equilibrium (the first order conditions), and CAPM is silent on how that equilibrium is reached. One possible *story* is that β_i is driven by the firm characteristics (industry, size, leverage, etc.) and that equilibrium is reached as follows: high β assets are in low demand since they are too procyclical (pay off at the wrong time) which means that (in equilibrium) the share price will be low. For a given dividend stream, this means a higher dividend/price ratio, which contributes to a high average return.

Clearly, *CAPM relies on very strong assumptions*, in particular, the assumption about all investors having the same beliefs. Also, it rules out that investors face other types of financial risks (not just asset market risks). These issues will be discussed in later chapters.

6.1.5 Summarizing MV and CAPM: CML and SML

According to MV analysis, and assuming that the market portfolio equals the tangency portfolio, average return of all optimal (effective) portfolios (denoted opt) obey

$$\mathbb{E} R_{opt} = R_f + \sigma_{opt} SR_m. \quad (6.12)$$

The plot of $\mathbb{E} R_{opt}$ against σ_{opt} is called the *capital market line*. See Figure 6.3 for an illustration.

In contrast, according to CAPM, the average return on all portfolios (optimal or not), obey the beta representation (6.6)

$$\mathbb{E} R_i = R_f + \beta_i \mu_m^e. \quad (6.13)$$

The plot of $\mathbb{E} R_i$ against β_i (for different assets, i) is called the *security market line*. Again, see Figure 6.3 for an illustration.

6.2 More Properties of CAPM

6.2.1 Heterogeneous Beliefs*

This section discusses different risk aversions and heterogeneous beliefs about the mean returns. Both these allow for fairly straightforward aggregation. In contrast, handling

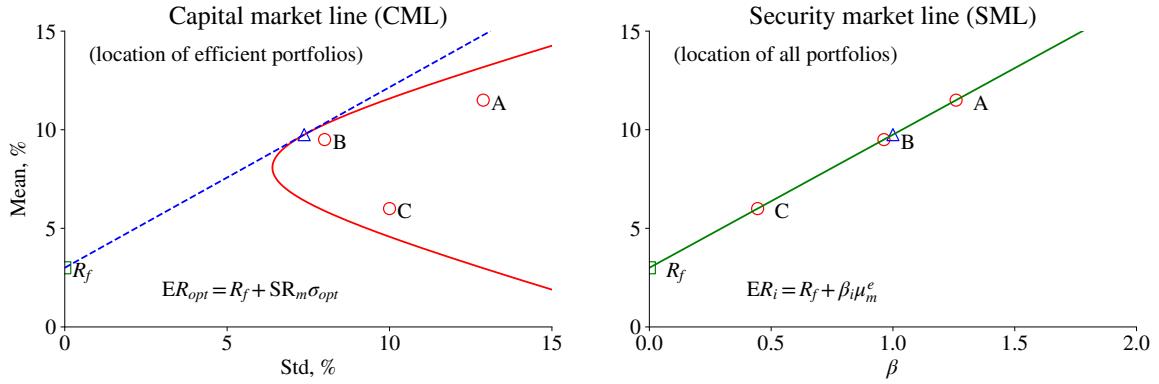


Figure 6.3: CML and SML. The properties of the investable assets (A, B, and C) are shown in Table 6.1.

different beliefs about the variance-covariance matrix are much harder.

We further assume that all investors have the same capital to invest (or more generally, if the capital of a investor is uncorrelated to the beliefs and risk aversion). Let superscript z to indicate the investor, so $E^z R^e$ is the vector of his/her expected returns of the n assets and k^z is the risk aversion. Then, define the consensus (average) expectations across the Z investors as

$$E^* R^e = \sum_{z=1}^Z E^z R^e k^*/(k^z Z), \text{ where } k^* = 1/[\sum_{z=1}^Z 1/(k^z Z)]. \quad (6.14)$$

Notice that k^* can be interpreted as an aggregate risk aversion. In the average expectations, investors with high risk aversion (and thus less inclined to invest in risky assets) have less weight. It follows that the market portfolio is the optimal portfolio ($v_m = \Sigma^{-1} E^* R^e / k^*$) for an investor with the expectations in (6.14).

This gives a beta equation for consensus beliefs

$$E^* R_i^e = \beta_i E^* R_m^e, \quad (6.15)$$

where $E^* R_i^e$ is an average (“consensus”) expectation of the excess return on asset i , and m signifies the market portfolio. (See below for a proof.) Overall, this analysis suggests that CAPM might hold, at least as an approximation, for some types of heterogeneous beliefs.

Proof of (6.14)–(6.15). For investor z the first order conditions (6.1) are $E^z R^e / k^z = \Sigma v^z$. Beliefs about the variance-covariance matrix Σ are assumed to be shared by all investors. Average the foc across the Z investors, $\sum_{z=1}^Z \Sigma v^z / Z$, to get $\Sigma v_m = \sum_{z=1}^Z E^z R^e / (k^z Z)$, where v_m is the market portfolio. Define an aggregate risk aver-

sion as in (6.14), and a consensus expectation as in the same equation. Then, the previous equation can be written $\Sigma v_m = E^* R^e / k^*$, which can be solved for v_m . Notice that $E^* R^e$ is a consensus (or average) belief. Row i of this expression is $\sigma_{im} = E^* R_i^e / k^*$, since row i of Σv_m is the covariance of asset i and the v_m portfolio, σ_{im} . For the v_m portfolio, this gives $\sigma_m^2 = E^* R_m^e / k^*$. Combine the last two expressions as $(\sigma_{im}/\sigma_m^2) E^* R_m^e = E^* R_i^e$.

□

6.2.2 CAPM and Stochastic Discount Factors*

For future reference, we here notice that the CAPM expression (6.11)) can also be written in terms of a “stochastic discount factor” (SDF) model. This model implies

$$E R_i^e M = 0, \text{ where } M = a - b R_m^e, \text{ with } b > 0. \quad (6.16)$$

Many asset pricing models can be written on a similar form, as will be discussed in later chapters.

Proof of (6.16) giving (6.11). Recall that $\text{Cov}(x, y) = E xy - E x \times E y$, so $E xy = 0$ can be rearranged as $E y = -\text{Cov}(x, y)/E x$. Applying to (6.16) gives

$$E R_i^e = b \sigma_{im} / (a - b E R_m^e)$$

We can, of course, apply this expression to the market excess return (instead of asset i) to get

$$E R_m^e = b \sigma_m^2 / (a - b E R_m^e).$$

Solve for $b/(a - b E R_m^e)$ and use that in the first equation to get the CAPM expression (6.11). □

6.2.3 Back to Prices (The Gordon Model)*

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

$$1 + R_{t+1} = (D_{t+1} + P_{t+1})/P_t, \quad (6.17)$$

where P_t is the asset price and D_{t+1} is the dividend it gives at the beginning of the next period. If we assume that expected returns are constant across time (denoted R , for instance 10%) and that dividends are expected to grow at the rate g (for instance, 2%) after period $t + 1$, then it is straightforward to show that the asset price is

$$P_t = E_t D_{t+1} / (R - g). \quad (6.18)$$

(See a later appendix on discounted cash flow.) Clearly, higher (expected) dividends and/or a higher growth rate increases the asset price. In addition, a lower expected (“required”) *future return* also increases *today’s asset price*.

In CAPM, a lower expected return could be driven by a lower beta or by a lower risk-free rate. One way of interpreting this is as follows. Suppose the asset (suddenly) gets a lower beta, which means that it has less systematic risk than before. It is therefore more useful in portfolio formation and thus more demanded—so the price level increases. With a higher price level, the dividend yield is lower, which contributes to a lower return (recall the return is the dividend yield plus the capital gains yield). The valuation in (6.18) and CAPM are then consistent.

6.3 Testing CAPM

6.3.1 Testing a Single Asset

The basic implication of CAPM is that the expected excess return of an asset ($E R_{it}^e$) is linearly related to the expected excess return on the market portfolio ($E R_{mt}^e$) according to (6.11). This could be tested by the regression (6.7), but where we use the market return to proxy for the tangency portfolio return.

In particular, take average (over time) of the regression to get

$$\bar{R}_i^e = \hat{\alpha}_i + \hat{\beta}_i \bar{R}_m^e, \quad (6.19)$$

where \bar{R}_i^e is the average excess return on asset i in the sample ($\bar{R}_i^e = \sum_{t=1}^T R_{it}^e / T$). Notice that $\bar{\epsilon}_i = 0$ by construction.

The OLS estimate of β_i is the sample analogue to the true β_i . It is then clear that *CAPM implies* that the intercept (α_i) of the regression should be zero, which is what empirical tests of CAPM focus on.

However, this interpretation relies on a big *assumption*: that market expectations are (on average) well represented by the sample data. A rejection of CAPM in the test could therefore be driven either by some false assumptions behind CAPM or that the sample is systematically different from the market beliefs during the same period.

The test of the null hypothesis that $\alpha_i = 0$ uses the fact that, under fairly mild conditions, the t-statistic has an asymptotically normal distribution, that is

$$\frac{\hat{\alpha}_i}{\text{Std}(\hat{\alpha}_i)} \xrightarrow{d} N(0, 1) \text{ under } H_0 : \alpha_i = 0. \quad (6.20)$$

In this expression, $\hat{\alpha}_i$ is the estimate of the intercept in (6.7) and $\text{Std}(\hat{\alpha}_i)$ its standard deviation (error), for instance, from the usual Gauss-Markov results for OLS. See also Remark 6.8 for a discussion of heteroskedasticity and autocorrelation. We reject the null hypothesis ($\alpha_i = 0$) when the t-statistic is very negative or very positive (for instance, lower than -1.64 or higher than 1.64 for the 10% significance level, and $-1.96/1.96$ for the 5% level).

The test assets are often portfolios of firms with similar characteristics, for instance, small size or having their main operations in the retail industry. There are several reasons for testing the model on such portfolios: individual stocks are extremely volatile and firms can change substantially over time (so the beta changes). Moreover, it is of interest to see how the deviations from CAPM are related to firm characteristics (size, industry, etc), since that can possibly suggest how the model needs to be extended.

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

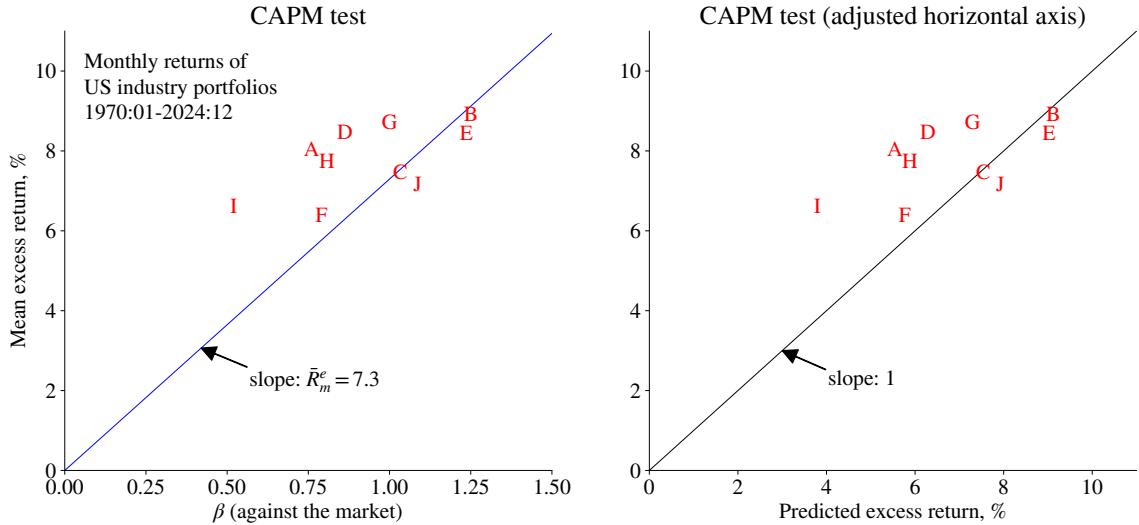
Empirical Example 6.7 Figure 6.4 shows some results for US industry portfolios, while Table 6.2 and Figures 6.5–6.6 for US size/book-to-market portfolios. In these figures, the results are plotted in one of two different ways:

$$\begin{array}{c} \text{horizontal axis} \\ \hline 1 : \beta_i & \bar{R}_i^e \\ 2 : \beta_i \bar{R}_m^e & \bar{R}_i^e, \end{array} \quad (6.21)$$

where \bar{R}_i^e indicates the (time) average excess return of asset i . In the first approach, CAPM says that all data points (different assets, i) should cluster around a straight line with a slope equal to the average market excess return, \bar{R}_m^e . In the second approach, CAPM says that all data points should cluster around a 45-degree line. In either case, the vertical distance to the line is α_i (which should be zero according to CAPM).

6.3.2 Testing Several Assets

In most cases there are several (n) test assets, and we actually want to test if all the α_i (for $i = 1, 2, \dots, n$) are zero, because that is the implication of CAPM. Ideally we then want to take into account the correlation of the different alpha estimates.



	α (ann.)	t-stat	σ (ann.)
A (NoDur)	2.44	2.06	8.73
B (Durbl)	-0.25	-0.11	17.26
C (Manuf)	-0.12	-0.14	6.43
D (Enrgy)	2.13	0.93	16.84
E (HiTec)	-0.63	-0.41	11.17
F (Telcm)	0.57	0.38	10.99
G (Shops)	1.38	1.13	8.97
H (Hlth)	1.81	1.22	10.94
I (Utils)	2.77	1.74	11.75
J (Other)	-0.80	-0.85	6.97

$$\begin{aligned} \text{CAPM: } & R_i^e = \alpha_i + \beta_i R_m^e + e_i \\ \text{Predicted excess return: } & \beta_i \bar{R}_m^e \\ \text{10% crit. value (Bonferroni): } & 2.58 \\ \text{Test if all } \alpha_i = 0: & \\ \text{Wald stat} & 10.55 \\ \text{5% crit val} & 18.31 \\ \text{p-value} & 0.39 \end{aligned}$$

Figure 6.4: CAPM regressions on US industry indices

While it is straightforward to construct such a test, it requires setting up a system of regression equations and test across regressions. See Remark 6.8 for details. Alternatively, we can apply a *Bonferroni adjustment* of the individual t-stats: reject CAPM at the 10% significance level only if the largest t-stat (in absolute terms) exceeds the critical value at the $0.10/n$ significance level. For instance, with $n = 25$, the critical value from a standard normal distribution would be 2.88 instead of 1.64. The motivation for this is that repeated single-asset testing (using a traditional critical value) will, by pure randomness, reject 10% of the cases—even if the null hypothesis is true. The Bonferroni adjustment takes this into account to correct the “family-wise” false rejection rate.

Remark 6.8 (*Variance-covariance matrix if OLS estimate of α) The “system estimation” is actually n separate OLS regressions. Assuming residuals have no autocorrelation or heteroskedasticity (the standard Gauss-Markov assumption), it is straightforward to show

that the variance-covariance matrix of the vector of the estimated $\hat{\alpha}$ -values ($\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_n$) is $V = \Omega(1 + SR_m^2)/T$, where Ω is the variance-covariance matrix of the residuals, SR_m is the Sharpe ratio of the market and T the (time) length of the sample. This holds also for a single asset as in (6.20). For monthly or longer return periods, autocorrelation is rarely a problem and a heteroskedasticity-robust V is typically similar (for α , not for β). The joint hypothesis that all alphas are zero can be tested with an F - or χ^2 -test. The latter has the test statistic $\hat{\alpha}'V^{-1}\hat{\alpha}$, which is distributed as a χ_n^2 variable under the null hypothesis.

	1	2	3	4	5
1	-3.45	-0.09	0.44	2.11	2.52
2	-2.33	0.53	1.43	2.30	1.84
3	-2.19	1.28	1.20	2.22	2.22
4	-0.86	0.39	1.21	2.06	1.54
5	0.24	1.23	1.17	0.11	0.97

Table 6.2: t-stats for α in CAPM, 25 FF portfolios 1970:01-2024:12. NW uses 1 lag. The Bonferroni adjusted 10% and 5% critical values are 2.88 and 3.09.

A quite different approach to study a cross-section of assets is to first perform a CAPM regression for each asset, and use the estimated betas as regressors in the following cross-sectional regression

$$\bar{R}_i^e = \gamma + \lambda \hat{\beta}_i + u_i, \quad (6.22)$$

where \bar{R}_i^e is the (sample) average excess return on asset i . Notice that the estimated betas are used as regressors and that there are as many data points as there are assets (n).

There are severe econometric problems with this regression equation since the regressor contains measurement errors (it is only an estimate), which typically tend to bias the slope coefficient towards zero (“errors in variables”). To get the intuition for this bias, consider an extremely noisy measurement of the regressor: it would be virtually uncorrelated with the dependent variable (noise isn’t correlated with anything), so the estimated slope coefficient would be close to zero.

If we could overcome this bias (and we can by being careful), then the testable implications of CAPM is that $\gamma = 0$ and that λ equals the average market excess return. We also want (6.22) to have a high R^2 —since it should be unity in a very large sample (if CAPM holds).

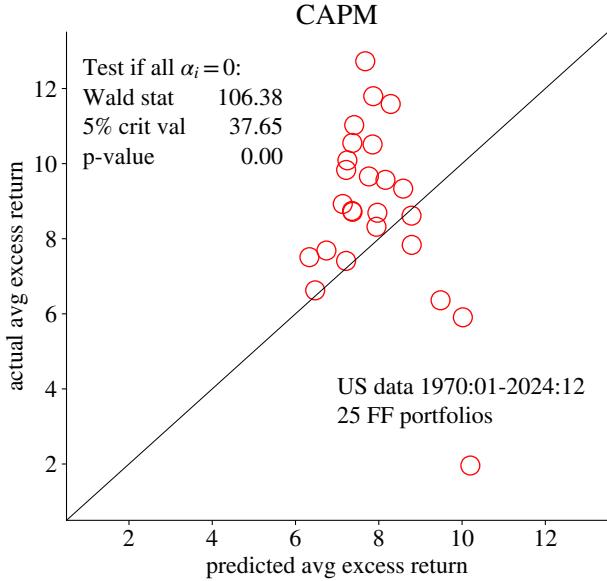


Figure 6.5: CAPM, FF portfolios

6.3.3 Representative Results of the CAPM Test

One of the more interesting studies is Fama and French (1993) (see also Fama and French (1996)). They construct 25 stock portfolios according to two characteristics of the firm: the size (by market capitalization) and the book-value-to-market-value ratio (BE/ME).

They run a traditional CAPM regression on each of the 25 portfolios (monthly data 1963–1991)—and then study if the expected excess returns are related to the betas as they should according to CAPM. However, it is found that there is almost no relation between \bar{R}_i^e and β_i (there is a cloud in the $\beta_i \times \bar{R}_i^e$ space). This is due to the combination of two features of the data. First, *within a BE/ME quintile*, there is a positive relation (across size quantiles) between \bar{R}_i^e and β_i —as predicted by CAPM. Second, *within a size quintile* there is a negative relation (across BE/ME quantiles) between \hat{R}_i^e and β_i —in stark contrast to CAPM. See Figures 6.5–6.6 for results from more recent data.

6.3 Appendix – Discounted Cash Flow*

6.3.1 Fundamental Asset Value

A *present value* is a sum of discounted future cash flows. A higher discount rate and longer time until the cash flow reduces the present value.

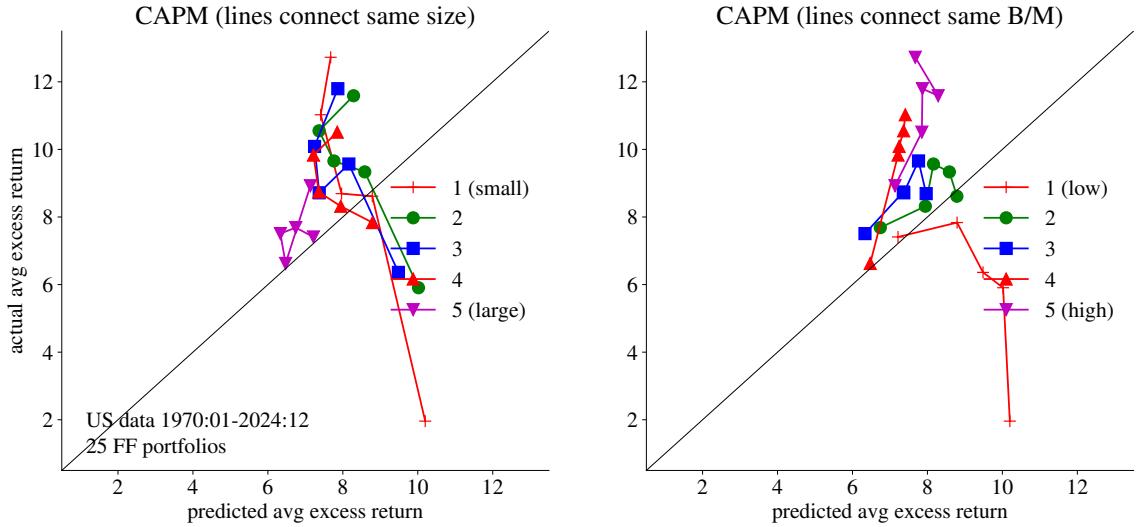


Figure 6.6: CAPM, FF portfolios

Remark 6.9 If the cash flow is -150 in t , 100 in $t + 1$ and 130 in $t + 2$, and the discount rate $R = 0.1$ then

$$-150 + \frac{100}{1+R} + \frac{130}{(1+R)^2} \approx 48.3 \text{ for } R = 0.1.$$

Many assets are long-lived. A fundamental valuation of the asset is that its (fair) price equals the present value of future cash flow

$$P_t = \sum_{s=1}^{\infty} \frac{E_t D_{t+s}}{(1+R)^s}, \quad (6.23)$$

where D_{t+s} are the future cash flows to the investor. In this expression subscripts refer to time periods.

For shares the cash flows are the dividend payments, while for bonds they are the coupon and face value payments. In this section, the discount rate R is given (and here assumed to be constant). In general, the discount rate depends on both the risk-free rate and the risk of the asset. In project evaluation, the discount rate is often a weighted average (“WACC”) of the required return on equity and the after tax borrowing rate.

Remark 6.10 (What if the company cancels dividends in order to invest more?*) Suppose the investment project generates an annual return of ROE—and all earnings are paid out

in period 3:

$$\begin{aligned} \text{Old plan: } P_0 &= \frac{E_0 D_1}{1+R} + \frac{E_0 D_2}{(1+R)^2} + \frac{E_0 D_3}{(1+R)^3} + \dots \\ \text{New plan: } \tilde{P}_0 &= \frac{\mathbf{0}}{1+R} + \frac{E_0 D_2}{(1+R)^2} + \frac{E_0 D_3 + E_0 D_1(1+ROE)^2}{(1+R)^3} + \dots \end{aligned}$$

Same value ($\tilde{P}_0 = P_0$) if $ROE = R$.

In general, *dividends* reduce the stock price on the ex-dividend day (when the next dividend belongs to the seller, rather than the buyer of the stock) by an amount equal to the dividend. In contrast, a *stock repurchase* does not directly affect the stock price, but clearly reduces the number of outstanding (floating) shares. Both methods (if of same size) are likely to reduce the market value of the firm with the same amount (Taxes, changes in risk and beliefs about future cash flows can complicate this story.) See [Fabozzi, Neave, and Zhou \(2012\)](#) for a discussion.

Remark 6.11 (*Dividends and stock repurchases**) Suppose the total value of a firm is 100, of which 90 is the present value of future earnings and 10 is cash. With 10 outstanding shares, the share price is 10 (100/10). If the firm distributes the cash as dividends, then the remaining total value of the firm is 90 so the share price is now 9. Overall the share holders have this period (assuming no other news) received a zero return (dividend yield + capital gain). Instead, if the firm buys back one share at the price of 10, then the total firm value becomes 90 and there are now 9 outstanding shares, so the share price would be unchanged at 10. Again, the return is zero.

Remark 6.12 (*Valuation in terms of earnings instead of dividends**) Earnings can be spent on dividends or kept on the balance sheet as cash or some other asset (an “investment”): $E = D + I$. The firm value is

$$P_0 = \frac{\overbrace{E_0 D_1}^{E_1 - I_1}}{1+R} + \frac{\overbrace{E_0 D_2}^{E_2 - I_2}}{(1+R)^2} + \frac{\overbrace{E_0 D_3}^{E_3 - I_3}}{(1+R)^3} + \dots$$

This shows that the firm value equals the present value of future earnings minus the present value of new investment expenditures used to generate those earnings.

Remark 6.13 (*From income to cash flow**) To calculate the cash flow start with the earnings before interests and taxes (EBIT) from the income statement, subtract taxes on

EBIT (they are costs...), add back the depreciations (it is just an accounting item), subtract the capital expenditure (buying machines takes cash, even if it is not booked as a cost) and also subtract the change in the net working capital (current assets minus current liabilities, booked as income but you have not received it yet). All financial transactions are disregarded, so the cash flow must be used to pay all bond and equity holders.

Remark 6.14 (Internal Rate of Return) *The IRR is the R that makes the net present value of a cash flow process zero. For instance, if the cash flow is -150 in t (an investment), 100 in $t + 1$ and 130 in $t + 2$, then*

$$-150 + \frac{100}{1+R} + \frac{130}{(1+R)^2} \approx 0 \text{ for } R = 0.32.$$

Typically we have to solve for the IRR by numerical methods. Notice that there may be more than one IRR if the cash flow process changes sign more than once.

6.3.2 “Speculative” Valuation

An alternative view of the asset value is the present of the next dividend plus what you expect to resell the asset for

$$P_t = \frac{E_t D_{t+1} + E_t P_{t+1}}{1+R}. \quad (6.24)$$

This is the same as the fundamental valuation (6.23) if you expect to resell it at your (expected next period) fundamental valuation. Otherwise not.

Proof of fundamental = speculative asset value, if $E_t P_{t+1}$ follows fundamental valuation. Use (6.23) to write

$$P_{t+1} = \frac{E_{t+1} D_{t+2}}{1+R} + \frac{E_{t+1} D_{t+3}}{(1+R)^2} + \dots$$

Take expectations as of period t and use in (6.24)

$$P_t = \frac{E_t D_{t+1}}{1+R} + \frac{E_t E_{t+1} D_{t+2}}{(1+R)^2} + \frac{E_t E_{t+1} D_{t+3}}{(1+R)^3} + \dots$$

Recall that $E_t(E_{t+1} D_{t+s}) = E_t D_{t+s}$ (the “law of iterated expectations.”) to complete the proof. \square

Remark 6.15 (Law of iterated expectations) *The law of iterated expectations implies that $E_t(E_{t+1} y_{t+2}) = E_t y_{t+2}$. To see why, let $y_{t+2} = E_{t+1} y_{t+2} + \varepsilon_{t+2}$, so ε_{t+2} is a surprise in $t + 2$. The equation above can then be written $E_t(y_{t+2} - \varepsilon_{t+2}) = E_t y_{t+2}$, which holds if $E_t \varepsilon_{t+2} = 0$. That is, the surprise in $t + 2$ cannot be predicted by any information in*

period t . Basically, this is the same as saying that we know more, not less, as time goes by.

6.3.3 Fundamental Valuation and Returns

The return from holding the asset from t to $t + 1$ is

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} - 1. \quad (6.25)$$

If the discount rate in (6.23) is constant over time, then it equals the expected return

$$\mathbb{E}_t R_{t+1} = R. \quad (6.26)$$

It follows that if there is no news between t and $t + 1$ (so expectations are unchanged, $\mathbb{E}_t D_{t+s} = \mathbb{E}_{t+1} D_{t+s}$), then

$$R_{t+1} = R \text{ (if no news).} \quad (6.27)$$

Notice that this return does *not* depend on the level or growth rate of the dividends. Old information is in P_t , and does not affect R_{t+1} .

Proof of (6.27)–(6.26). Use (6.23) to write

$$P_{t+1} = \frac{\mathbb{E}_{t+1} D_{t+2}}{1+R} + \frac{\mathbb{E}_{t+1} D_{t+3}}{(1+R)^2} + \dots$$

Use in the realized return (6.25) and take expectations as of t to get (using $\mathbb{E}_t \mathbb{E}_{t+1} D_{t+s} = \mathbb{E}_t D_{t+s}$)

$$\mathbb{E}_t R_{t+1} = \frac{\mathbb{E}_t D_{t+1} + \frac{\mathbb{E}_t D_{t+2}}{1+R} + \frac{\mathbb{E}_t D_{t+3}}{(1+R)^2} + \dots}{\frac{\mathbb{E}_t D_{t+1}}{1+R} + \frac{\mathbb{E}_t D_{t+2}}{(1+R)^2} + \frac{\mathbb{E}_t D_{t+3}}{(1+R)^3} + \dots} - 1 = R.$$

In addition, if expectations are unchanged, then $R_{t+1} = \mathbb{E}_t R_{t+1}$. (This can also be proved directly by substituting for P_{t+1} in (6.25).) \square

6.3.4 Asset Price with constant Cash Flow Growth

With *constant dividend growth forever* (growing perpetuity), $\mathbb{E}_t D_{t+s+1} = (1+g) \mathbb{E}_t D_{t+s}$, so (6.23) becomes

$$P_t = (\mathbb{E}_t D_{t+1}) \sum_{s=1}^{\infty} \frac{(1+g)^{s-1}}{(1+R)^s} = \frac{\mathbb{E}_t D_{t+1}}{R-g}. \quad (6.28)$$

This is the “Gordon model.” The asset price (6.28) is high when: (a) dividends are expected to be high; (b) the growth rate (g) is believed to be high; and (c) when discounting (R) is low.

Inverting this formula to get the discount rate (“cost of equity capital”)

$$R = \frac{E_t D_{t+1}}{P_t} + g. \quad (6.29)$$

Example 6.16 (Asset price as sum of discounted cash flows) With $D_1 = 100$, $R = 0.1$ and $g = 2\%$,

$$P_0 = 100/(0.1 - 0.02) = 1250$$

Proof of (6.28) Write the first equality of (6.28) as $P_t = \frac{E_t D_{t+1}}{1+R} \sum_{s=0}^{\infty} (\frac{1+g}{1+R})^s$. Recall the fact that for a geometric series, $\sum_{s=0}^{\infty} r^s = 1/(1 - r)$ if $|r| < 1$. Apply this on $r = (1 + g)/(1 + R)$, to get that

$$P_t = \frac{E_t D_{t+1}}{1 + R} \frac{1}{1 - (1 + g)/(1 + R)} = \frac{E_t D_{t+1}}{R - g}.$$

□

6.3.5 Valuation Multiples

The *price-earnings ratio* (p/e) is

$$\text{“p/e”} = \frac{P}{e}, \quad (6.30)$$

where e is short for earnings per share.

If dividends are proportional to earnings, $D_t = k \times e_t$ in each period and earnings grow at the rate g , $e_{t+1} = (1 + g)e_t$, then

$$(p/e)_t = \frac{P_t}{e_t} = \frac{\overbrace{ke_{t+1}}^{D_{t+1}} / (R - g)}{e_t} = k \frac{1 + g}{R - g}.$$

Example 6.17 $R = 0.1$, $g = 2\%$ and $k = 1$ (a “cash cow”)

$$p/e = 1 \times \frac{1.02}{0.1 - 0.02} = 12.75$$

Instead, with $g = 5\%$ we have $p/e = 21$. This shows that p/e is very sensitive to assumptions about the growth rate.

The *multiples approach* is to use a comparison with a peer group (in the market or recent M&A transactions) in order price an asset (here denoted i). It has the advantage that

we do not need to specify growth or discount rate. The *equity value method* is calculate the share value of company i as

$$P_{i,t} = \left(\frac{P_t}{e_t} \right)_{peers} \times e_{i,t}, \text{ so } \frac{P_{i,t}}{e_{i,t}} = \left(\frac{P_t}{e_t} \right)_{peers}. \quad (6.31)$$

As alternatives to e , use cash flow and book value. In general, this approach makes sense if firm i and the peers have similar growth and risk, while the dividends might differ.

Remark 6.18 (*The discounted cash flow model vs. the multiples approach*) To simplify, assume $D = e$ and assume constant growth. This means that $P = (1 + g)e/(R - g)$ for both i and peers. To have $P_i/e_i = (P/e)_{peers}$ as in (6.31), the following must hold

$$\frac{P_i}{e_i} = \left(\frac{1 + g}{R - g} \right)_i = \left(\frac{1 + g}{R - g} \right)_{peers} = \left(\frac{P}{e} \right)_{peers}.$$

This shows that the discount and growth rates must be similar, or somehow counterbalance each other.

Chapter 7

Downside Risk Measures

The mean-variance framework is often criticized for failing to distinguish between the downside of the return distribution (considered to represent risk) and upside (considered to represent potential). This chapter introduces several commonly used downside risk measures and also explores more general methods of describing return distributions.

These measures can be embedded into portfolio choice models (see later chapters), and would then have to be constructed from the investor's beliefs. However, several of them can also be used as descriptive statistics, based on ex post (realized) data.

7.1 Value at Risk

The Value at Risk (VaR) measures the downside risk by focusing on a quantile of the return (or loss) distribution,

Remark 7.1 (*Quantile of a distribution*) *The 0.05 quantile is the value x such that there is only a 5% probability of a lower number, $\Pr(R \leq x) = 0.05$.*

The 95% Value at Risk ($\text{VaR}_{95\%}$) is a number such that there is only a 5% chance that the loss as a fraction of the investment (which is the negative of the return, $-R$) is larger

$$\Pr(-R \geq \text{VaR}_{95\%}) = 5\%. \quad (7.1)$$

Here, 95% is the confidence level of the VaR. For instance, with $\text{VaR}_{95\%} = 18\%$, then we are 95% sure that we will not lose more than 18% of our investment.

To work with the return distribution, not the loss distribution, we notice that (7.1) is the same as

$$\Pr(R \leq -\text{VaR}_{95\%}) = 5\%, \quad (7.2)$$

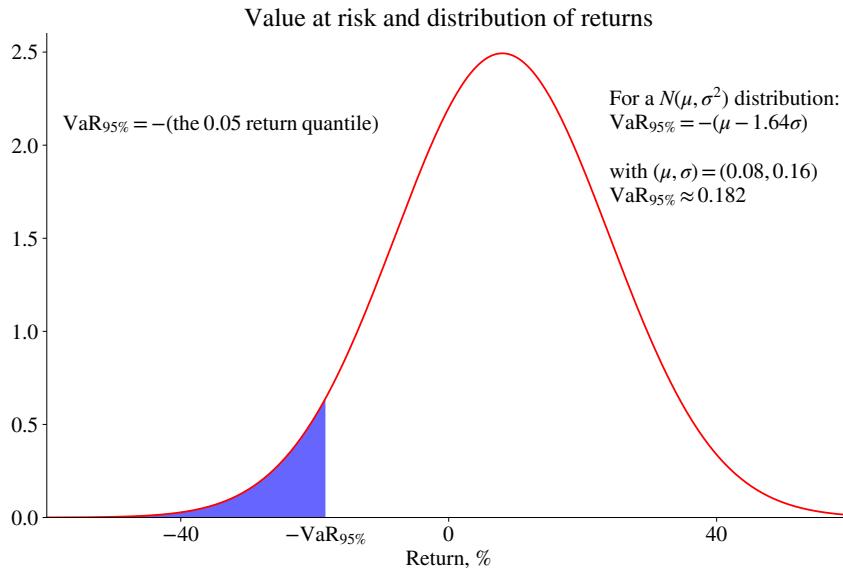


Figure 7.1: Value at risk

so $-\text{VaR}_{95\%}$ is the 0.05 *quantile* of the return distribution. More generally, the VaR for confidence level α (instead of 0.95) is

$$\text{VaR}_\alpha = -(\text{the } 1 - \alpha \text{ quantile of the } R \text{ distribution}). \quad (7.3)$$

If the return is *normally distributed*, $R \sim N(\mu, \sigma^2)$, then

$$\text{VaR}_\alpha = -(\mu + c\sigma), \quad (7.4)$$

where c is the $1 - \alpha$ quantile of a $N(0, 1)$ distribution. For instance, c is approximately $(-1.64, -1.96, -2.33)$ for the $(0.05, 0.025, 0.01)$ levels, respectively. See Figures 7.1–7.2. Since $c < 0$, the VaR is here strictly increasing the standard deviation, which will later be important when we consider portfolio choice based on a VaR.

Example 7.2 (VaR and regulation of bank capital) *Bank regulations have used 3 times the 99% VaR for 10-day returns as the required bank capital.*

Note that the return distribution depends on the *investment horizon*; therefore, the VaR is typically calculated for a particular return period (for instance, one day). Multi-period VaRs are calculated by either explicitly constructing the distribution of multi-period returns, or by making simplifying assumptions about the relation between returns in different periods (for instance, that they are uncorrelated). If the returns are iid, then

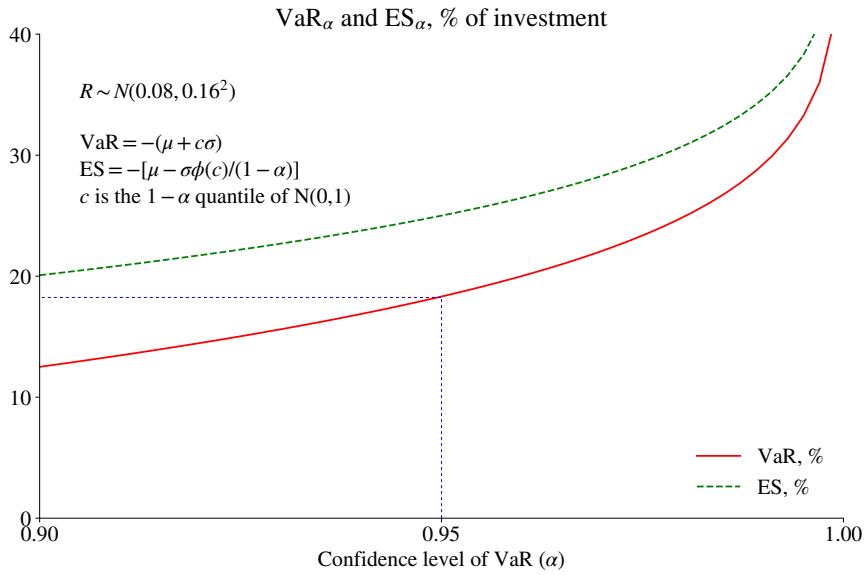


Figure 7.2: Value at risk and expected shortfall, different confidence levels

a q -period return has the mean $q\mu$ and variance $q\sigma^2$, where μ and σ^2 are the mean and variance of the one-period returns respectively. If the mean is zero, then the q -day VaR is \sqrt{q} times the one-day VaR.

Example 7.3 (*The London whale*) *The broad outline of the “London whale” (JPM) story is as follows: at the end of 2011, top management instructed the division to bring down the RWA (risk weighted asset) exposure to credit derivatives. However, (a) that would have caused high execution costs and (b) the portfolio had recently performed well. At this time a new VaR method was developed and quickly implemented. The division went on to triple the positions (and lose \$719 million in 2012Q1). Interestingly, the two VaR models (old and new) show divergent paths for the value at risk, with the new suggesting much lower risk.*

To use VaR as a risk control, or more generally, in portfolio choice, we first need to formulate beliefs (estimates) of the future return distribution. The next section is aimed at assessing how well some approaches do in capturing the future (ex post) return movements.

7.1.1 Backtesting a VaR model

Backtesting a VaR model amounts to comparing how well the VaR model can describe the 5% quantile (say) of the ex post data. This is done by determining whether the returns

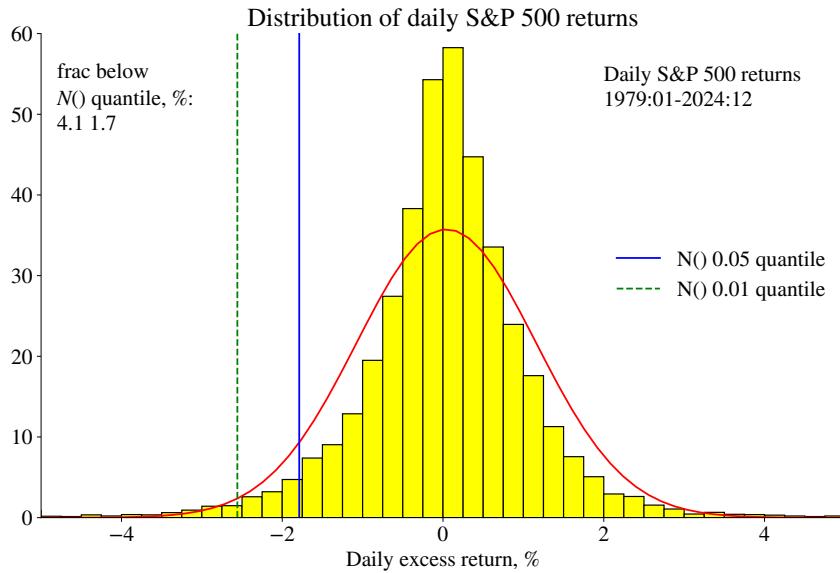


Figure 7.3: Return distribution and VaR for S&P 500

fall below the $\text{VaR}_{95\%}$ approximately 5% of the times in the sample. This could be a long sample period or a set of subsamples. In particular, a model with extended periods of under- and then over-performance which average (in the full sample) to roughly 5% is unlikely to be a useful model.

Empirical Example 7.4 Figure 7.3 shows the distribution and static VaRs (using constant parameters) for the daily S&P 500 returns for a long sample. The results indicate that the $N()$ -based model has a reasonable coverage for the 95%, but perhaps not for the 99% confidence level.

Empirical Example 7.5 Figure 7.4 shows backtesting on many subsamples. The results indicate that a static VaR model for S&P 500 has long cycles of under- and then over-performance.

7.1.2 A Simple Dynamic VaR

It is well known that financial *volatility changes over time*, which needs to be embedded in a reliable VaR model. One particularly simple approach is to estimate means and

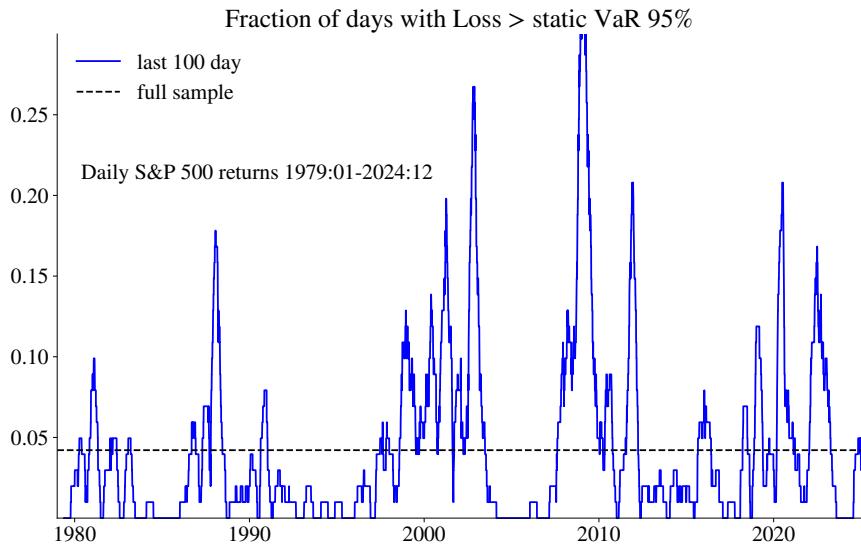


Figure 7.4: Backtesting a static VaR model on a moving data window

variances, using the recursive formulas (the RiskMetrics approach, see JP Morgan (1996))

$$\mu_t = \lambda\mu_{t-1} + (1 - \lambda)R_{t-1} \quad (7.5)$$

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1 - \lambda)(R_{t-1} - \mu_{t-1})^2, \quad (7.6)$$

where $0 < \lambda < 1$ and often high (around $0.90 - 0.95$ for daily data). The estimate of the mean is an update of yesterday's estimate, using yesterday's return for the update. This is the same as a weighted average of past returns (actually, an exponentially weighted moving average, EWMA), with recent data having higher weights than old data. The variance is a similar, with updating using the square of yesterday's surprise.

Empirical Example 7.6 Figure 7.5 illustrates the VaR calculated from a time series model for daily S&P returns. In this case, the VaR changes from day to day as both the mean return (the forecast) as well as the standard error (of the forecast error) do. Figures 7.5–7.6 show results from backtesting a VaR model which assumes that one-day returns are normally distributed, but where the mean and volatility are time varying. The evidence suggests that this model works relatively well at the 95% confidence level and that it is important to account for the time-varying volatility.

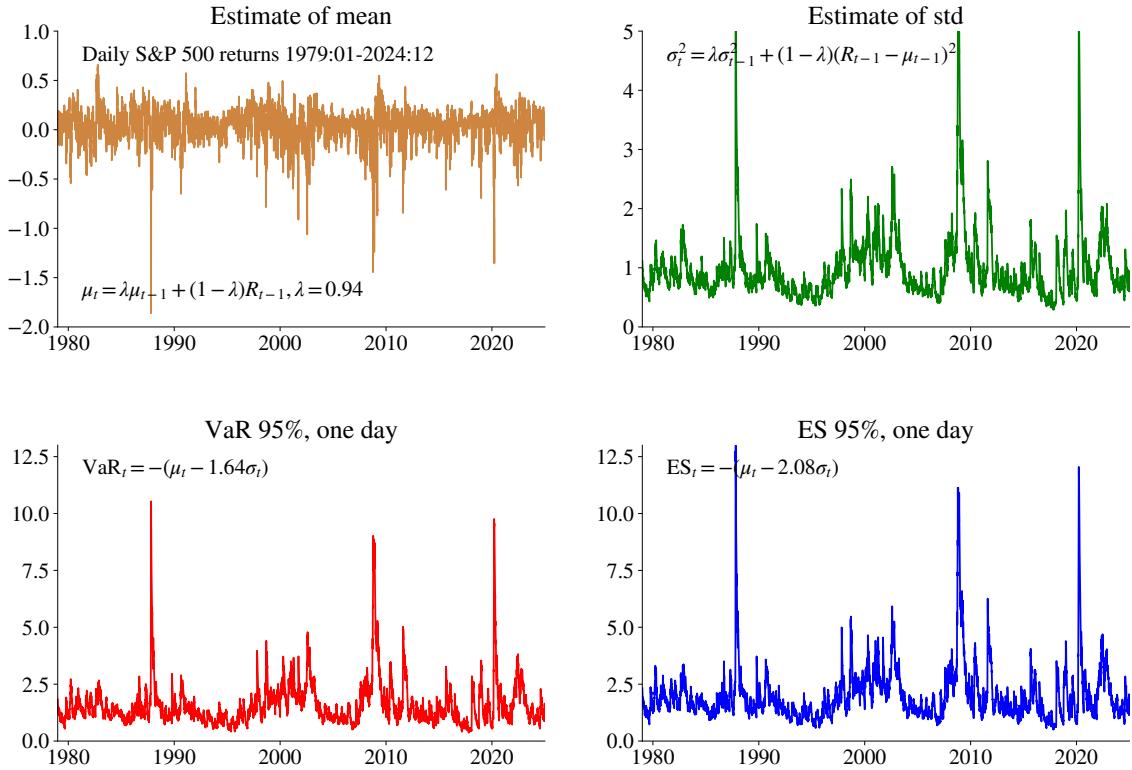


Figure 7.5: A dynamic VaR model

7.1.3 Value at Risk of a Portfolio

The general way of calculating the VaR of a portfolio is the same as for an individual asset: first calculate (or estimate) the parameters of the distribution, then find the quantile.

However, in some special cases, it is possible to directly translate the VaR values of the individual assets into a portfolio VaR.

Remark 7.7 Suppose the assets in the portfolio are jointly normally distributed with zero means, so the VaR of asset i is $\text{VaR}_i = 1.64\sigma_i$. (The index on VaR here indicates the asset, not a confidence level.) Let v be a vector where $v_i = w_i \text{VaR}_i$, where w_i is the portfolio weight. Then, $\text{VaR}_p = [v' \text{Corr}(R)v]^{1/2}$, where $\text{Corr}(R)$ is the correlation matrix of the assets. (To see this, recall that $\text{VaR}_p = 1.64\sigma_p$ and that we can calculate σ_p from the σ_i values and correlations.)

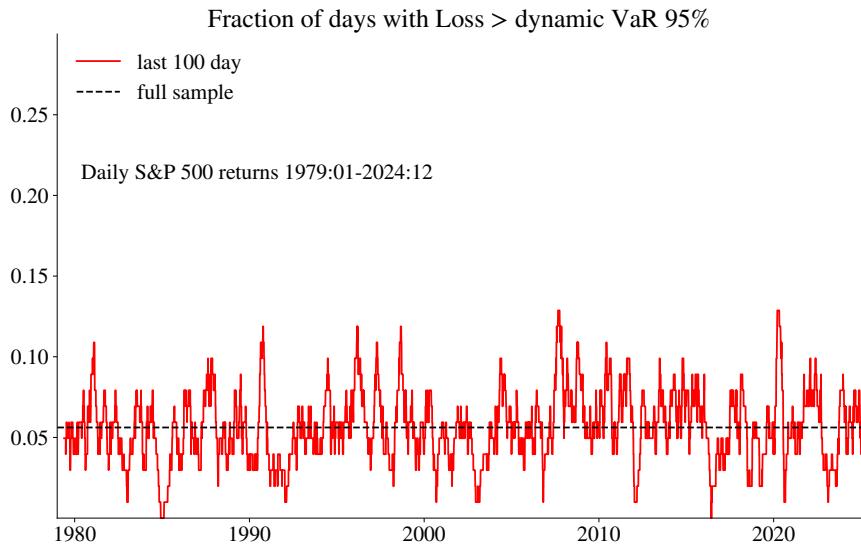


Figure 7.6: Backtesting a dynamic VaR model on a moving data window

7.1.4 Index Models for Calculating the Value at Risk

Consider a multi-index model

$$R_t = a + b' I_t + e_t, \quad (7.7)$$

where b is a $k \times 1$ vector of the b_i coefficients and I_t is a $k \times 1$ vector of indices. As usual, we assume $\text{E } e_t = 0$ and $\text{Cov}(e_t, I_t) = 0$. This model can be used to generate the inputs to a VaR model. For instance, the mean and standard deviation of the return are

$$\begin{aligned} \mu &= a + b' \text{E } I_t \\ \sigma &= \sqrt{b' \text{Cov}(I_t)b + \text{Var}(e_t)}, \end{aligned} \quad (7.8)$$

which can be used in (7.4). If the return is of a well diversified portfolio and the indices include the key market indices, then the idiosyncratic risk $\text{Var}(e)$ is close to zero. The RiskMetrics approach is to make this assumption.

A *stand-alone VaR* assesses the contribution of different factors (indices) on the overall VaR. For instance, the indices in (7.7) could include: equity indices, interest rates, exchange rates and perhaps also a few commodity indices. Then, an *equity VaR* is calculated by setting all elements in b , except those for the equity indices, to zero. Often, the intercept, a , is also set to zero. Similarly, an *interest rate VaR* is calculated by setting all elements

in b , except referring to the interest rates, to zero. And so forth for an *FX VaR* and a *commodity VaR*. Clearly, these different VaRs do not add up to the total VaR, but they still give an indication of where the main risk comes from.

If an asset or a portfolio is a non-linear function of the indices, then (7.7) can be thought of as a first-order Taylor approximation where b_i represents the partial derivative of the asset return with respect to index i . For instance, an option is a non-linear function of the underlying asset value and its volatility. This approach, when combined with the normal assumption in (7.4), is called the *delta-normal method*.

7.2 Expected Shortfall

While the value at risk is a useful risk measure, it has the strange property that it only considers whether an outcome is in the tail of the return distribution, not how far out.

In addition, the VaR concept has been criticized for having poor aggregation properties. In particular, the VaR of a portfolio is not necessarily (weakly) lower than the portfolio of the VaRs even if the assets all have the same volatility, which contradicts the notion of diversification benefits. (To get this unfortunate property, the return distributions must be heavily skewed.) See [McNeil, Frey, and Embrechts \(2005\)](#) and [Alexander \(2008\)](#) for more detailed discussions.

The expected shortfall (ES, also called conditional VaR, average value at risk and expected tail loss) has better properties. It is the expected loss when the return actually is below the VaR_α , that is,

$$\text{ES}_\alpha = -\mathbb{E}(R|R \leq -\text{VaR}_\alpha). \quad (7.9)$$

See Figure 7.7 for an illustration.

Empirical Example 7.8 *See Figure 7.5 for an empirical estimate of ES, based on the dynamic estimates of the mean and variance in (7.5)–(7.6).*

Empirical Example 7.9 *See Table 7.1 for an empirical comparison of the VaR, ES and some alternative downside risk measures (discussed below) for two stock indices.*

For a normally distributed return $R \sim N(\mu, \sigma^2)$ we have

$$\text{ES}_\alpha = -[\mu - \phi(c)\sigma/(1 - \alpha)], \quad (7.10)$$

where $\phi()$ is the pdf of a $N(0, 1)$ variable and c is the $1 - \alpha$ quantile of a $N(0, 1)$ distribution. For instance, with $1 - \alpha = 0.05$ and thus $c \approx -1.64$. In this case, the

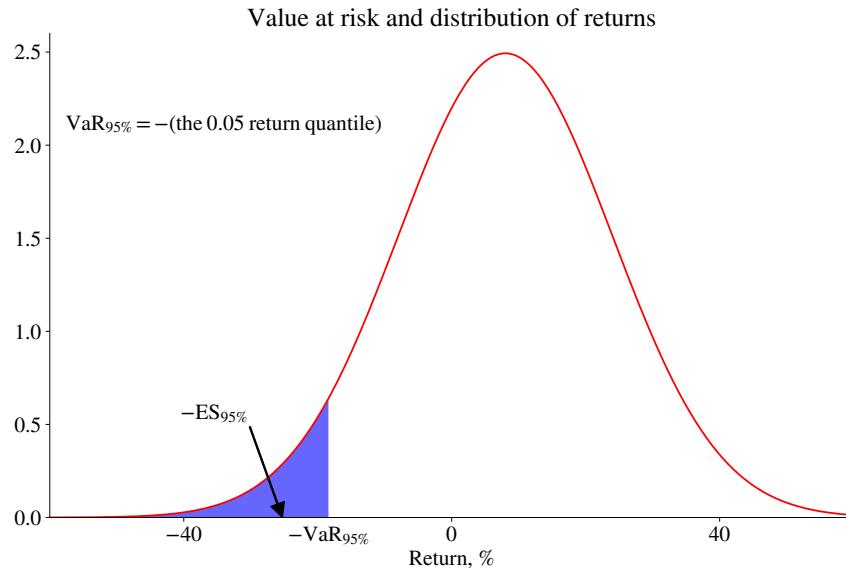


Figure 7.7: Value at risk and expected shortfall

	Small growth	Large value
Std	8.1	5.8
VaR (95%)	12.9	8.8
ES (95%)	17.9	13.1
SemiStd	5.7	3.9
Drawdown	78.4	63.2

Table 7.1: Risk measures of monthly returns of two stock indices (%), US data 1970:01-2024:12.

ES is strictly increasing in the standard deviation, which will later be important when we consider portfolio choice. See Figure 7.2 for an illustration.

Example 7.10 (ES) If $\mu = 8\%$ and $\sigma = 16\%$, the 95% expected shortfall is $ES_{95\%} = -(0.08 - 2.08 \times 0.16) \approx 0.25$ (since $\phi(-1.64)/0.05 \approx 2.08$) and the 97.5% expected shortfall is $ES_{97.5\%} = -(0.08 - 2.34 \times 0.16) \approx 0.29$ (since $\phi(-1.96)/0.025 \approx 2.34$)

Proof of (7.10). If $x \sim N(\mu, \sigma^2)$, then it is well known that $E(x|x \leq b) = \mu - \sigma\phi(b_0)/\Phi(b_0)$ where $b_0 = (b - \mu)/\sigma$ and where $\phi()$ and $\Phi()$ are the pdf and cdf of a $N(0, 1)$ variable respectively. To apply this, use $b = -\text{VaR}_\alpha = \mu + c\sigma$ so $b_0 = c$. Clearly, $\Phi(c) = 1 - \alpha$, so $E(R|R \leq -\text{VaR}_\alpha) = \mu - \sigma\phi(c)/(1 - \alpha)$. Multiply by -1 . \square

To estimate the average shortfall from a sample, calculate the average $-R_t$ for observations where $R_t \leq -\text{VaR}_\alpha$

$$\text{ES}_\alpha = -\sum_{t=1}^T \delta_t R_t / (\sum_{t=1}^T \delta_t), \text{ where } \delta_t = \delta(R_t \leq -\text{VaR}_\alpha) \quad (7.11)$$

and where $\delta(q) = 1$ if q is true and zero otherwise. (In this expression δ_t is a dummy variable whose value depends on the $\delta()$ function.) This can be used in backtesting an ES model.

Empirical Example 7.11 See Figure 7.8 for a back testing of the dynamic ES model previously shown in Figure 7.5.

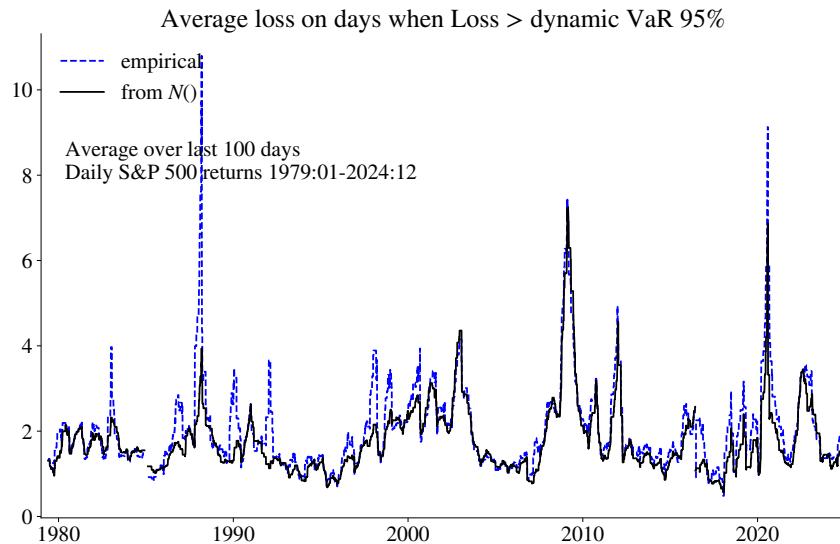


Figure 7.8: Backtesting a dynamic ES model on a moving data window

7.3 Target Semi-Variance

The target semi-variance (TSV, also called the lower partial 2nd moment) is defined as

$$\sigma_-^2(h) = E[\min(R - h, 0)^2], \quad (7.12)$$

where h is a “target level” chosen by the investor. Also, $\sqrt{\sigma_-^2(\mu)}$ is called the semi-standard deviation. In comparison with the variance, $\sigma^2 = E(R - E R)^2$, the target semi-variance differs in two aspects: (i) it uses the target level h as a reference point instead of the mean

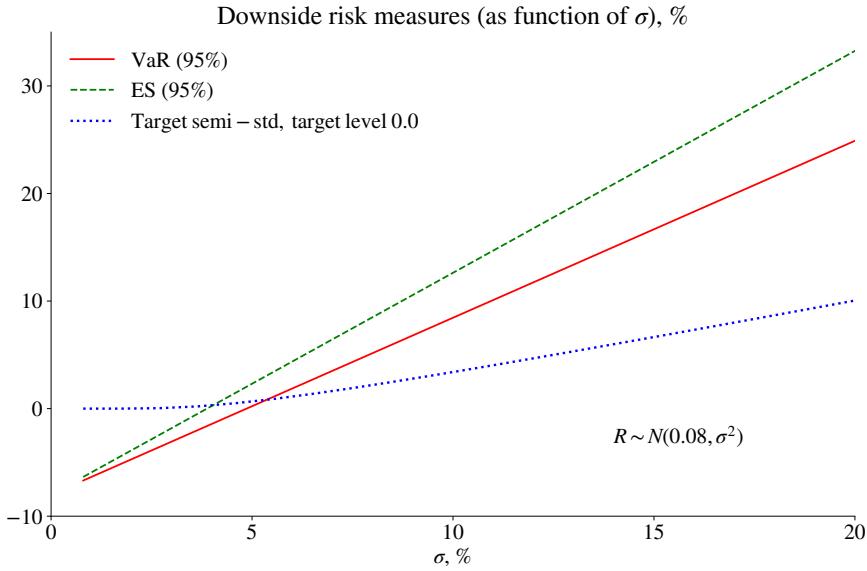


Figure 7.9: Downside risk measures as functions of the standard deviation for a $N(\mu, \sigma^2)$ variable

μ : and (ii) only negative deviations from the reference point are given any weight (see [Bawa and Lindenberg \(1977\)](#) and [Nantell and Price \(1979\)](#)).

For a normally distributed variable, the target semi-variance $\sigma_-^2(h)$ is increasing in the standard deviation, see Remark 7.13, which will later be important when we consider portfolio choice. See also Figure 7.9 for an illustration.

To estimate the target semi-variance from the empirical return distribution (for back-testing), use

$$\sigma_-^2(h) = \frac{1}{T} \sum_{t=1}^T \delta_t (R_t - h)^2, \text{ where } \delta_t = \delta(R_t \leq h) \quad (7.13)$$

and where $\delta(q) = 1$ if q is true and zero otherwise.

Remark 7.12 (*Alternative scaling of $\sigma_-^2(h)$) Some analysts define $\sigma_-^2(h)$ by just including those observations for which $R_t \leq h$. This means multiplying (7.13) by $T / \sum_{t=1}^T \delta(R_t \leq h)$, which is actually estimating $E[(R - h)^2 | R_t \leq h]$.

Remark 7.13 (Target semi-variance calculation for normally distributed variable*) For an $N(\mu, \sigma^2)$ variable, target semi-variance around the target level h is

$$\sigma_-^2(h) = \sigma^2 a \phi(a) + \sigma^2 (a^2 + 1) \Phi(a), \text{ where } a = (h - \mu)/\sigma,$$

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

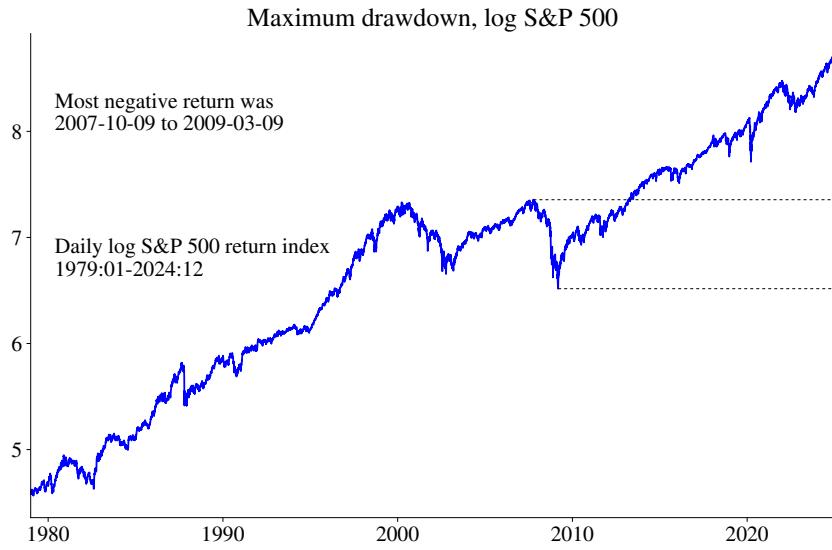


Figure 7.10: Maximum drawdown over the full sample

$\sigma^2_-(\mu) = \sigma^2/2$. It is straightforward to show that $d\sigma^2_-(h)/d\sigma = 2\sigma\Phi(a)$, so the target semi-variance is a strictly increasing function of the standard deviation.

Remark 7.14 (*Sortino ratio*) The Sortino ratio is an alternative to the Sharpe ratio as a measure of performance. It is $(E R - h)/\sqrt{\sigma^2_-(h)}$.

Empirical Example 7.15 See Table 7.2 for an empirical rank correlation of the different risk measures for 25 FF portfolios. Most of the risk measures have strong rank correlations, meaning that they give very similar ranking of “riskiness” of these 25 assets. However, max drawdown is different, mostly likely since it is focused on the extreme left tail of the distribution.

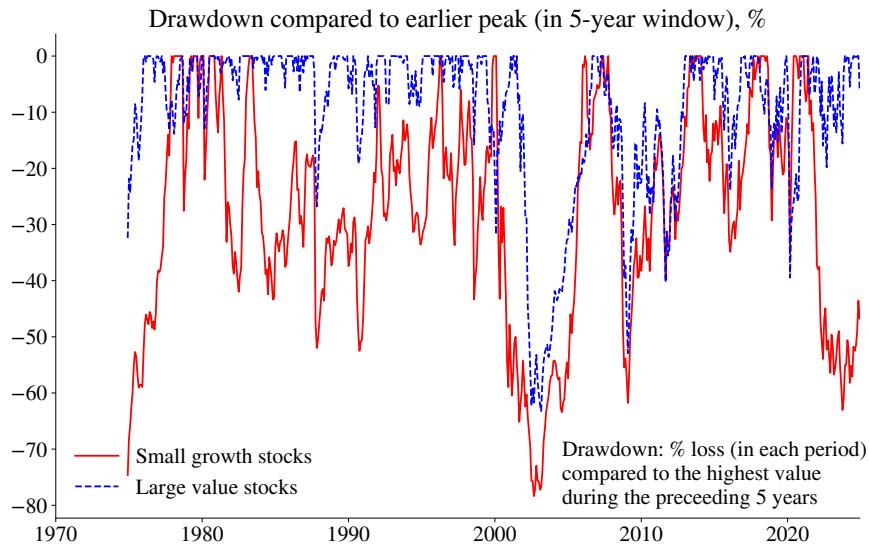


Figure 7.11: Drawdown

	Std	VaR (95%)	ES (95%)	SemiStd	Drawdown
Std	1.00	0.95	0.97	0.98	0.61
VaR (95%)	0.95	1.00	0.95	0.96	0.65
ES (95%)	0.97	0.95	1.00	0.97	0.64
SemiStd	0.98	0.96	0.97	1.00	0.60
Drawdown	0.61	0.65	0.64	0.60	1.00

Table 7.2: Correlation of rank of risk measures across the 25 FF portfolios (%), US data 1970:01-2024:12. The VaR and ES are based on the empirical return distribution. The max drawdown is calculated over a moving 5-year data window.

7.4 Maximum Drawdown

An alternative measure is the (percentage) *maximum drawdown* over a given horizon, for instance, 5 years, say. This is the largest loss from peak to bottom within the given horizon, see Figure 7.10. This is a useful measure when the investor do not know exactly when he/she has to exit the investment—since it indicates the worst (peak to bottom) outcome over the sample.

Empirical Example 7.16 See Figure 7.11 for a comparison of the max drawdown of two return series. The results suggest that small growth stocks are considerably more risky than large value stocks.

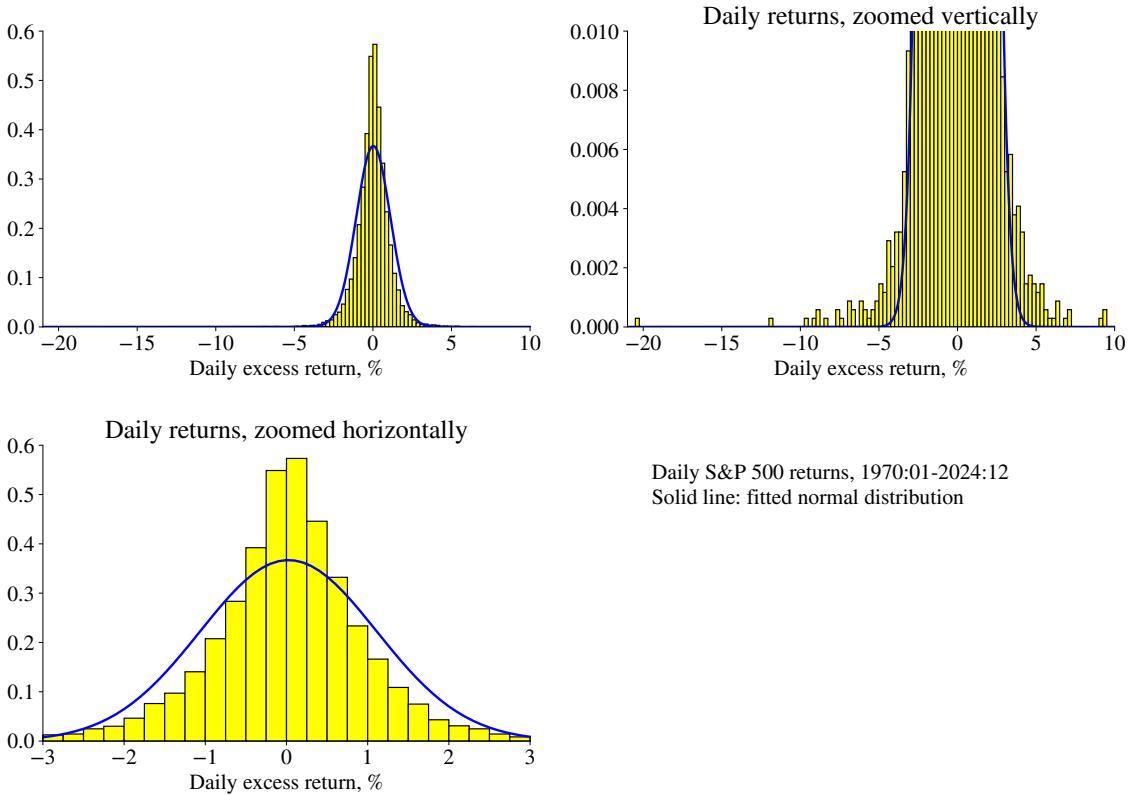


Figure 7.12: Distribution of daily S&P returns

7.5 Empirical Return Distributions

Are returns normally distributed? Mostly not, but it depends on the asset type and on the data frequency. Options returns have very non-normal distributions (in particular, since the return is -100% on many expiration days). Stock returns are typically distinctly non-normal at short horizons, but may appear approximately normal over longer horizons. This may (or may not) carry over to the beliefs held by investors.

To assess the normality of realized 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.

Remark 7.17 (*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 data. Each point in the *QQ* plot shows a specific percentile (quantile) according to the empirical as well as

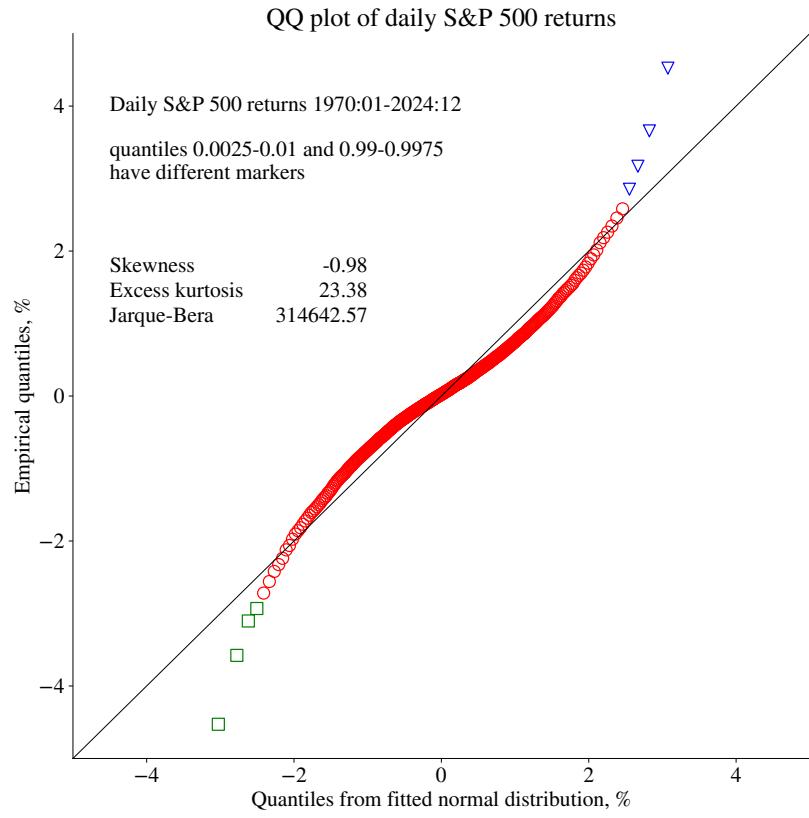


Figure 7.13: Quantiles of daily S&P returns

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

Empirical Example 7.18 See Figures 7.12–7.14 for empirical histograms and QQ-plots of S&P 500 returns. It is observed, among other findings, that empirical returns distributions exhibit more extreme negative returns than a normal distribution would suggest, and that the return distribution looks closer to a normal distribution as the return horizon increases.

These methods can be applied to both data on returns and to residuals from a statistical model. For instance, the $(R_t - \mu_t)/\sigma_t$ where (μ_t, σ_t) are estimated by (7.5)–(7.6).

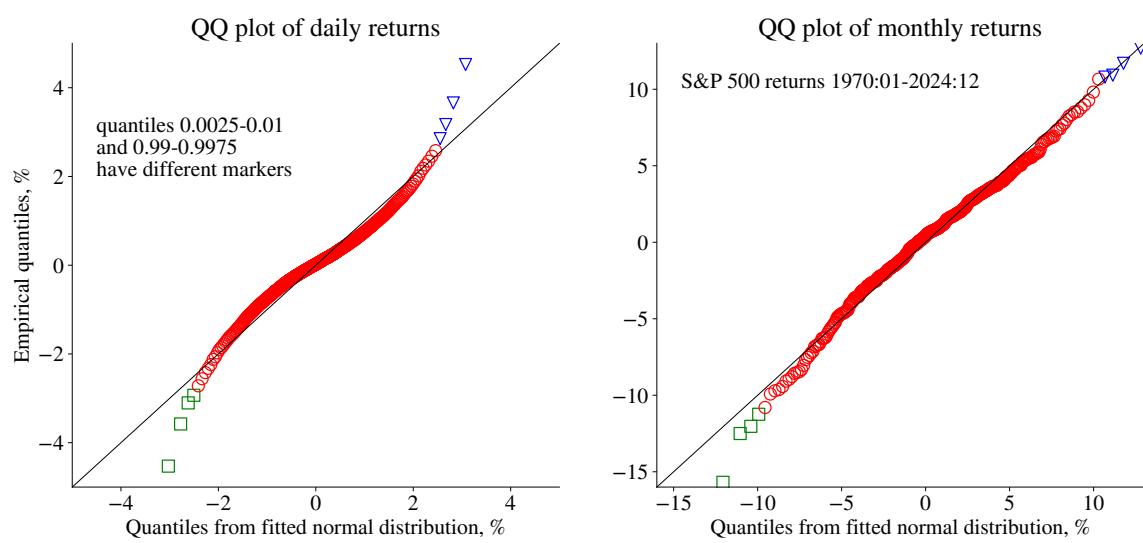


Figure 7.14: Distribution of S&P returns (different horizons)

Chapter 8

Utility-Based Portfolio Choice

8.1 Utility Functions and Risky Investments

Any model of portfolio choice must embody a notion of “what is best?” In finance, that often means a portfolio that strikes a good balance between expected return and its variance. However, in order to make sense of that idea—and to go beyond it—we must refer to economic utility theory.

8.1.1 Specification of Utility Functions

In finance, the key features of utility functions are as follows. *First*, utility is a function of a scalar argument, $U(x)$. This argument (x) can be end-of-period wealth, a consumption basket or the *real* (inflation adjusted) portfolio return. In one-period investment problems, this choice of x is irrelevant since consumption equals wealth, which is proportional to the portfolio return.

Second, uncertainty is incorporated by letting investors maximize expected utility, $E U(x)$. The reason is that returns (and therefore wealth and consumption) are uncertain. Hence, we need a way to rank portfolios at the time of investment, before the uncertainty is resolved. For instance, if there are S possible states with outcomes x_1, x_2, \dots, x_S and probabilities $\pi_1, \pi_2, \dots, \pi_S$, then expected utility is

$$E U(x) = \sum_{s=1}^S \pi_s U(x_s). \quad (8.1)$$

The outcomes could represent portfolio returns, which depend on the state *and* the portfolio weights. That is, x_s is not a fixed list, but rather functions of the investor’s choices. As usual, the expectation is based on the investor’s beliefs.

Example 8.1 ($E U(W)$) Suppose there are two states of the world: W (wealth) will be

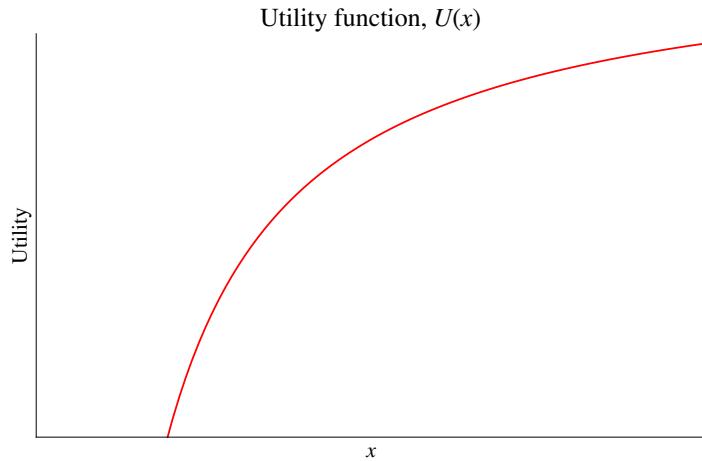


Figure 8.1: A utility function

either 0.85 or 1.15 with probabilities 1/3 and 2/3. If $U(W) = \ln W$, then $\text{E } U(W) = 1/3 \times \ln 0.85 + 2/3 \times \ln 1.15 \approx 0.039$. If the investor had picked another portfolio the outcomes would instead be 0.9 and 1.05, with expected utility $1/3 \times \ln 0.9 + 2/3 \times \ln 1.05 \approx -0.003$.

Third, the functional form of the utility function is such that more is better (the function is increasing) and the function is concave. The latter means that investors are risk averse. See Figure 8.1 for an illustration.

Remark 8.2 (*Expected utility theorem**) Expected utility, $\text{E } U(W)$, is the right thing to maximize if the investors' preferences $U(W)$ are (1) complete: can rank all possible outcomes (that is, we know what we like); (2) transitive: if A is better than B , and B is better than C , then A is better than C (a form of consistency); (3) independent: if X and Y are equally preferred, and Z is some other outcome, then the following gambles are equally preferred (a) X with prob π and Z with prob $1 - \pi$ and (b) Y with prob π and Z with prob $1 - \pi$ (this is the key assumption); and (4) such that every gamble has a certainty equivalent (a non-random outcome that gives the same utility, fairly trivial).

8.1.2 Basic Properties of Utility Functions: (1) More is Better

The idea that *more is better* (non-satiation) is trivial. It means that the utility function is upward sloping. If $U(W)$ is differentiable, then this is the same as marginal utility being positive, $U'(W) > 0$.

Example 8.3 (*Logarithmic utility*) $U(W) = \ln W$ so $U'(W) = 1/W > 0$ (assuming $W > 0$).

8.1.3 Basic Properties of Utility Functions: (2) Risk is Bad

With expected utility, *risk aversion* (uncertainty is considered to be bad) is captured by the concavity of the utility function.

In contrast, a linear utility function implies risk-neutrality, which we rule out because investors appear to care about risk. (Some may seem not to do so, but they are often not gambling with their own money.)

As an example, consider Figure 8.2. It shows a case where the portfolio (or wealth, or consumption,...) of an investor will pay either x^- or x^+ with equal probabilities. The utility function embodies risk aversion since the utility of getting the expected payoff for sure, $U(E x)$, is higher than the expected utility from owning the uncertain asset

$$U(E x) > 0.5U(x^-) + 0.5U(x^+) = E U(x). \quad (8.2)$$

Remark 8.4 (**Risk aversion and “marginal utility”*) Rearranging (8.2) gives

$$U(E x) - U(x^-) > U(x^+) - U(E x),$$

which says that moving from a low to a mid value of x (left hand side) counts for more than moving from a mid value to a high value (right hand side). Another way of phrasing the same thing is that a poor person appreciates an extra dollar more than a rich person. This is a key property of a concave utility function.

The lowest price, P , the investor is willing to sell this risky portfolio for is the certain amount that gives the same utility as $E U(x)$, that is, the value of P that solves the equation

$$U(P) = E U(x). \quad (8.3)$$

This price (P), called the *certainty equivalent* of the portfolio, is less than the expected payoff

$$P < E x = 0.5x^- + 0.5x^+. \quad (8.4)$$

(The result follows from $U(P) < U(E x)$ and that $U()$ is an increasing function.) Again, see Figure 8.2 for an illustration.

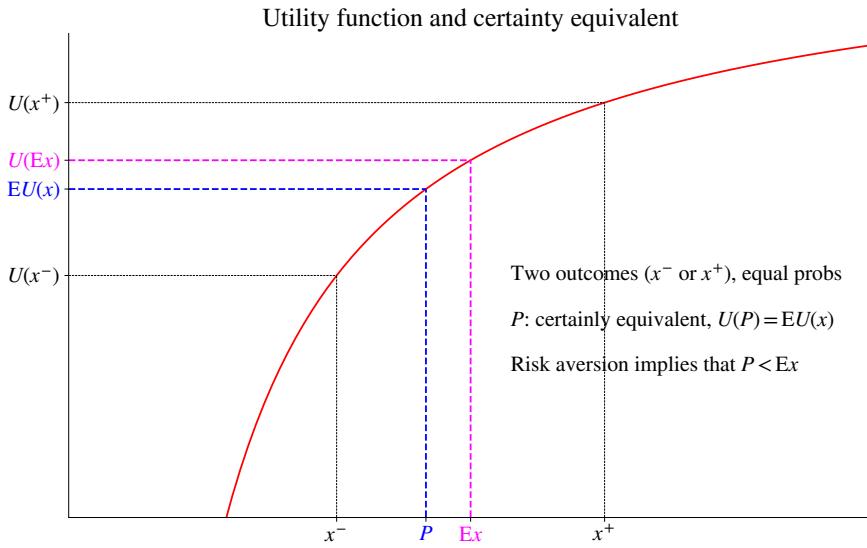


Figure 8.2: Certainty equivalent

Example 8.5 (*Certainty equivalent*) Suppose you have a CRRA utility function, $x^{1-\gamma}/(1-\gamma)$, and own an asset that gives either 0.85 or 1.15 with equal probabilities. What is the certainty equivalent (that is, the lowest price you would sell this asset for)? The answer is the P that solves

$$\frac{P^{1-\gamma}}{1-\gamma} = 0.5 \frac{0.85^{1-\gamma}}{1-\gamma} + 0.5 \frac{1.15^{1-\gamma}}{1-\gamma}.$$

For instance, with $\gamma = 0, 2, 5, 10$, and 25 we have $P \approx 1, 0.977, 0.947, 0.912$, and 0.875. Notice that P goes from the average payoff (1) to the lowest outcome (0.85) as risk aversion increases.

This means that the *expected net return* on the risky portfolio that the investor demands is

$$E R_x = \frac{E x}{P} - 1 > 0, \quad (8.5)$$

which is greater than zero. This “required return” is higher if the investor is very risk averse (since P is lower). Notice that this analysis applies to the portfolio return (or wealth, or consumption,...), that is, the argument of the utility function—not to any individual asset. To analyse an individual asset, we need to study how it changes the argument of the utility function, so the covariances with the other assets play a key role.

Example 8.6 (*Risk premium in a simple case*) Using the $k = 2$ case in Example 8.5 we get the expected net return (8.5) $1/0.977 - 1 \approx 2.4\%$, since $E x = 1$. Instead, with

$k = 25$ we get $1/0.875 - 1 \approx 14.3\%$.

8.2 Utility-Based Portfolio Choice and MV Frontiers

8.2.1 Utility-Based Portfolio Choice with a Single Risky Asset

Suppose the investor maximizes expected utility from the portfolio return by choosing between a risky and a risk-free asset

$$\max_v \mathbb{E} U(R_p), \text{ with } R_p = vR^e + R_f, \quad (8.6)$$

where R^e denotes the excess return of the risky asset.

The first order condition with respect to the weight on risky assets is

$$\begin{aligned} \frac{d \mathbb{E} U(vR^e + R_f)}{dv} &= 0 \text{ or} \\ \mathbb{E}[U'(R_p)R^e] &= 0, \end{aligned} \quad (8.7)$$

where $U'(R_p)$ is the marginal utility evaluated at $R_p = vR^e + R_f$. Notice that the order of \mathbb{E} and ∂ are different in the first and second expressions. This is permissible since \mathbb{E} defines a sum, and a derivative of a sum is the sum of derivatives, see below for a remark. Also, notice that the second expression is the expectation of the *product* of marginal utility and the excess return.

Remark 8.7 (*Stochastic discount factor models**) Equation (8.7) is on the same form as a stochastic discount factor (SDF) to asset pricing, where $\mathbb{E} MR^e = 0$ is a key condition.

As an example, with a CRRA utility function the first order condition (8.7) can be written

$$\mathbb{E} \frac{R^e}{(vR^e + R_f)^\gamma} = 0, \quad (8.8)$$

which is an expectation of a non-linear expression.

Remark 8.8 (*Interchanging the order of \mathbb{E} and ∂) Consider expected utility in (8.1) and let the outcomes be functions of a portfolio weight v , as in $x_s(v)$. Differentiating wrt. v then gives

$$\frac{d \mathbb{E} U(x)}{dv} = \sum_{s=1}^S \pi_s \frac{dU(x_s)}{dx} \frac{dx_s}{dv} = \mathbb{E} \frac{dU(x)}{dv},$$

where the last expression uses a short hand notation for how utility depends on v .

Clearly, the first order condition (8.7) defines one equation in one unknown (v). Unfortunately, it can be complicated. The expectation requires integration and marginal utility might be non-linear, together requiring numerical methods. Explicit solutions are only possible in a few simple cases.

Example 8.9 (*Portfolio choice with log utility and two states*) Suppose $U(R_p) = \ln(R_p + 1)$, and that there is one risky asset and a risk-free asset. The excess return on the risky asset R^e is either a low value R^{e-} (with probability π) or a high value R^{e+} (with probability $1 - \pi$). The optimization problem is

$$\max_v \pi \ln(vR^{e-} + R_f + 1) + (1 - \pi) \ln(vR^{e+} + R_f + 1).$$

The first order condition ($\partial \mathbb{E} U(R_p)/\partial v = 0$) is

$$0 = \pi \frac{R^{e-}}{vR^{e-} + R_f + 1} + (1 - \pi) \frac{R^{e+}}{vR^{e+} + R_f + 1}, \text{ so}$$

$$v = -(1 + R_f) \frac{\pi R^{e-} + (1 - \pi) R^{e+}}{R^{e-} - R^{e+}}.$$

See Figure 8.3 for an illustration. As a special case, consider $R_f = 0$ and $R^{e-} = -1$, so the bad state means losing the entire investment (bet). In this case, $v = (1 - \pi) - \pi/R^{e+}$. This is often used to illustrate betting/log utility (the “Kelly” criterion).

Remark 8.10 (*When to put all investments in the risk-free asset?*) Suppose $v = 0$ would be an optimal decision, then the portfolio return equals the risk-free rate which is not random. The first order condition (8.7) can then be written

$$\mathbb{E}[U'(R_f)R^e] = U'(R_f)\mathbb{E} R^e = 0$$

which holds only if $\mathbb{E} R^e = 0$. This shows that it is optimal to make zero investment in the risky asset when its expected excess return is zero, which is intuitively reasonable.

8.2.2 Utility-Based Portfolio Choice with Several Risky Assets

We now consider the case with n risky assets and a risk-free asset. The optimization problem is

$$\max_{v_1, v_2, \dots} \mathbb{E} U(R_p), \text{ where} \tag{8.9}$$

$$R_p = \sum_{i=1}^n v_i R_i^e + R_f. \tag{8.10}$$

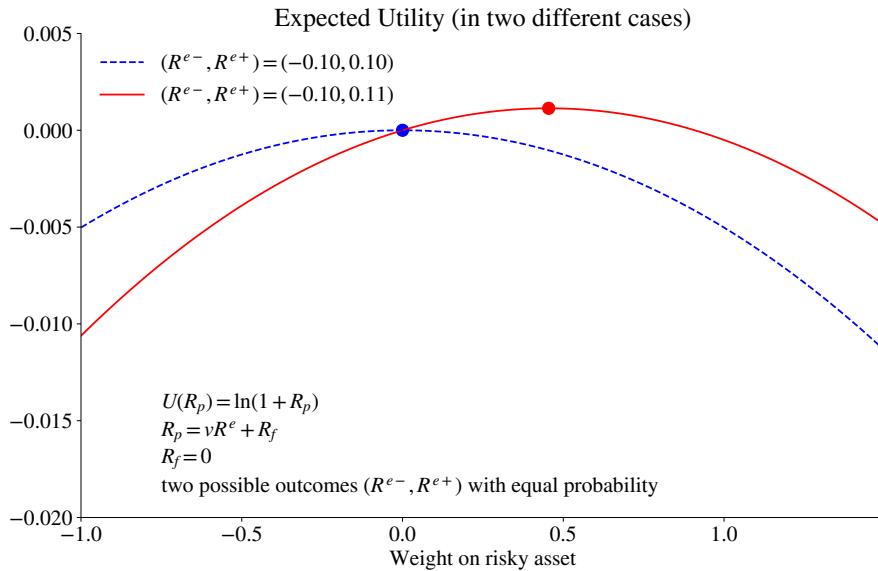


Figure 8.3: Example of portfolio choice with a log utility function

The first order conditions for the portfolio weights on the risky assets are

$$E[U'(R_p)R_i^e] = 0 \text{ for } i = 1, 2, \dots, n. \quad (8.11)$$

This is similar to the case with one risky asset, but now there are n (non-linear) equations in n unknowns: v_1, v_2, \dots, v_n . For instance, with a CRRA utility function we get

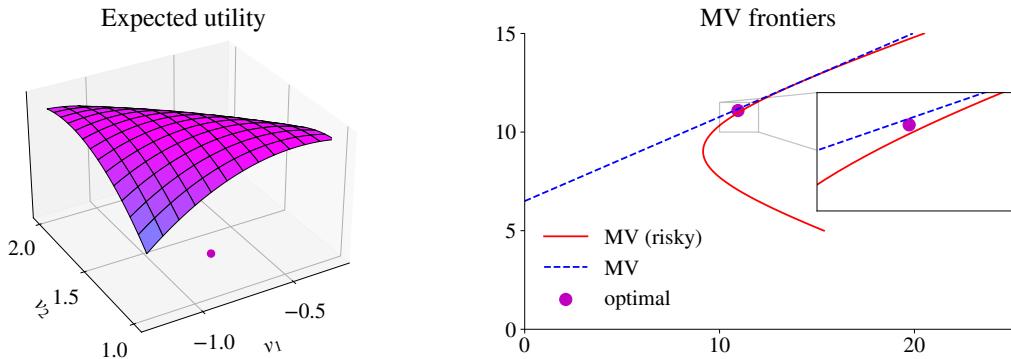
$$E \frac{R_i^e}{(\sum_{i=1}^n v_i R_i^e + R_f)^\gamma} = 0 \text{ for } i = 1, 2, \dots, n. \quad (8.12)$$

Notice that calculating the expectation involves integrating over n dimensions. See Figures 8.4 –8.6 for illustrations. The (explicit or numerical) solution is often hard to obtain—so it would be convenient if we could simplify the problem.

8.2.3 Is the Optimal Portfolio on the Mean-Variance Frontier?

There are important cases where we can side-step most of the problems with solving the general portfolio choice problem (8.11). In particular, sometimes we can show that the portfolio will be on the mean-variance frontier.

The optimal portfolio is on the mean-variance frontier when optimisation problem can be rewritten as a function in terms of the expected return (positive derivative) and the



$$\text{Utility function: } (1+R)^{1-\gamma}/(1-\gamma), \gamma=5$$

Two risky assets (1 and 2) and one risk-free asset

Three states with equal probability, returns in %:

	state 1	state 2	state 3
asset 1	-3.0	8.0	20.0
asset 2	-4.0	22.0	15.0
risk-free	6.5	6.5	6.5

	optimal	MV
asset 1	-0.73	-0.34
asset 2	1.32	0.94
risk-free	0.41	0.41

Figure 8.4: Example of when the optimal portfolio is (slightly) off the MV frontier

variance (negative derivative) only

$$\max_v V(\mathbb{E} R_p, \text{Var}(R_p)), \quad (8.13)$$

where $\partial V() / \partial \mathbb{E} R_p > 0$ and $\partial V() / \partial \text{Var}(R_p) < 0$. In this case, we should interpret $V()$ as incorporating the preferences, all relevant restrictions and also the features of the return distribution. See [Danthine and Donaldson \(2005\)](#) 4–6 and [Huang and Litzenberger \(1988\)](#) 4–5 for more detailed discussions.

This means that Figure 8.5 summarizes the preferences and the possibility set (everything below the CML).

In contrast, Figures 8.4 and 8.6 show examples when (8.13) does not hold. For instance, the preferences may include concerns about the skewness of the portfolio returns, at the same time as the return distribution exhibits non-trivial skewness. In such cases, the optimal portfolio may be off the MV frontier.

8.2.4 Special Cases

This section outlines special cases when the utility-based portfolio choice problem can be rewritten in terms of mean and variance only as in (8.13), so that the optimal portfolio is

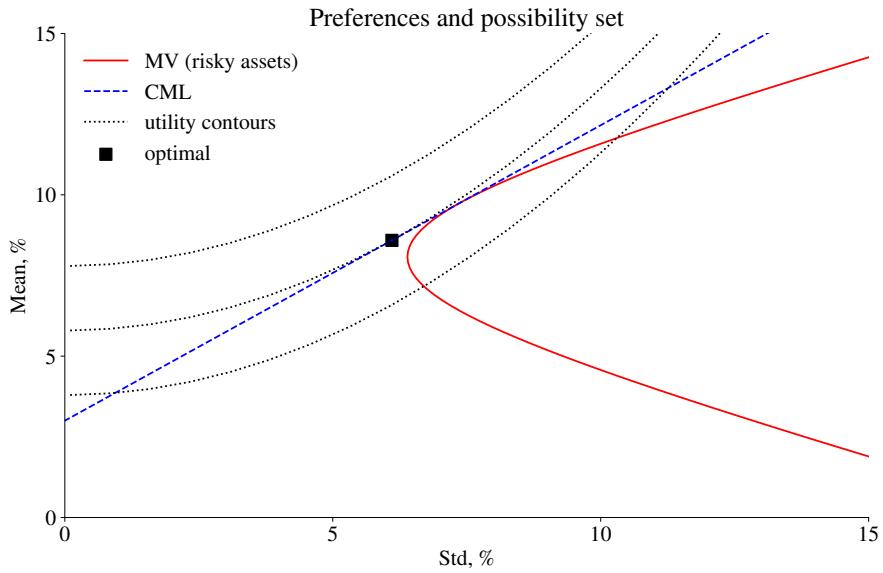


Figure 8.5: Iso-utility curves. The calculations use the properties of the assets in Table 8.1.

	$\mu, \%$	Σ, bp		
		A	B	C
A	11.5	166	34	58
B	9.5	34	64	4
C	6.0	58	4	100

Table 8.1: Characteristics of the three assets in some examples. Notice that $\mu, \%$ is the expected return in % (that is, $\times 100$) and Σ, bp is the covariance matrix in basis points (that is, $\times 100^2$).

on the mean-variance frontier.

Case 1: Mean-Variance Utility

We already know that if the investor maximizes $E R_p - \text{Var}(R_p)k/2$, then the optimal portfolio is on the mean-variance frontier. Clearly, this is the same as assuming that the utility function is $U(R_p) = R_p - (R_p - E R_p)^2 k/2$. (Evaluate $E U(R_p)$ to see this.)

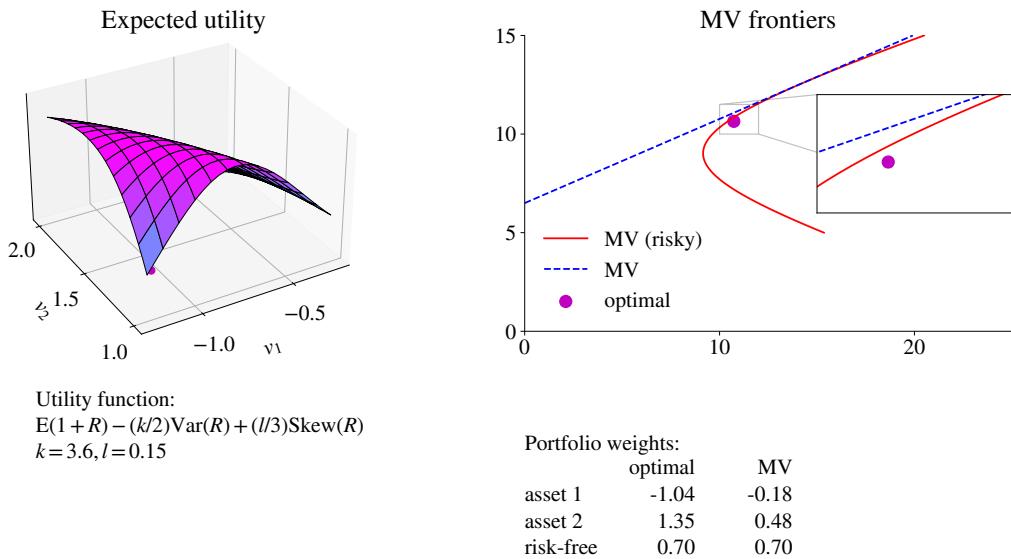


Figure 8.6: Example of when the optimal portfolio is (slightly) off the MV frontier

Case 2: Quadratic Utility

If utility is quadratic in the return (or equivalently, in wealth)

$$U(R_p) = R_p - kR_p^2/2, \quad (8.14)$$

then expected utility can be written

$$E U(R_p) = E R_p - k[\text{Var}(R_p) + (E R_p)^2]/2 \quad (8.15)$$

since $\text{Var}(R_p) = E R_p^2 - (E R_p)^2$. For $k > 0$ this function is decreasing in the variance, and increasing in the mean return as long as $k E R_p < 1$. In this case, the optimal portfolio is on the mean-variance frontier.

The main drawback of this utility function is that we have to make sure that we are on the portion of the curve where expected utility is increasing in $E R_p$ (below the so called “bliss point”). Moreover, the quadratic utility function has the strange property that the amount invested in risky assets decreases as wealth increases (increasing absolute risk aversion).

Case 3: Normally Distributed Returns

When the distribution (as perceived by the investor) of the investable assets is jointly normal, then all portfolio returns are normally distributed. In this case, maximizing $E U(R_p)$ will result in a mean variance portfolio, at least if the utility function is strictly increasing and concave. The reason is that the expected value of such a utility function is increasing in the mean and decreasing in the volatility, and that these two moments fully describe a normal distribution.

Proposition 8.11 *If the returns of all investable assets are jointly normally distributed and the utility function is strictly increasing and concave, maximizing $E U(R_p)$ will result in a mean variance portfolio.*

Actually, the same result holds for any elliptical distribution with finite second moments (for instance, a multivariate t -distribution with more than 2 degrees of freedom), see Owen and Rabinovitch (1983).

Normally distributed returns should be considered as just an approximation for three reasons. *First*, limited liability means that the net return can never be below -100% (the asset price cannot be negative). However, such returns are possible in a normal distribution, although they may have very low probabilities. *Second*, empirical evidence suggests that most asset returns have distributions with fatter tails and more skewness than implied by a normal distribution, especially when the returns are measured over short horizons. *Third*, some assets with non-linear payoffs, like options, have return distributions that must be non-normal.

As an example of what happens when we combine a normal distribution with a valid utility function, consider the next propositions. Further examples and applications, for instance, using the Telser criterion, are discussed in a separate section below.

Proposition 8.12 *If returns are normally distributed, then maximizing the expected value a utility function with constant absolute risk aversion (CARA) $k > 0$, $U(R_p) = -\exp(-R_p k)$, is the same as solving a mean-variance problem. (The proof is in the appendix.)*

Case 4: CRRA Utility and Lognormally Distributed Portfolio Returns

Proposition 8.13 *Consider a CRRA utility function, $(1 + R_p)^{1-\gamma}/(1 - \gamma)$, and suppose all log portfolio returns, $r_p = \ln(1 + R_p)$, are normally distributed. The solution is then, once again, on the mean-variance frontier. (The proof is in the appendix.)*

This result is especially useful in analysis of multi-period investments. Notice, however, that this should be thought of as an approximation since $1 + R_p = \alpha(1 + R_1) + (1 - \alpha)(1 + R_2)$ is not lognormally distributed even if both R_1 and R_2 are.

Proof (of Proposition 8.11) First, joint normality of all returns means that portfolio returns are normally distributed. Second, a normal distribution $N(\mu_p, \sigma_p^2)$ is fully described by the mean and variance. Third (following Ingersoll (1987)), write $R_p \sim N(\mu_p, \sigma_p^2)$ as $\mu_p + \sigma_p z$ where $z \sim N(0, 1)$. Expected utility is then

$$E U(R_p) = E U(\mu_p + \sigma_p z).$$

The derivative with respect to μ_p is

$$\partial E U(\mu_p + \sigma_p z)/\partial \mu_p = E U'(\mu_p + \sigma_p z),$$

which is positive since $U'(0) > 0$. Also, the derivative with respect to σ_p is

$$\partial E U(\mu_p + \sigma_p z)/\partial \sigma_p = E[U'(\mu_p + \sigma_p z)z]. \quad (*)$$

This must be negative since z has a symmetric distribution around zero, so for every term where $z = x$, there is also a term with $z = -x$. The expectation can thus be calculated as $\int_0^\infty [U'(\mu_p + \sigma_p x) - U'(\mu_p - \sigma_p x)]x\phi(x)dx$, where $\phi(x)$ is the $N(0, 1)$ pdf. The term in square brackets is negative since marginal utility is decreasing, that is, utility is concave ($U''(0) < 0$). For an alternative proof, using a Taylor series expansion of $E U()$, see Pennacchi (2008) 2. \square

8.2.5 Application of Normal Returns

This section gives a few examples of how fairly non-standard preferences, combined with normally distributed portfolio returns, give optimal portfolios on the mean-variance frontier.

The down-side risk measure Value at Risk (VaR) is just a quantile of the loss distribution, while Expected Shortfall (ES) is the average loss in case the loss is beyond the VaR. Target semivariance (TSV) is the average squared deviation around a target, but only counting the downside. Another chapter discusses the details and shows that, when returns are normally distributed, then *all three measures are increasing in the standard deviation*. Remark 8.14 summarises the key features, and details are in another chapter.

Remark 8.14 (VaR, ES and TSV with normally distributed returns) *If the return is normally distributed, $R \sim N(\mu, \sigma^2)$, then $VaR_\alpha = -(\mu + c\sigma)$, where c is the $1-\alpha$ quantile of a $N(0, 1)$ distribution (for instance, -1.64 for 5%). Also, $ES_\alpha = -[\mu - \phi(c)\sigma/(1 - \alpha)]$, where $\phi()$ is the pdf for a $N(0, 1)$ variable. Finally, it can be shown that the TSV $\lambda(h)$ is a*

strictly increasing function of the standard deviation, $d\lambda_p(h)/d\sigma = 2\sigma\Phi(a)$, where $\Phi()$ is the distribution function of a standard normal and $a = (h - \mu)/\sigma$.

With normally distributed returns, the VaR, ES and TSV are strictly increasing functions of the variance. In this case, the portfolio that *minimizes the VaR, ES or TS* at a given average return will be on the mean-variance frontier.

Another portfolio choice approach is to *use the value at risk (VaR) as a restriction*. For instance, the *Telser criterion* maximizes the expected portfolio return subject to the restriction that the value at risk does not exceed a given level V^*

$$\max_v E R_p \text{ st. } \text{VaR}_\alpha < V^*. \quad (8.16)$$

When returns are normally distributed, Remark 8.14 shows that the restriction can be rewritten as

$$E R_p > -V^* - c \text{ Std}(R_p), \quad (8.17)$$

where c is, for instance, -1.64 when the VaR_α has a confidence level $\alpha = 95\%$.

Example 8.15 With a VaR confidence level of 95% and $V^* = 0.1$, then (8.17) gives $E R_p > -0.1 + 1.64 \text{ Std}(R_p)$.

This optimization problem is illustrated in Figure 8.7. The objective is to find the portfolio with the highest expected return that satisfies the VaR restriction, which means that it has to be on or above the line defined by (8.17). Also, only points on or below the CLM (the mean-variance frontier based on both risky assets and a risk-free asset) are feasible.

The optimal portfolio is therefore where the restriction intersects the CLM: the Telser criterion applied to normally distributed returns gives a mean-variance portfolio. To be precise, the optimal portfolio puts

$$w = -\frac{R_f + V^*}{\mu_T^e + c\sigma_T}. \quad (8.18)$$

in the tangency portfolio (with average excess return μ_T^e and standard deviation σ_T) and the rest $(1 - w)$ in the risk-free asset.

We could instead use a restriction on expected shortfall or target semivariance, which define areas in a MV figure similar to that in Figure 8.7.

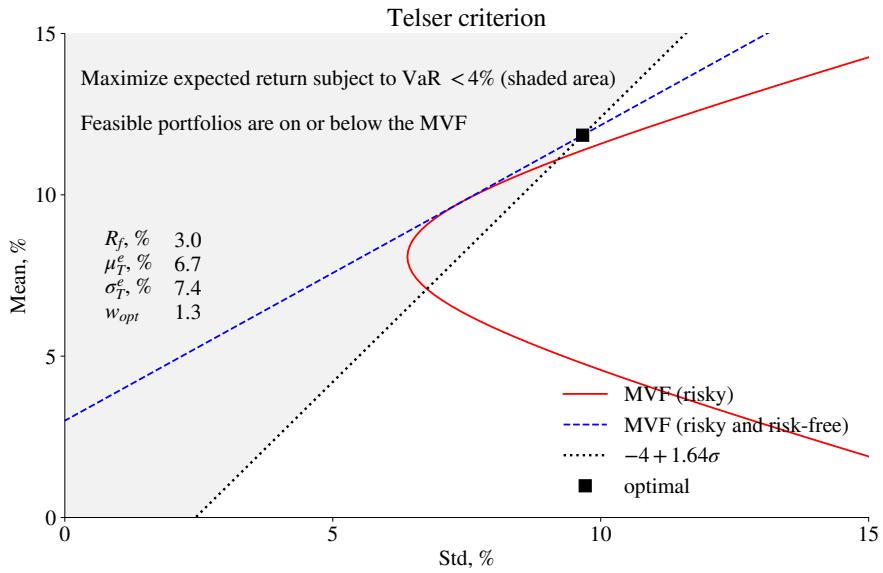


Figure 8.7: Telser criterion and VaR. The calculations use the inputs from Table 8.1.

Example 8.16 (Optimal portfolio. Telser) Let $\mu_T^e = 6.7\%$, $\sigma_T = 7.4\%$ and $R_f = 3\%$. The optimal portfolio with $V^* = 4\%$ is then

$$w = -\frac{0.03 + 0.04}{0.067 - 1.64 \times 0.074} \approx 1.3.$$

Instead, if the restriction is that $\text{VaR} < 2\%$, then the weight is $w \approx 0.9$. Figure 8.7.

Proof of (8.18). As usual, the average return on a portfolio p of the CML is

$$\mu_p = R_f + SR_T \sigma_p,$$

where SR_T is the Sharpe ratio of the tangency portfolio. This equals the mean return required by the VaR restriction (8.17) when

$$\sigma_p = -\frac{R_f + V^*}{SR_T + c}.$$

Since $\sigma_p = w\sigma_T$ (assuming $w \geq 0$), the optimal portfolio weight on the tangency portfolio is (8.18). \square

8.3 Behavioural Finance

Reference: Elton, Gruber, Brown, and Goetzmann (2014) 20; Forbes (2009); Shefrin (2005)

There is relatively little direct evidence on investor's preferences (utility). For obvious reasons, we can't know for sure what people really like. The evidence we do have is from two sources: "laboratory" experiments designed to elicit information about the test subject's preferences for risk, and a lot of indirect information.

8.3.1 Evidence on Utility Theory

The laboratory experiments are typically organized at university campuses (mostly by psychologists and economists) and involve only small compensations—so the test subjects are those students who really need the monetary compensation for taking part or those that are interested in this type of psychological experiments. The results vary quite a bit, but a main theme is that the key assumptions in utility-based portfolio choice might be reasonable. There are, however, some important systematic deviations from these assumptions.

For instance, investors seem to be unwilling to realize losses, that is, to sell off assets which they have made a loss on (often called the "disposition effect"). They also seem to treat the investment problem much more on an asset-by-asset basis than suggested by mean-variance analysis which pays a lot of attention to the covariance of assets (sometimes called mental accounting). Discounting appears to be non-linear in the sense that discounting is higher when comparing today with dates in the near future than when comparing two dates in the distant future. (Hyperbolic discount factors might be a way to model this, but lead to time-inconsistent behaviour: today we may prefer an asset that pays off in $t + 2$ to an asset that pays off in $t + 1$, but tomorrow our ranking might be reversed.) Finally, the results seem to move towards tougher play as the experiments are repeated and/or as more competition is introduced—although the experiments seldom converge to ultra tough/egoistic behaviour (as typically assumed by utility theory).

The indirect evidence is broadly in line with the implications of utility-based theory—especially now that the costs for holding well diversified portfolios have decreased (mutual funds). However, there are clearly some systematic deviations from the theoretical implications. For instance, many investors seem to be too little diversified. In particular, many investors hold assets in companies/countries that are very strongly correlated to their labour income (local bias). Moreover, diversification is often done in a naive fashion and depends on the "menu" of choices. For instance, many pension savers seem to diversify by putting the fraction $1/n$ in each of the n funds offered by the firm/bank—irrespective of what kind of funds they are. There are, of course, also large chunks of wealth invested for

control reasons rather than for a pure portfolio investment reason (which explains part of the so called “home bias”—the fact that many investors do not diversify internationally).

8.3.2 Evidence on Expectations Formation (Forecasting)

In laboratory experiments (and studies of the properties of forecasts made by analysts), several interesting results emerge on how investors seem to form expectations. First, complex situations are often approached by treating them as a simplified representative problem—even against better knowledge (often called “representativeness”)—and stands in contrast to the idea of Bayesian learning where investors update and learn from their mistakes. Second (and fairly similar), difficult problems are often handled as if they were similar to some old/easy problem—and all that is required is a small modification of the logic (called “anchoring”). Third, recent events/data are given much higher weight than they typically warrant (often called “recency bias” or “availability”). Finally, most forecasters seem to be overconfident: they draw (too) strong conclusions from small data sets (“law of small numbers”) and overstate the precision of their own forecasts.

Notice, however, that it is typically difficult to disentangle (distorted) beliefs from non-traditional preferences. For instance, the aversion of selling off bad investments, may equally well be driven by a belief that past losers will recover.

8.3.3 Prospect Theory

The *prospect theory* (developed by Kahneman and Tversky) tries to explain several of these things by postulating that the utility function is concave over some reference point (which may shift), but convex below it. This means that gains are treated in a risk-averse way, but losses in a risk-loving way. For instance, after a loss (so we are below the reference point) an asset looks less risky than after a gain—which might explain why investors hold on to losing investments. Clearly, an alternative explanation is that investors believe in mean-reversion (losing positions will recover, winning positions will fall back). In general, it is hard to make a clear distinction between non-classical preferences and (potentially distorted) beliefs.

8.4 Appendix – Risk Aversion and the Level of Wealth*

This section discusses how risk aversion is related to the wealth level. (Notice that when we use the portfolio return as the argument of the utility function, then this amounts to

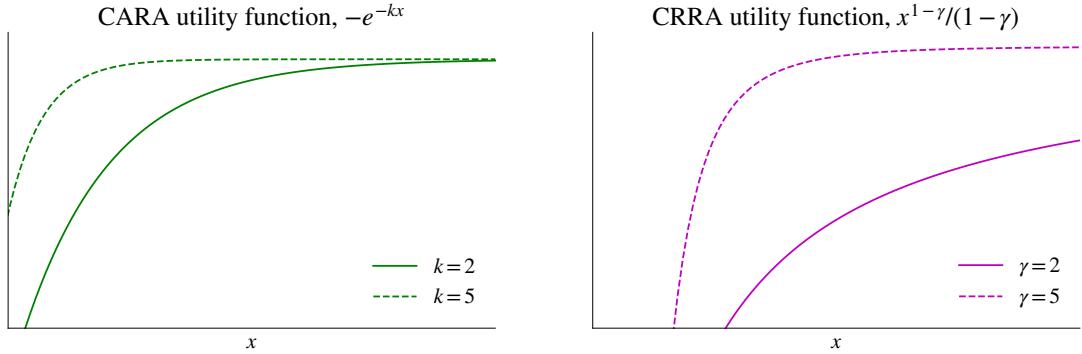


Figure 8.8: Examples of utility functions

disregarding differences across wealth levels.)

First, define *absolute risk aversion* as

$$A(W) = \frac{-U''(W)}{U'(W)}, \quad (8.19)$$

where $U'(W)$ is the first derivative and $U''(W)$ the second derivative. Second, define *relative risk aversion* as

$$R(W) = WA(W) = \frac{-WU''(W)}{U'(W)}. \quad (8.20)$$

These two definitions are strongly related to the attitude towards taking risk (see below).

Figure 8.8 demonstrates two commonly used utility functions, and the following discussion outlines their main properties.

The *CARA utility function* (constant absolute risk aversion), $U(W) = -e^{-kW}$, is quite simple to use (in particular when returns are normally distributed), but has the unappealing feature that the amount invested in the risky asset (in a risky/risk-free trade-off) is constant across wealth levels. This means, of course, that wealthy investors would have a lower portfolio weight on risky assets.

Remark 8.17 (*Risk aversion in CARA utility function*) $U(W) = -e^{-kW}$ gives $U'(W) = ke^{-kW}$ and $U''(W) = -k^2e^{-kW}$, so we have $A(W) = k$. This means an increasing relative risk aversion, $R(W) = Wk$, so a poor investor typically has a larger portfolio weight on the risky asset than a rich investor.

The *CRRA utility function* (constant relative risk aversion) is often harder to work with, but has the nice property that the portfolio weights are unaffected by the wealth level. This

fits with historical data which show no trends in portfolio weights or risk premia—in spite of investors having become much richer over time.

Remark 8.18 (*Risk aversion in CRRA utility function*) $U(W) = W^{1-\gamma}/(1-\gamma)$ gives $U'(W) = W^{-\gamma}$ and $U''(W) = -\gamma W^{-\gamma-1}$, so we have $A(W) = \gamma/W$ and $R(W) = \gamma$. The absolute risk aversion decreases with the wealth level in such a way that the relative risk aversion is constant. In this case, a poor investor typically has the same portfolio weight on the risky asset as a rich investor.

To understand the concepts of absolute and relative risk aversion, consider an investor with wealth W who can choose between taking on a zero mean risk Z (so $E Z = 0$) or pay a price P . The investor is indifferent if

$$E U(W + Z) = U(W - P). \quad (8.21)$$

If Z is a small risk, then we can use a second order approximation and solve for the price as

$$P \approx A(W) \text{Var}(Z)/2. \quad (8.22)$$

This says that the price the investor is willing to pay to avoid the risk Z is proportional to the *absolute risk aversion* $A(W)$.

Example 8.19 (*Willingness to pay to avoid a risk*) Suppose the investor has a CARA utility function with $A(W) = 5$ and that $\text{Var}(Z) = 1$. Then, $P = 5 \times 1/2 = 2.5$.

Proof of (8.22). First, approximate as

$$\begin{aligned} E U(W + Z) &\approx U(W) + U'(W) E Z + U''(W) E Z^2/2 \\ &= U(W) + U''(W) \text{Var}(Z)/2, \end{aligned}$$

since $E Z = 0$. Second, approximate $U(W - P) \approx U(W) - U'(W)P$. Finally, make the two approximations equal to get (8.22). \square

If we change the setting in (8.21)–(8.22) to make the risk proportional to wealth, that is $Z = Wz$ where z is the risk factor, then (8.22) directly gives

$$\begin{aligned} P &\approx A(W)W^2 \text{Var}(z)/2, \text{ so} \\ P/W &\approx R(W) \text{Var}(z)/2. \end{aligned} \quad (8.23)$$

This says that the fraction of wealth (P/W) that the investor is willing to pay to avoid the risk (z) is proportional to the *relative risk aversion* $R(W)$.

Example 8.20 (*Willingness to pay to avoid a risk*) Suppose the investor has a CRRA utility function with $R(W) = 5$ and that $\text{Var}(z) = 0.2$. Then, $P/W = 5 \times 0.2/2 = 0.5$.

These results mostly carry over to the portfolio choice: high absolute risk aversion typically implies that only small *amounts* are invested in risky assets, whereas a high relative risk aversion typically leads to small *portfolio weights* of risky assets.

8.5 Appendix – Portfolio Choice with $N()$ Returns*

8.5.1 Case 3 and 4: Proofs

Proof of Proposition 8.12. First, recall that if $x \sim N(\mu, \sigma^2)$, then $E e^x = e^{\mu + \sigma^2/2}$. Therefore, rewrite expected utility as

$$E U(R_p) = E[-\exp(-R_p k)] = -\exp[-E R_p k + \text{Var}(R_p)k^2/2].$$

Notice that the assumption of normally distributed returns is crucial for this result. Second, recall that if x maximizes $f(x)$, then it also maximizes $g[f(x)]$ if g is a strictly increasing function. The function $-\ln(-z)/k$ is defined for $z < 0$ and it is increasing in z . We can apply this function by letting z be the right hand side of the previous equation to get

$$-\ln(-z)/k = E R_p - \text{Var}(R_p)k/2.$$

Therefore, maximizing the expected CARA utility or MV preferences (in terms of the returns) gives the same solution. \square

Proof of Proposition 8.13. Notice that

$$E(1 + R_p)^{1-\gamma}/(1-\gamma) = E \exp[(1-\gamma)r_p]/(1-\gamma), \text{ where } r_p = \ln(1 + R_p).$$

Since r_p is normally distributed, the expectation is (recall that if $x \sim N(\mu, \sigma^2)$, $E \exp(x) = \exp(\mu + \sigma^2/2)$)

$$E \exp[(1-\gamma)r_p]/(1-\gamma) = \exp[(1-\gamma)E r_p + (1-\gamma)^2 \text{Var}(r_p)/2]/(1-\gamma).$$

If $\gamma > 1$ ($0 < \gamma < 1$), then the function $\ln[z(1-\gamma)]/(1-\gamma)$ is defined for $z < 0$ ($z > 0$) and it is increasing in z . Let z be the right hand side of the previous equation (which is negative if $\gamma > 1$ and positive if $0 < \gamma < 1$) and apply the transformation to get

$$E r_p + (1-\gamma) \text{Var}(r_p)/2.$$

To express this in terms of the mean and variance of the return instead of the log return we use the following facts: if $\ln y \sim N(\mu, \sigma^2)$, then $E y = \exp(\mu + \sigma^2/2)$ and $\text{Std}(y)/E y = (\exp(\sigma^2) - 1)^{1/2}$. Using this fact in the previous expression and rearranging (express $E r_p$ and $\text{Var}(r_p)$ in terms of $E R_p$ and $\text{Var}(R_p)$) gives

$$\ln(1 + E R_p) - \gamma \ln[\text{Var}(R_p)/(1 + E R_p)^2 + 1]/2,$$

which is increasing in $E R_p$ and decreasing in $\text{Var}(R_p)$. We therefore get a mean-variance portfolio. \square

-

Chapter 9

Multi-Factor Models

This chapter is an *introduction* to multi-factor models. It provides theoretical motivations of some simple multi-factor models and performs an empirical test of a commonly used model (Fama and French (1993)). Detailed treatments of this important topic is found elsewhere (for instance, Cochrane (2005) and Back (2010)).

9.1 Factor Investment

A number of factor returns related to, for instance, firm characteristics like size and profitability, have shown good performance over long periods of time. It is therefore common to base investment strategies on those characteristics—and a large number of funds and other investment vehicles have been developed for this purpose. This approach is called *factor investing* or “smart beta.” It is essentially a dynamic trading strategy since firm characteristics change over time.

In studies of investment fund performance, it is often found that the abnormal performance (α from a CAPM regression) can be explained by a fairly small set of factors. It seems that many fund managers have indeed been able to invest in those characteristics that have historically paid off. This suggests that the market might have moved beyond CAPM and that a multi-factor model would be empirically more appropriate.

Empirical Example 9.1 (Fama-French factors) Figure 9.1 illustrate several of the factors discussed by Fama and French (1993) and Fama and French (2015), while Table 9.1 summarises the return patterns. Many factors have positive excess returns, which is especially interesting since they are long-short in equities. It is also clear that several of the factors are virtually uncorrelated with the market excess return, so the α values are considerable (and similar to the average excess returns).

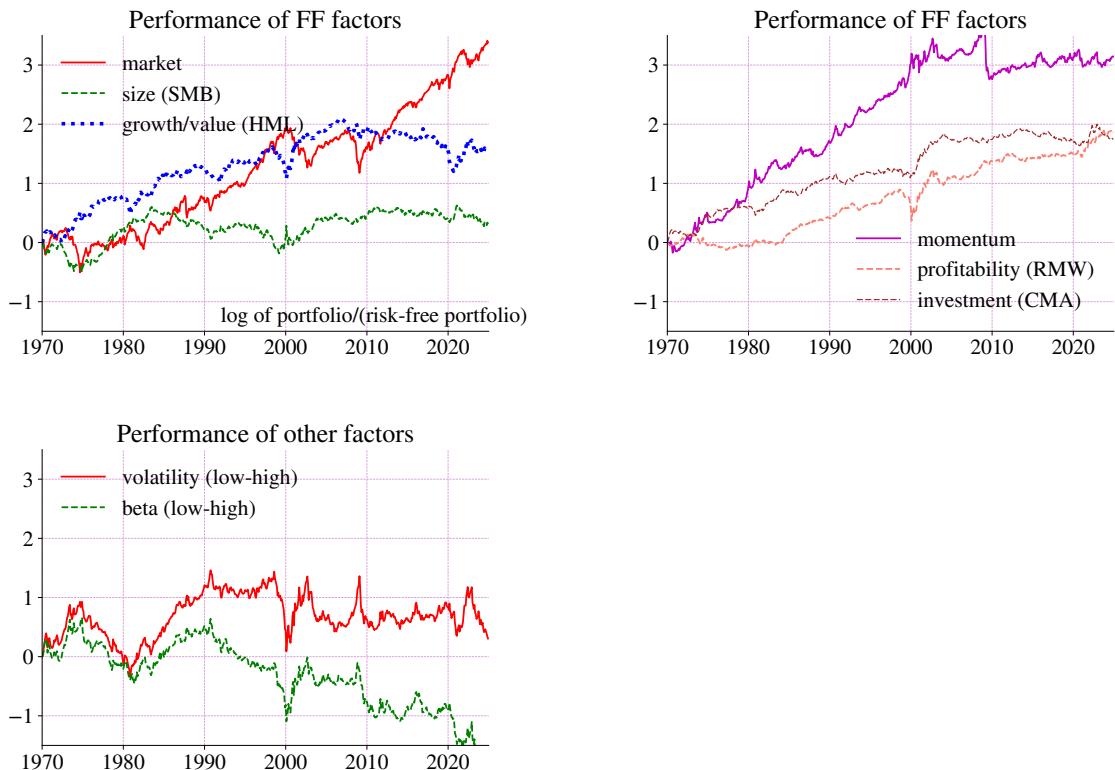


Figure 9.1: Cumulated returns of important equity factors

9.1.1 Portfolio Sorts

Many of the factors in Table 9.1 are based on portfolio sorts, which can be thought of as dynamic trading strategies. This approach is often used to study how asset characteristics relate to returns, sometimes as an alternative to regressions.

In a *univariate sort*, we could rank (sort) the assets according to z_{it} and then construct portfolios (say, three equally weighted portfolios: low, mid and high). Then, we measure and study the returns of the portfolios. Typically, the sorting is repeated at regular time intervals (every day or perhaps every June), so z_{it} should be understood as the value used in a particular period.

Empirical Example 9.2 (*Sorting on recent returns*) See Table 9.2 for an empirical example where the 25 FF portfolios are sorted into low/low recent 22-day returns, with 5 portfolios in each. The results indicate strong momentum.

Bivariate (double) sorts are used when there are two important characteristics (here called x and z) and you want to study how z affects returns, holding x “constant”. This

	μ , %	σ , %	SR	β	α , %
market	7.30	15.92	0.46	1.00	-0.00
size	1.18	10.66	0.11	0.19	-0.20
growth/value	3.44	10.73	0.32	-0.14	4.49
momentum	6.94	14.99	0.46	-0.18	8.24
profitability	3.68	7.89	0.47	-0.10	4.41
investment	3.51	7.18	0.49	-0.17	4.72
volatility (low-high)	3.26	22.70	0.14	-0.88	9.69
beta (low-high)	-2.26	20.41	-0.11	-0.83	3.83

Table 9.1: Descriptive statistics of excess returns of different US equity portfolios (including the Fama-French factors and more), annualised. Monthly data 1970:01-2024:12.

	Portfolio returns
Low 22-day return	4.15 (0.20) [-5.36]
High 22-day return	13.71 (0.74) [5.16]
Difference (H-L)	9.55 (0.85) [10.52]

Table 9.2: Average excess returns, (Sharpe ratios) and $[\alpha]$ for 3 portfolios from a univariate sort on recent (22-day) returns (5/5 assets). Annualized figures. Daily data on 25 FF portfolios 1979:01-2024:12

may be important if x and z are correlated. An *independent bivariate sort* first does a univariate sort based on x_{it} (say, forming 3 categories: growth, neutral or value), then it makes another univariate sort according to the other sorting variable z_{it} (say, forming two categories: small or big). Then, we find the intersections of the two sorts. In contrast, in a *dependent bivariate sort* we first sort according to x_{it} as before. Then, *within* an x_t category we sort according to z_{it} . This allows us to control the number of assets in each group.

9.2 An Overview of Multi-Factor Models

This section gives a short introduction to multi-factor models. Model details and proofs are in later sections.

A multi-factor model extends the market model by allowing more factors to explain the return on an asset. For instance, a two-factor model is

$$R_{it}^e = \alpha + \beta_{im} R_m^e + \beta_{ic} R_c^e + \varepsilon_{it}, \quad (9.1)$$

where R_m^e is the excess return on the market and R_c^e is the excess return on some other portfolio. As usual, we require $E \varepsilon_{it} = 0$, and that ε_{it} is uncorrelated to all regressors.

The pricing implication is a multi-beta model

$$E R_i^e = \beta_{im} \mu_m^e + \beta_{ic} \mu_c^e. \quad (9.2)$$

Notice that there is no intercept, so α in (9.1) should be zero.

Remark 9.3 (*When factors are not excess returns**) Equation (9.2) assumes that the factor can be expressed as an excess return, but that is not always the case. For instance, it could be that the second factor is a macro variable like inflation surprises. Then there are two possible ways to proceed. First, find that portfolio which mimics the movements in the inflation surprises best and use the excess return of that (factor mimicking) portfolio in (9.1) and (9.2). Second, we could instead (a) estimate the betas (β_{im}, β_{ic}) by a time series regression of (9.1), but allowing for an intercept; and (b) estimate the factor risk premia (μ_m^e, μ_c^e) by a cross-section of (9.2) where the dependent variable is the historical average returns of different assets ($i = 1, 2, \dots, n$) and the regressors are the betas from the first step.

This chapter will discuss *theoretical* multi-factor models: (a) CAPM with background risk as well as (b) a consumption-based model.

There are also many *empirically motivated* multi-factor models that have been found to work well in practice. For instance, Fama and French (1993) estimate a three-factor model (capturing the market, the difference between small and large firms and the difference between value firms and growth firms) and show that it empirically performs much better than CAPM. The more recent Fama and French (2015) extends this to a five-factor model. Also, the multi-factor model by MSCI Barra (MSCI Inc. (2024)) is widely used in the financial industry. It uses a set of firm characteristics as factors, for instance, size, volatility,

price momentum, and industry/country (see Stefek (2002)). These models suggest that CAPM may not be enough.

9.3 Portfolio Choice with Background Risk

This section discusses the portfolio problem when there is “background risk” or non-traded assets (see Mayers (1972)), for instance, labour income, real estate, a private business, or a liability.

The existence of background risk/non-traded assets will affect portfolio choice, and therefore perhaps also asset prices.

9.3.1 Portfolio Choice with Background Risk: One Risky Asset

Consider a mean-variance investor who forms a financial portfolio by choosing between a risky asset (henceforth called “equity”) with return R and a risk-free asset with return R_f . The investor also has a background risk in the form of an endowment (positive or negative) of a non-traded asset with return R_c . This could, for instance, be labour income or a house (positive endowment) or a liability (negative endowment).

The non-traded asset accounts for the fraction ϕ of total wealth, while the financial portfolio accounts for $1 - \phi$. When the non-traded asset is a liability, then $\phi < 0$. The return on the total portfolio, R_p , is

$$R_p = (1 - \phi)R_{Fin} + \phi R_c, \text{ with} \quad (9.3)$$

$$R_{Fin} = wR^e + R_f, \quad (9.4)$$

where R_c is the return (change of value, payoff) of the non-traded asset.

The investor chooses w to maximize

$$E U(R_p) = E R_p - \frac{k}{2} \text{Var}(R_p), \quad (9.5)$$

and the optimal value is

$$w = \frac{\mu^e/k - \phi S_c}{(1 - \phi)\sigma^2}, \quad (9.6)$$

where σ^2 is the variance of equity and S_c is the covariance of equity and the non-traded asset.

Proof of (9.6). To simplify the notation, write the portfolio return as $R_p = vR^e +$

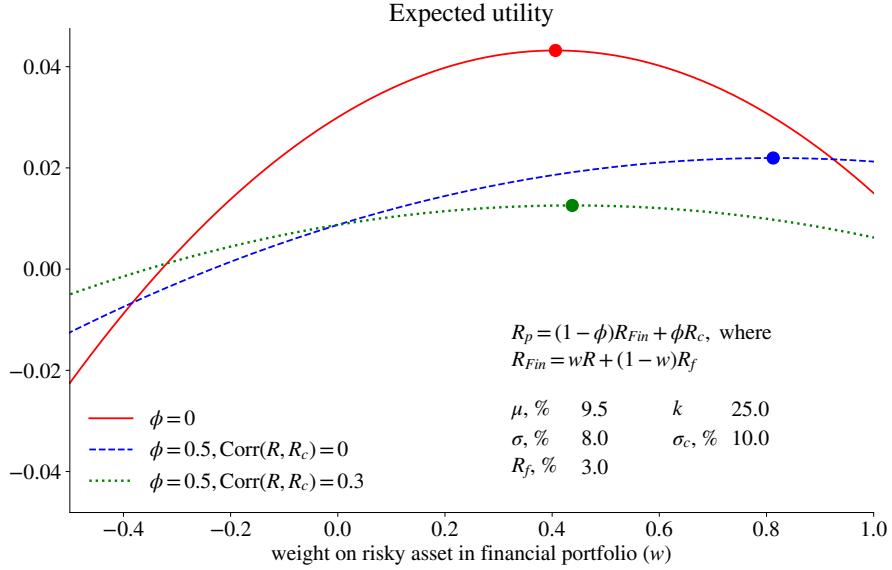


Figure 9.2: Portfolio choice with non-traded assets

$\phi R_c^e + R_f$, where $v = (1 - \phi)w$. Use in the objective function to get

$$E U(R_p) = v\mu^e + \phi\mu_c^e + R_f - \frac{k}{2}(v^2\sigma^2 + \phi^2\sigma_c^2 + 2v\phi S_c),$$

The first order condition wrt. v is $\mu^e - k(v\sigma^2 + \phi S_c) = 0$, solve for v and divide by $1 - \phi$ to get (9.6). \square

The second term in the optimal portfolio weight, *the hedging term*, depends on how important the non-traded asset is in the portfolio (ϕ) and also on the covariance term (S_c). Clearly, if there is no non-traded asset in the total portfolio ($\phi = 0$), then we are back in a traditional MV case.

Remark 9.4 (*Interpreting the hedging term**) The hedging term in (9.6) is related to the slope coefficient in a regression of the non-traded asset on equity, $R_{ct}^e = \alpha + \gamma R_{it}^e + \varepsilon_t$, since $\gamma = \sigma_{ic}/\sigma_i^2$, which corresponds to S_c/σ^2 in the notation above.

Several things can be noticed. First, when the correlation is zero ($\text{Corr}(R, R_c) = 0$), then the equity weight is increasing in the amount of non-traded assets (ϕ), while the opposite holds for the risk-free asset; see Figure 9.2 for an illustration. The intuition is that a zero correlation means that the non-traded asset is quite similar to a bond: having an endowment of a bond-like asset in the overall portfolio means that the financial portfolio should be tilted towards equity.

Second, *when the correlation is positive* ($\text{Corr}(R, R_c) > 0$) and we have a positive exposure to the non-traded asset ($\phi > 0$), then the hedging term will reduce the equity weight. Again, see Figure 9.2. The intuition is that the overall portfolio now includes a lot of “equity like” assets, so the financial portfolio should be tilted towards the risk-free asset. The opposite holds when the exposure to the non-traded asset is negative (a liability, $\phi < 0$) or when the non-traded asset is negatively correlated with equity.

Example 9.5 (*Portfolio choice of young and old*) Consider the common portfolio advice that young investors (with labour income) should invest relatively more in stocks than old investors. In this case, the non-traded asset is an endowment of “human capital,” that is, the present value of future labour income—and current labour income can loosely be interpreted as its return. The analysis in this section suggests that a low correlation of stock returns and wages means that the young investor is endowed with a bond-like asset, so the financial portfolio should be tilted towards equity. Old investors less so.

9.3.2 Portfolio Choice with Background Risk: Several Risky Assets

With several risky assets, the financial portfolio return is

$$R_{Fin} = w' R^e + R_f, \quad (9.7)$$

where w now is a vector of portfolio weights, R a vector of returns on the risky assets and $\mathbf{1}$ is a vector of ones. The optimal portfolio is now

$$w = \Sigma^{-1} \frac{\mu^e/k - \phi S_c}{1 - \phi}, \quad (9.8)$$

where Σ is the variance-covariance matrix of the risky assets (not including the non-traded asset) and S_c is a vector of covariances of the assets with the non-traded asset. The portfolio weights of the financial subportfolio will (as long as $\phi S_c \neq 0$) give a return that is *off the mean-variance frontier*—and will differ across investors if the non-traded asset does: the *two-fund separation theorem is no longer valid*. See Figure 9.3 for an illustration. Note that the optimal portfolio tends to have lower weights on assets that are positively correlated with the non-traded asset and vice versa (compare with Table 9.3).

Proof of (9.8). The portfolio return (9.7) can be written $R_p = v' R^e + \phi R_c^e + R_f$, where $v = (1 - \phi)w$. The investor solves

$$\max_v v' \mu^e + \phi \mu_c^e + R_f - \frac{k}{2} (v' \Sigma v + \phi^2 \sigma_c^2 + 2\phi v' S_c),$$

	$\underline{\mu, \%}$		$\underline{\Sigma, \text{bp}}$			$\underline{\rho_{x,c}}$
	A	B	C	c		
A	11.5	166	34	58	161	0.50
B	9.5	34	64	4	180	0.90
C	6.0	58	4	100	-25	-0.10
c	10.0	161	180	-25	625	1.00

Table 9.3: Characteristics of the assets in the example of MV with background risk. Notice that $\mu, \%$ is the expected return in % (that is, $\times 100$) and Σ, bp is the covariance matrix in basis points (that is, $\times 100^2$). $\rho_{x,c}$ are the correlations of each asset with the background risk.

with first order conditions

$$\mu^e - k(\Sigma v + \phi S_c) = 0.$$

Solve for v and divide by $1 - \phi$ to get (9.8). \square

Remark 9.6 (*Interpreting the hedging terms**) The hedging terms are related to the slope coefficients from a regression of R_c^e on the vector of investable risky assets (R^e) $R_{ct}^e = \alpha + \gamma' R_t^e + \varepsilon_t$, since $\gamma = \Sigma^{-1} S_c$.

Example 9.7 (*Portfolio choice of a pharmaceutical engineer*) Suppose asset 1 is an index of pharmaceutical stocks, and asset 2 is the rest of the equity market. Consider a person working as a pharmaceutical engineer: the covariance of her labour with asset 1 is likely to be high, while the covariance with asset 2 might be more modest. This person should therefore tilt the financial portfolio away from pharmaceutical stocks: the market portfolio is not the best for everyone.

Remark 9.8 (*Transformed assets**) However, the optimal portfolio w is on the mean-variance frontier of some transformed assets with returns Z_i . We can rewrite the portfolio return as

$$R_p = w' Z + (1 - 1' w) Z_f, \text{ where}$$

$$Z_i = (1 - \phi) R_i + \phi R_c \text{ and } Z_f = (1 - \phi) R_f + \phi R_c.$$

Notice that all these transformed assets (also Z_f) are risky. The optimal portfolio will be on the mean-variance frontier of $Z = (Z_i, Z_f)$. See Figure 9.4. (The “proof” is that maximizing the objective function (9.5) subject to this new definition of the portfolio return is a traditional mean-variance problem—but in terms of the transformed assets, Z .)

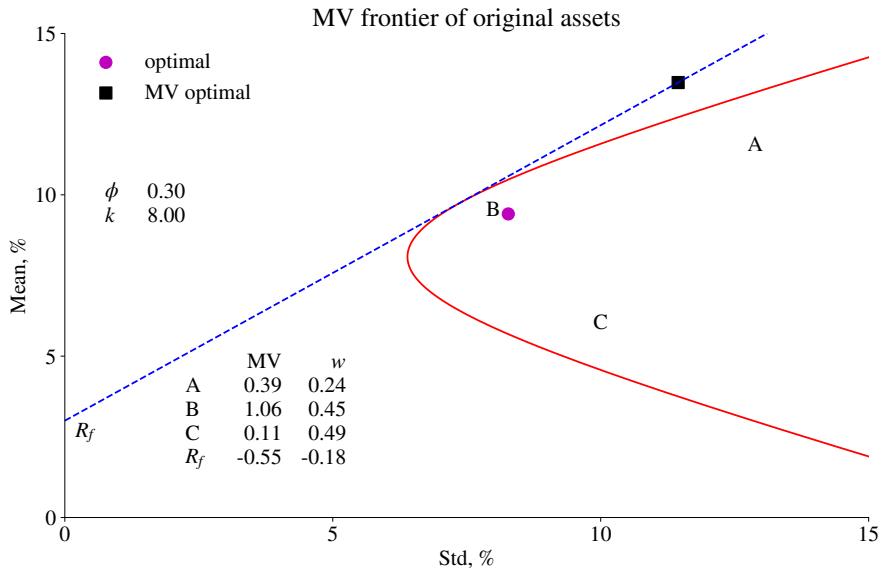


Figure 9.3: Portfolio choice with background risk. The properties of the assets are shown in Table 9.3.

9.4 Asset Pricing Implications

Background risk that varies greatly across investors is unlikely to affect asset prices. Hedging it is then rather similar to hedging idiosyncratic (asset specific) risk.

In contrast, if the background risk affects the portfolio choice of a *large fraction* of the investors, then it is also likely to influence asset prices. For instance, an asset that provides an effective hedge against a common background risk will be greatly demanded—and therefore generate low returns.

To create tractable pricing expressions, we use a *factor mimicking portfolio* λ (or factor replicating portfolio, see Cochrane (2005) 6 and Back (2010) 6) in place of the true factor (here, the background risk). This is the portfolio with the highest squared correlation (R^2) with the true factor. In fact, it is the fitted value (minus the intercept) from a linear regression of the background risk on all excess returns. In practice, an approximate factor mimicking portfolio might be used to facilitate the implementation.

Notice that a factor mimicking portfolio is not a new asset, so it does not change any prices. Rather, it is just a convenient way of characterizing the pricing process.

The market portfolio m is here chosen to be the optimal portfolio for an investor with a risk aversion (k) such that the portfolio weight on the risk-free asset is zero. This, together with the factor mimicking portfolio will define a two-factor model.

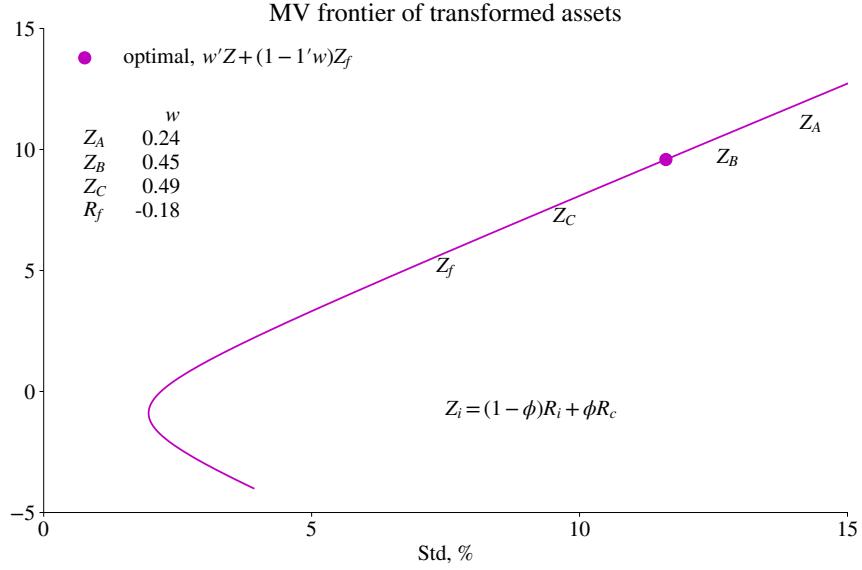


Figure 9.4: Portfolio choice with background risk, transformed assets. The properties of the assets are shown in Table 9.3.

Remark 9.9 (*Factor mimicking portfolio*) Regress $R_c^e = \delta + R^{e'}w_\lambda + \varepsilon$ and notice that $w_\lambda = \Sigma^{-1}S_c$, where Σ is the variance-covariance matrix of the assets and S_c the vector of covariances of the assets with the background risk. The factor mimicking portfolio has the excess return $R_\lambda^e = w_\lambda' R^e$. It is clear that, for any portfolio w_p , $\sigma_{pc} = \sigma_{p\lambda}$, since ε is uncorrelated with R^e .

Example 9.10 (*Numerical example I*) Table 9.4 shows results based on the asset properties in Table 9.3: the portfolio weights on the risky assets for the factor mimicking portfolio (λ), the market portfolio (m), and a randomly picked portfolio (p) that will be used to test the pricing ability of the two-factor model below.

The theoretical implication of the background risk is a *multi-beta model*

$$E R_p^e = \beta_{pm}\mu_m^e + \beta_{p\lambda}\mu_\lambda^e, \quad (9.9)$$

where μ_m^e and μ_λ^e are the average excess returns on the two factors, and the betas are from the linear regression

$$R_p^e = \alpha + \beta_{pm}R_m^e + \beta_{p\lambda}R_\lambda^e + \varepsilon, \quad (9.10)$$

where we regress the excess return of a portfolio p on the excess returns on the two factors.

	w_m	w_λ	w_p
A	0.19	0.74	0.50
B	0.33	2.47	0.40
C	0.48	-0.78	0.10

Table 9.4: Portfolio weights on the market portfolio (w_m), the factor mimicking portfolio (w_λ), and a randomly picked portfolio w_p . The remainder is invested in the riks-free asset. The market portfolio is the optimal portfolio when $k = 8.691$. Based on the asset properties in Table 9.3.

Clearly, the expected excess return on portfolio p in (9.9) depends on how it is related to both the (financial) market and the background risk. Notice that there is no intercept in (9.9), so the testable implication is that $\alpha = 0$ in (9.10). (The formal proof is in the Appendix.)

Example 9.11 (*Numerical example II*) Table 9.5 shows the implied expected excess returns of (m, λ) , their variance-covariance matrix, and also their covariances with portfolio p . The betas, shown in the last column, are for the regression. According to these results, the expected excess return of portfolio p should be

$$\mathbb{E} R_p^e = 0.8 \times 5.22\% + 0.15 \times 19.98\% \approx 7.17\%.$$

The actual value is 7.15% (see the caption of the table). The difference is due to rounding.

	$\mu^e, \%$	$\text{Cov}([m, \lambda]), \text{bp}$	$\text{Cov}(x, p), \text{bp}$	β
m	5.22	51.98	78.81	53.23
λ	19.98	78.81	582.44	150.03

Table 9.5: Properties of the market (m), factor mimicking (λ), and the randomly selected portfolio (p). Expected returns are in percent and variances and covariances are in basis points. The β are from the regression $R_p^e = \alpha + \beta' [R_m^e, R_\lambda^e] + \varepsilon$. Also, $\mu_p^e = 7.15\%$. Based on the asset properties in Table 9.3 and results in Table 9.4.

Remark 9.12 (*Calculating the properties of $(m, \lambda, p)^*$*) With the portfolio weights, we can calculate the variance-covariance matrix of (m, λ) and also the vector of covariance of p with (m, λ) as

$$\begin{bmatrix} \sigma_{mm} & \sigma_{m\lambda} \\ \sigma_{m\lambda} & \sigma_{\lambda\lambda} \end{bmatrix} = \begin{bmatrix} w'_m \Sigma w_m & w'_m \Sigma w_\lambda \\ w'_m \Sigma w_\lambda & w'_\lambda \Sigma w_\lambda \end{bmatrix} \text{ and } \begin{bmatrix} \sigma_{pm} \\ \sigma_{p\lambda} \end{bmatrix} = \begin{bmatrix} w'_p \Sigma w_m \\ w'_p \Sigma w_\lambda \end{bmatrix}$$

Excess returns are calculated as $\mu_x^e = w'_x \mu^e$, where μ^e is the vector of excess returns for the investable assets. The slope coefficients in (9.10) are

$$\begin{bmatrix} \beta_{pm} \\ \beta_{p\lambda} \end{bmatrix} = \begin{bmatrix} \sigma_{mm} & \sigma_{m\lambda} \\ \sigma_{m\lambda} & \sigma_{\lambda\lambda} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{pm} \\ \sigma_{p\lambda} \end{bmatrix}.$$

9.4.1 Asset Pricing Implications II: Reinterpreting alpha

Consider the standard CAPM regression

$$R_{it}^e = \alpha_i + \beta_i R_{mt}^e + \varepsilon_{it}, \quad (9.11)$$

where R_{mt}^e is the market excess return in period t . We use time series data to estimate it. The intercept α is likely to be non-zero if the R_{it} returns are driven by a multi-factor model. That is, (9.11) suffers from an omitted variable bias.

To be precise, suppose the two-factor model (9.10) holds with a zero intercept. Then, the OLS estimate of α_i from (9.11) is

$$\hat{\alpha}_i = \hat{\theta}_0 \hat{\beta}_{ic}, \quad (9.12)$$

where $\hat{\beta}_{ic}$ is the beta in the two-factor model regression (9.10) and $\hat{\theta}_0$ is the estimate of the intercept in

$$R_{ct}^e = \theta_0 + \theta_1 R_{mt}^e + \eta_t. \quad (9.13)$$

(To show this, apply Remark 9.13.) Together, these two equations suggest that non-zero alphas from CAPM regression may be explained by a combination of (1) a missing factor ($\beta_{ic} \neq 0$); (2) and that factor is not “priced” by the market returns alone ($\theta_0 \neq 0$).

Remark 9.13 (Omitted variable bias in OLS) Suppose the correct regression model is $y_t = x'_t \beta + h_t \gamma + u_t$, but we omit the h_t regressor and estimate $y_t = x'_t \delta + \varepsilon_t$ by OLS. It is well known that the OLS estimate is $\hat{\delta} = \hat{\beta} + \hat{\theta} \hat{\gamma}$, where $\hat{\theta}$ is from regressing $h_t = x'_t \theta + \eta_t$.

9.5 Joint Portfolio and Savings Choice

The basic *consumption-based* multi-period investment problem assumes that the investor derives utility from consumption in every period and that the utility in one period is additively separable from the utility in other periods. For instance, if the investor plans

for 2 periods (labelled 1 and 2), then the task is to maximize expected utility

$$\max U(c_1) + \delta E_1 U(c_2), \text{ subject to} \quad (9.14)$$

$$c_1 + I_1 = W_1 \quad (9.15)$$

$$c_2 + I_2 = (1 + R_p)I_1 + y_2, \text{ where} \quad (9.16)$$

$$R_p = v' R^e + R_f. \quad (9.17)$$

In equation (9.14), c_t is consumption in period t . The current period (when the portfolio is chosen) is period 1—so all expectations are made on the basis of the information available then. The constant δ is the time discounting, with $0 < \delta < 1$ indicating impatience. (In an equilibrium without risk, we will get a positive real interest rate if investors are impatient.)

Equation (9.15) is the budget constraint for period 1: an initial wealth (including exogenous income), W_1 , is split between consumption, c_1 , and investment, I_1 . Equation (9.16) is the budget constraint for period 2: consumption plus investment must equal the wealth at the beginning of period 2 plus (exogenous) income, y_2 . The wealth at the beginning of period 2 equals the investment in period 1, I_1 , times the gross portfolio return—which in turn (see (9.17)) depends on the portfolio weights v chosen in period 1 as well as on the returns on the assets.

Obtaining closed-form solutions is typically difficult. However, we can gain some insights by studying the first order conditions. The optimization problem involves maximizing with respect to the investment level (mostly a macro topic, but summarized in a Remark below) and how to form the investment portfolio, which is the focus here.

9.5.1 Optimal Portfolio Choice

This section studies the *portfolio choice*, that is, the portfolio weights in the vector v .

$$E_1[U'(c_2)R_i^e] = 0 \text{ for } i = 1, \dots, n. \quad (9.18)$$

The expression says that excess returns should be “orthogonal” to marginal utility. This is similar to earlier results on utility based portfolio choice, with the difference that marginal utility now depends on consumption rather than the portfolio return.

Proof (of (9.18)) The first order condition for v_i is $\delta E_1[U'(c_2)\partial c_2 / \partial v_i] = 0$. Differentiates the budget constraint (9.16) to get $\partial c_2 / \partial v_i = I_1 R_i^e$ and simplify the resulting expression by dividing both sides by δI_1 (which is known in t and therefore can be moved outside the expectation). \square

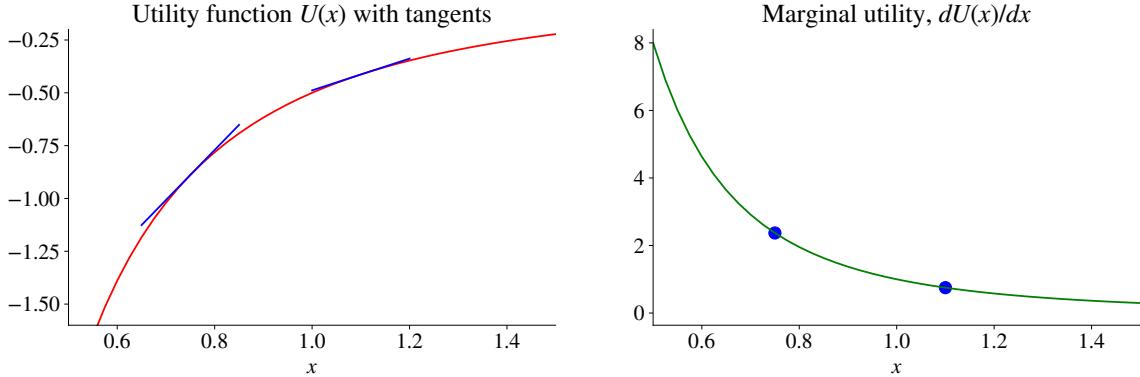


Figure 9.5: Utility function

The first order conditions (9.18) still contain some useful information, especially if we rewrite them as

$$E_1 R_i^e = -\text{Cov}_1[U'(c_2), R_i^e]/E_1 U'(c_2), \quad (9.19)$$

where the time subscripts on the expectation and covariance operators indicate that they are conditional on the information in period 1.

Proof (of (9.19)) Recall that, by definition, $\text{Cov}(x, y) = E xy - E x \times E y$. $E xy = 0$, so $E y = -\text{Cov}(x, y)/E x$. Set $y = R_i^e$, $x = U'(c_2)$ and notice that (9.18) says that $E xy = 0$. \square

First, the denominator is positive (marginal utility always is). Second, suppose the return is procyclical, $\text{Cov}(c_2, R_i^e) > 0$. This will make $\text{Cov}[U'(c_2), R_i^e] < 0$, since marginal utility $U'(c_2)$ is a decreasing function of the consumption level as the utility function is concave, see Figure 9.5. Together, this creates a positive risk premium, $E R_i^e > 0$. That is, *an asset is risky if it is procyclical*. (Recall that risky assets have high risk premia since otherwise no one would like to buy those assets.)

Remark 9.14 (*Linearizing $U'(c)$) A first-order Taylor approximation of marginal utility around \bar{c} is $U'(\bar{c}) \approx U'(\bar{c}) + U''(\bar{c})(c - \bar{c})$. The numerator in (9.19) can thus be written $-\text{Cov}[U'(c_2), R_i^e] \approx -U''(\bar{c}) \text{Cov}(c_2, R_i^e)$, where $-U''(\bar{c}) > 0$ since the utility function is concave.

Although these results were derived from a two-period problem, it can be shown that a problem with more periods gives the same first-order conditions. In this case, the objective function is

$$U(c_1) + \delta E_1 U(c_2) + \delta^2 E_1 U(c_3) + \dots \delta^{T-1} E_1 U(c_T). \quad (9.20)$$

Remark 9.15 (*Investment level in period 1, I_1) The first order condition for I_1 is that the derivative of (9.14) wrt I_1 is zero, $-U'(c_1) + \delta E_1[U'(c_2)(1 + R_p)] = 0$, where $U'(c_t)$ is the marginal utility in period t . This says that consumption should be planned such that the marginal loss of utility from investing (decreasing c_1) equals the discounted expected marginal gain of utility from increasing c_2 by the gross return on the investment. For instance, with logarithmic utility we get $E_1 c_2/c_1 = \delta(1 + R_f)$, which says that when R_f is high, then the expected (planned) consumption path is upward sloping. (It is also clear that $I_2 = 0$ since investing in period 2 is just a waste.)

9.5.2 The Stochastic Discount Factor or Pricing Kernel*

More advanced asset pricing theory often work with a *stochastic discount factor* (SDF) or pricing kernel. For the optimization problem (9.14)–(9.17) this could be

$$M_2 = \delta U'(c_2) / U'(c_1). \quad (9.21)$$

Rewrite the first order condition for v_i (9.18) (notice that $\delta/U'(c_1)$ makes no difference since it is known at the time of investment) and the expression for expected excess returns (9.19) as

$$E(MR_i^e) = 0, \text{ and} \quad (9.22)$$

$$E R_i^e = -\text{Cov}(M, R_i^e) / E M, \quad (9.23)$$

where we (for convenience) drop all time subscripts. (Similarly, the foc in Remark 9.15 can be written $E[M(1 + R_p)] = 1$.) Notice that the *same* SDF is used for each asset i . Such SDFs can be derived in many ways (here we used a consumption plan approach), but they typically imply an expression like (9.22).

Remark 9.16 (Pricing with a stochastic discount factor, SDF*) Let M_{t+1} be an SDF and x_{t+1} the payoff of an asset in $t + 1$. Most asset pricing theories imply that the price today of the asset today (P_t) must satisfy (a) $P_t = E_t M_{t+1} x_{t+1}$. This implies that the gross return must satisfy (b) $E_t M_{t+1}(1 + R_{t+1}) = 1$ and the excess returns must satisfy (c) $E_t M_{t+1} R_{t+1}^e = 0$.

9.5.3 The Equity Premium Puzzle*

Remark 9.17 (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 $\text{Cov}[x, h(y)] = \text{Cov}(x, y) E[h'(y)]$.

With CRRA utility, $c^{1-\gamma}/(1-\gamma)$, the SDF is

$$M_2 = \delta(c_2/c_1)^{-\gamma}$$

If the excess return, R^e , and consumption growth, Δc , have a bivariate normal distribution, then by using Stein's lemma, we can rewrite the risk premium (9.23) as (again dropping time subscripts)

$$\mathbb{E} R^e = \text{Cov}(R^e, \Delta c)\gamma \quad (9.24)$$

$$= \text{Corr}(R^e, \Delta c) \text{Std}(R^e) \text{Std}(\Delta c)\gamma. \quad (9.25)$$

The “equity premium puzzle” is that, over a long U.S. sample of the equity market and consumption per capita, $\mathbb{E} R^e \approx 0.08$, $\text{Corr}(R^e, \Delta c) \approx 0.15$, $\text{Std}(R^e) \approx 0.2$ and $\text{Std}(\Delta c) \approx 0.02$, so an implausibly high risk aversion ($\gamma \approx 133$) is required to account for the high risk premia on the equity market. Basically, consumption is not volatile enough to explain the risk premium. See [Cochrane \(2005\)](#) for an extensive analysis.

Proof of (9.24). Stein's lemma gives $\text{Cov}[R^e, \exp(\ln M)] = \text{Cov}(R^e, \ln M) \mathbb{E} M$. (In terms of Stein's lemma, $x = R^e$, $y = \ln M$ and $h() = \exp()$.) Finally, notice that $\text{Cov}(R^e, \ln M) = -\gamma \text{Cov}(R^e, \Delta c)$. \square

9.5.4 From a Consumption-Based Model to CAPM

Suppose the marginal utility (or stochastic discount factor) in equilibrium, is an affine function of the market excess return

$$U'(c) = a - bR_m^e, \text{ with } b > 0. \quad (9.26)$$

This would, for instance, be the case in a Lucas model where consumption equals the market return and the utility function is quadratic—but it could be true in other cases as well. From (9.23) some rearrangements we get

$$\mathbb{E} R_i^e = \beta_i \mathbb{E} R_m^e, \text{ where } \beta_i = \sigma_{im}/\sigma_m^2, \quad (9.27)$$

which is the standard CAPM expression. This means that CAPM is consistent with (some) multi-period utility based portfolio choice models.

Proof of (9.27). Using (9.26) in (9.19) gives $\mathbb{E} R_i^e = b\sigma_{im}/\mathbb{E}(a - bR_m^e)$. We can, of course, apply this expression to the market excess return (instead of asset i) to get $\mathbb{E} R_m^e = b\sigma_m^2/\mathbb{E}(a - bR_m^e)$. Solve for $b/\mathbb{E}(a - bR_m^e)$ and use that in the first equation to

get (9.27). \square

9.5.5 From a Consumption-Based Model to a Multi-Factor Model

The consumption-based model is really a one-factor model, in terms of consumption. However, the relevant consumption level can be tricky to measure: both the treatment of durables and the identification of those who are actively investing are non-trivial. However, it may be reasonable to assume that we can approximate this (hard to observe) factor with a linear combination of other variables. Many macro models would suggest that a small number of (more easily measured) variables can provide an approximation. Also, the models may have missing factors that could be correlated with the state of the economy. For instance, marginal utility (or the stochastic discount factor) in equilibrium could be

$$U'(c) = ay + br, \quad (9.28)$$

where y denotes output and r the real interest rate.

It is then possible to write (9.23) as

$$\mathbb{E} R_i^e = \beta_{iy} \mu_y^e + \beta_{ir} \mu_r^e, \quad (9.29)$$

where (β_{iy}, β_{ir}) are from a multiple regression of R_{it}^e on excess returns on assets that are perfectly correlated with y and r respectively (“factor mimicking portfolios”), while (μ_y^e, μ_r^e) are the corresponding average excess returns. (The proof is in the Appendix.) The more general insight is that when the marginal utility (or stochastic discount factor) is linear in K factors, then we get a K -beta model for average returns.

9.6 Testing Multi-Factors Models

Let R_{ot}^e be a vector of factor *excess returns*. Testing whether $\alpha = 0$ in the regression

$$R_{it}^e = \alpha + \beta' R_{ot}^e + \varepsilon_{it} \quad (9.30)$$

is then the key approach to assess the model. (This test is invalid if some factors are not excess returns.)

The t-test of the null hypothesis that $\alpha_i = 0$ uses the fact that, under fairly mild

conditions, the t-statistic has an asymptotically normal distribution, that is

$$\frac{\hat{\alpha}_i}{\text{Std}(\hat{\alpha}_i)} \xrightarrow{d} N(0, 1), \quad (9.31)$$

under the null hypothesis ($\alpha_i = 0$). The standard error could be based on iid assumption or (better) account for heteroskedasticity.

Fama and French (1993) 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). Although this three-factor model is rejected at traditional significance levels, it still captures a fair amount of the variation of expected returns and may thus be a useful model.

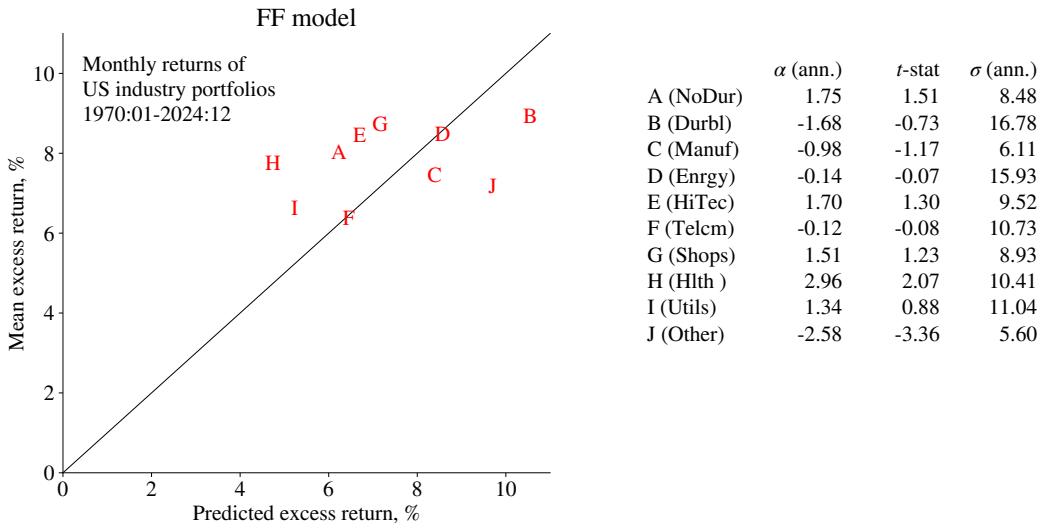
Remark 9.18 (*Fama-French factors*) *Fama and French (1993) use three factors: the market excess return, the return on a portfolio of small stocks minus the return on a portfolio of big stocks (SMB), and the return on a portfolio with a high ratio of book value to market value minus the return on a portfolio with a low ratio (HML). All three are excess returns (although only the first is in excess of a risk-free return), since they are long-short portfolios. He and Ng (1994) find that SMB is related to macroeconomic risks, but HML less so.*

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. For such (non-return) factors it is common to use factor mimicking portfolios: the excess return on portfolios strongly correlated with the factors.

Empirical Example 9.19 *Figure 9.6 shows some results for the Fama-French model on US industry portfolios and Figures 9.7–9.8 on the 25 Fama-French portfolios, both for more recent samples than in the original articles. The results indicate that the FF model is a considerable improvement compared to CAPM for the 25 FF portfolios, but perhaps not so much for the industry portfolios. Even for the 25 FF portfolios, strict statistical tests reject also the FF model, but the fit of the average returns is clearly better than CAPM.*

9.7 Appendix – The Asset Pricing Implications*

Proof (of (9.9)). Assume k in (9.8) is such that w equals the (financial) “market” portfolio, w_m in (9.8). For any portfolio with portfolio weights w_p , the covariance with the market



Fama-French model
 Factors: US market, SMB (size), and HML (book-to-market)
 Predicted excess return: $\beta_m \bar{R}_m^e + \beta_{SMB} \bar{R}_{SMB} + \beta_{HML} \bar{R}_{HML}$

10% crit. value (Bonferroni): 2.58

Test if all $\alpha_i = 0$:
 Wald stat 23.28
 5% crit val 18.31
 p-value 0.01

Figure 9.6: Fama-French regressions on US industry indices

return (times $1 - \phi$) is

$$\begin{aligned}(1 - \phi) \sigma_{pm} &= (1 - \phi) w_p' \Sigma w_m \\ &= w_p' \Sigma \Sigma^{-1} (\mu^e / k - S_c \phi) \text{ using (9.8)} \\ &= \mu_p^e / k - \phi \sigma_{pc},\end{aligned}$$

since $w_p' \mu^e = \mu_p^e$ and $w_p' S_c = \sigma_{pc}$. Rearrange to put μ_p^e / k on the LHS and rewrite as

$$\mu_p^e / k = \begin{bmatrix} 1 - \phi & \phi \end{bmatrix} \begin{bmatrix} \sigma_{pm} \\ \sigma_{pc} \end{bmatrix} \quad (*)$$

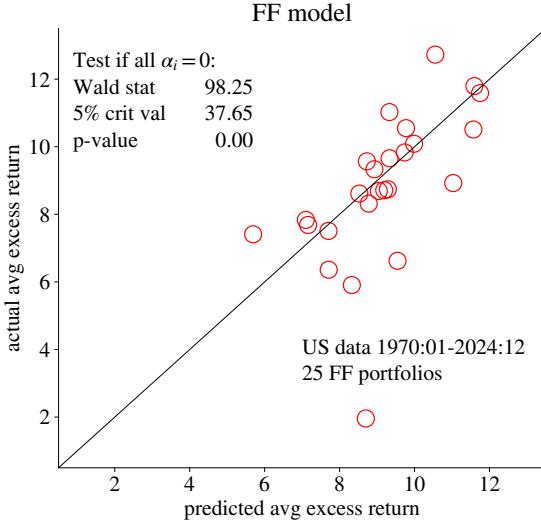


Figure 9.7: FF, FF portfolios

Recall that $\sigma_{p\lambda} = \sigma_{pc}$ where λ indicates the factor mimicking portfolio with excess return R_λ^e (see Remark 9.9) and rewrite as

$$\begin{aligned}\mu_p^e/k &= [1 - \phi \quad \phi] \begin{bmatrix} \sigma_{mm} & \sigma_{m\lambda} \\ \sigma_{m\lambda} & \sigma_{\lambda\lambda} \end{bmatrix} \begin{bmatrix} \sigma_{mm} & \sigma_{m\lambda} \\ \sigma_{m\lambda} & \sigma_{\lambda\lambda} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{pm} \\ \sigma_{p\lambda} \end{bmatrix} \\ &= [1 - \phi \quad \phi] \begin{bmatrix} \sigma_{mm} & \sigma_{m\lambda} \\ \sigma_{m\lambda} & \sigma_{\lambda\lambda} \end{bmatrix} \begin{bmatrix} \beta_{pm} \\ \beta_{p\lambda} \end{bmatrix},\end{aligned}\quad (**)$$

where $(\beta_{pm}, \beta_{p\lambda})$ are the coefficients from regressing R_p on (R_m^e, R_λ^e) . For the market return when $p = m$, $(\beta_{mm}, \beta_{m\lambda}) = (1, 0)$, so $(**)$ gives

$$\mu_m^e/k = [1 - \phi \quad \phi] \begin{bmatrix} \sigma_{mm} \\ \sigma_{m\lambda} \end{bmatrix}.$$

For the factor mimicking portfolio when $p = \lambda$, $(\beta_{\lambda m}, \beta_{\lambda\lambda}) = (0, 1)$, so $(**)$ gives

$$\mu_\lambda^e/k = [1 - \phi \quad \phi] \begin{bmatrix} \sigma_{m\lambda} \\ \sigma_{\lambda\lambda} \end{bmatrix}.$$

Use these last two equations to substitute for the first two terms on the RHS of $(**)$ and cancel the $1/k$ factors to get

$$\mu_p^e = [\mu_m^e \quad \mu_\lambda^e] \begin{bmatrix} \beta_{pm} \\ \beta_{p\lambda} \end{bmatrix},$$

which is (9.9). \square

*

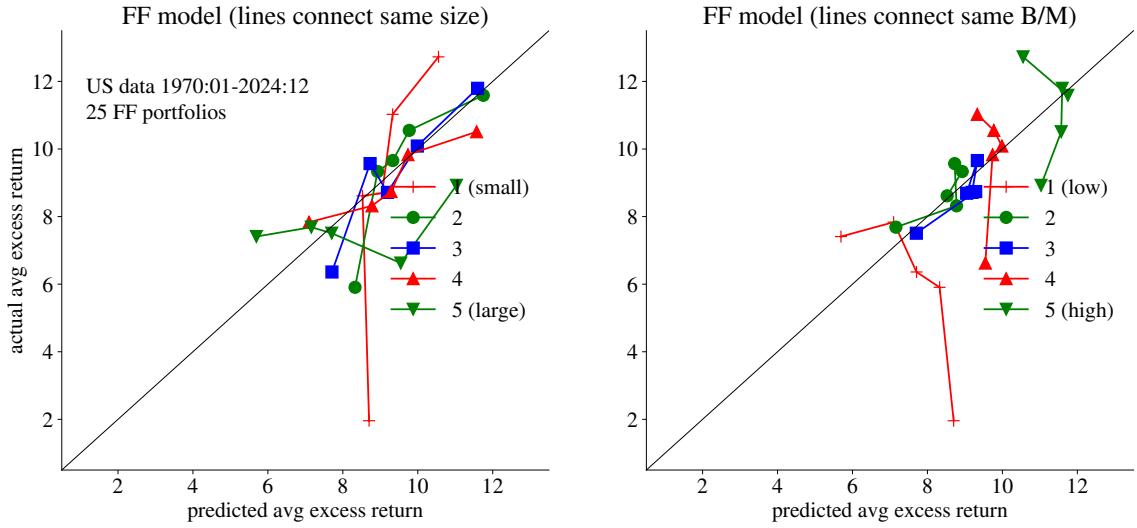


Figure 9.8: FF, FF portfolios

Proof (of (9.29)). Rewrite

$$E R_i^e = \frac{a\sigma_{iy} + b\sigma_{ir}}{-E(ay + br)} = \frac{1}{-E(ay + br)} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_{iy} \\ \sigma_{ir} \end{bmatrix}$$

Recall that $\sigma_{iv} = \sigma_{iy}$ and $\sigma_{i\rho} = \sigma_{ir}$, where R_v^e (R_ρ^e) is the factor mimicking portfolio of y (r), see Remark 9.9). Rewrite as

$$\begin{aligned} E R_i^e &= \frac{1}{-E(ay + br)} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_{vv} & \sigma_{v\rho} \\ \sigma_{v\rho} & \sigma_{\rho\rho} \end{bmatrix} \begin{bmatrix} \sigma_{vv} & \sigma_{v\rho} \\ \sigma_{v\rho} & \sigma_{\rho\rho} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{iv} \\ \sigma_{i\rho} \end{bmatrix} \\ &= \frac{1}{-E(ay + br)} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_{vv} & \sigma_{v\rho} \\ \sigma_{v\rho} & \sigma_{\rho\rho} \end{bmatrix} \begin{bmatrix} \beta_{iv} \\ \beta_{i\rho} \end{bmatrix}. \end{aligned} \quad (+)$$

Apply (+) to R_v^e to get $(\beta_{iv}, \beta_{i\rho}) = (1, 0)$, so

$$\mu_y^e = \frac{1}{-E(ay + br)} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_{vv} \\ \sigma_{v\rho} \end{bmatrix}.$$

Similarly, apply (+) to R_ρ^e to get $(\beta_{iv}, \beta_{i\rho}) = (0, 1)$, so

$$\mu_r^e = \frac{1}{-E(ay + br)} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_{v\rho} \\ \sigma_{\rho\rho} \end{bmatrix}.$$

Use these two expressions to rewrite (+) as

$$E R_i^e = \begin{bmatrix} \mu_y^e & \mu_r^e \end{bmatrix} \begin{bmatrix} \beta_{iv} \\ \beta_{i\rho} \end{bmatrix}.$$

□

Chapter 10

Efficient Markets

10.1 The Efficient Market Hypothesis

The efficient market hypothesis (EMH) says that it is very *hard to predict future asset returns*. If this is true (evidence is discussed later), then active management, such as security analysis and market timing is of limited use and incurs costs (management fees, trading costs). Instead, it might be more practical to apply a passive approach that satisfies individual requirements (diversification, hedging background risk, appropriate risk level, etc). The implications of the EMH are thus significant.

10.1.1 Different Versions of the Classical Efficient Market Hypothesis

A precise formulation of the EMH needs to specify three things.

First, what type of information is used in making the forecasts? Is it price and trading volume data (referred to as the weak form of the EMH), all public information (the semi-strong form), or perhaps all public and private information (the strong form)? Most modern analysis is focused on the weak or semi-strong forms as private information is likely to have predictive power.

Second, what is supposed to be unpredictable? Most modern financial theory would focus on *excess returns*, since they represent risk compensation.

Third, what is the link between predictability and expectations? If an excess return is almost unpredictable, then *rational* investors would have nearly constant expected risk premia, and portfolio weights are likely to be fairly stable over time. The opposite holds if excess returns are straightforward to predict.

Rejection of the EMH, based on statistical studies of ex post samples, can have different sources: changes in risk or in risk aversion (both rational reasons) or in inefficiencies. It

is typically very hard to disentangle these.

This chapter will present methods and empirical results. The first sections deal with traditional in-sample methods, initially focusing on the return history of the same asset, but later broadening the scope to bring in other types of predictors (fundamental valuation ratios, lagged returns of other assets, etc.) Later sections will instead focus on out-of-sample methods (recursive regressions, trading strategies, etc), and some evidence on the performance of professional forecasters.

10.2 Autocorrelations and Autoregressions

Autocorrelations and autoregressions are tools for studying whether past and current returns can predict future returns (typically of the same asset).

10.2.1 Autocorrelation Coefficients

The autocovariances of the R_t process can be estimated as

$$\hat{\gamma}_s = \frac{1}{T} \sum_{t=1+s}^T (R_t - \bar{R})(R_{t-s} - \bar{R}), \text{ where} \quad (10.1)$$

$\bar{R} = \Sigma_{t=1}^T R_t / T$ is the sample average estimated from the full sample. (In time series analysis we typically divide by T in (10.1) even if there are only $T - s$ observations to estimate γ_s from.) In most of the applications of this chapter, R_t indicates either a return or an excess return.

Autocorrelations are then estimated as

$$\hat{\rho}_s = \hat{\gamma}_s / \hat{\gamma}_0. \quad (10.2)$$

The sampling properties of $\hat{\rho}_s$ are complicated, but there are several useful large sample results for Gaussian processes (these results typically carry over to processes which are similar to the Gaussian). When the true autocorrelations (for $s > 1$) are all zero, then for any lag s different from zero

$$\sqrt{T} \hat{\rho}_s \xrightarrow{d} N(0, 1), \quad (10.3)$$

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

Example 10.1 (*t-test*) *Reject the null hypothesis that $\rho_1 = 0$ on the 10% significance level if $\sqrt{T} |\hat{\rho}_1| > 1.64$.*

Empirical Example 10.2 Figures 10.1–10.2 show autocorrelations for daily U.S. equity data.

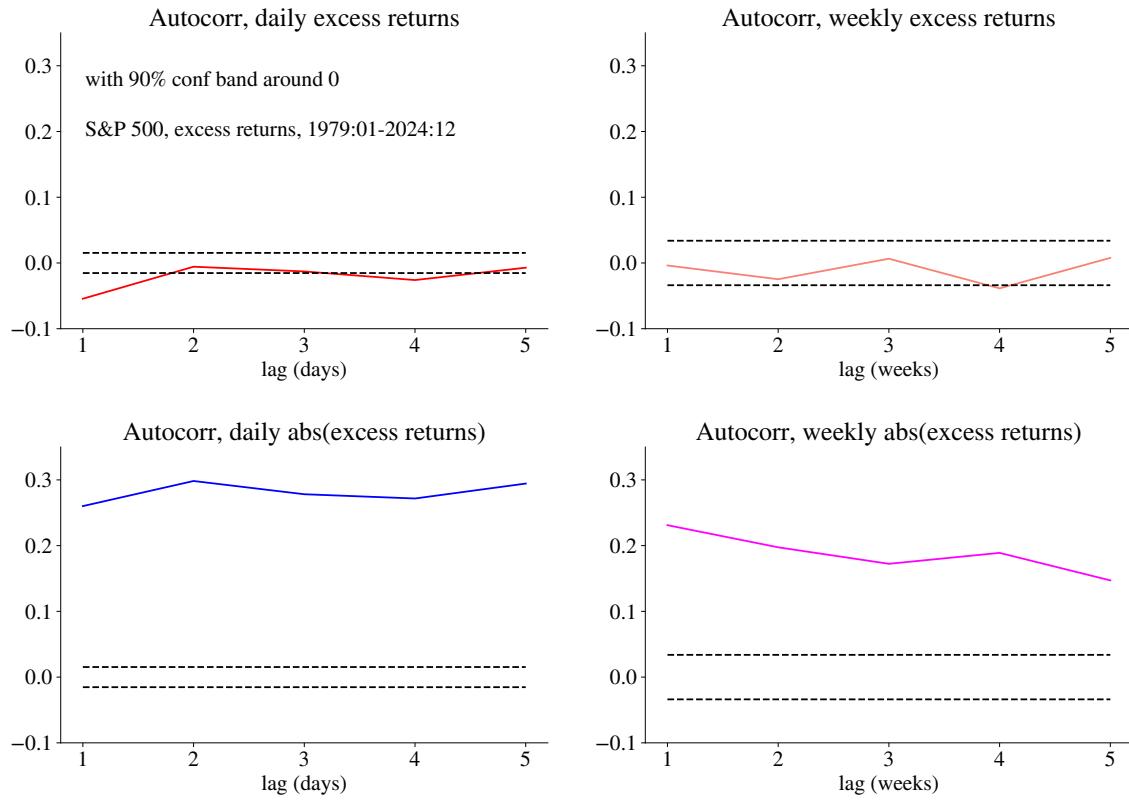


Figure 10.1: Predictability of US stock returns

10.2.2 Autoregressions

An alternative method for testing autocorrelations is to estimate an AR model

$$R_t = c + a_1 R_{t-1} + a_2 R_{t-2} + \dots + a_p R_{t-p} + \varepsilon_t, \quad (10.4)$$

and test if all slope coefficients (a_1, a_2, \dots, a_p) are zero with a χ^2 or F test. This approach is somewhat less general than testing if all autocorrelations are zero, but is easy to implement.

Empirical Example 10.3 Table 10.1 shows results from estimating an AR model on daily data for S&P 500 returns.

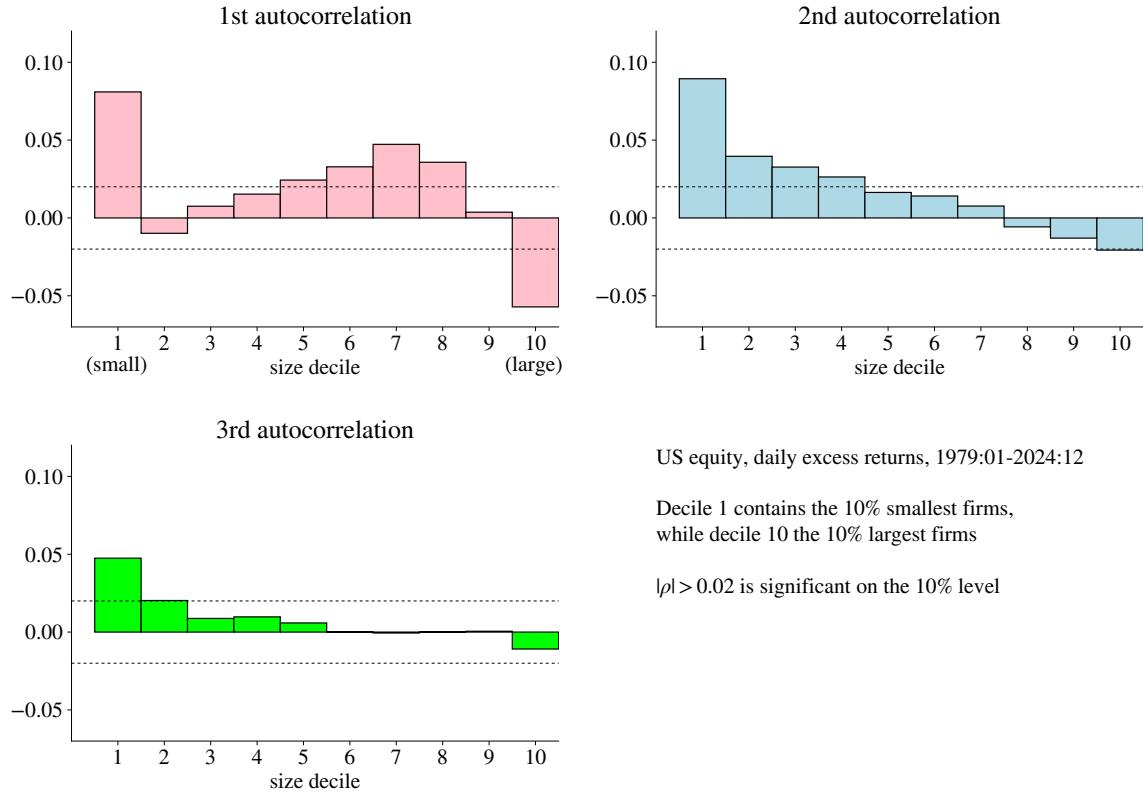


Figure 10.2: Predictability of US stock returns, size deciles

The autoregression can also handle non-linear patterns. For instance, consider an AR(1), but where the autoregression coefficient may be different depending on the sign of last period's return

$$R_t = \alpha + \beta Q_{t-1} R_{t-1} + \gamma(1 - Q_{t-1}) R_{t-1} + \varepsilon_t, \text{ where} \quad (10.5)$$

$$Q_{t-1} = 1 \text{ if } R_{t-1} < 0 \text{ and } 0 \text{ otherwise.}$$

Empirical Example 10.4 Figure 10.3 shows regression results from daily S&P 500 data. The reversal back after a negative shock is the most prominent finding.

Empirical Example 10.5 Figure 10.4 shows results from autoregressions for different investment horizons. For the business cycle frequency (3-4 years), there is some evidence of negative autocorrelation, that is, reversals. However, testing long-run returns is challenging because it requires a very long sample to have enough (non-overlapping) return periods, and it is unclear if data from long ago is informative about today's economy.

	Daily return
lag 1	−0.06 (−2.73)
lag 2	−0.02 (−0.70)
lag 3	−0.02 (−0.95)
lag 4	−0.03 (−1.95)
lag 5	−0.01 (−0.53)
c	0.03 (3.17)
R^2	0.01
All slopes	0.00
obs	10896

Table 10.1: AR(5) of daily S&P returns 1979:01-2024:12. Numbers in parentheses are t-stats, based on Newey-West with 3 lags. All slopes is the p-value for all slope coefficients being zero.

The empirical evidence reported in this section suggest little autocorrelation for daily returns for large-cap stocks (like those in S&P 500), but perhaps more for smaller firms. There is also some indication of non-linearity, with more autocorrelation in down markets. For longer return horizons, there appear to be some negative autocorrelation on the business cycle frequency.

10.3 Other Predictors and Methods

There are many other possible predictors of future stock returns. For instance, lagged returns on other assets might predict returns, and both the dividend-price ratio and interest rates have been used to predict long-run returns.

10.3.1 Lead-Lags

Stock indices have more positive autocorrelation than (most) individual stocks: there should therefore be fairly strong cross-autocorrelations among individual stocks. Indeed, this is also what is found in US data where returns of large size stocks forecast returns of

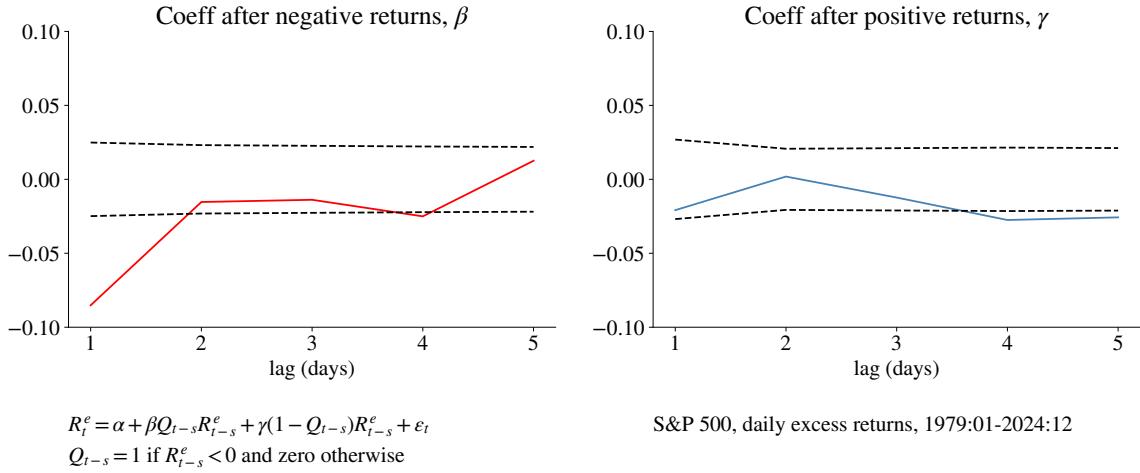


Figure 10.3: Predictability of US stock returns, results from a regression with interactive dummies

small size stocks.

Empirical Example 10.6 *Figure 10.5 shows (for different size deciles) the regressions coefficients on the 1-day own lag and the 1-day lag of large caps. The results suggest considerable spillover from large caps to the other size deciles.*

10.3.2 Earnings-Price Ratio as a Predictor

One of the most effective methods of forecasting long-run returns is a regression of future returns on the current earnings-price (or dividend-price) ratio

$$R_{s,t}^e = \alpha + \beta_q \ln(e_{t-s}/p_{t-s}) + \varepsilon_t, \quad (10.6)$$

where $R_{s,t}^e$ is the s -period excess return over the period $t - s$ to t .

Empirical Example 10.7 *Figure 10.6 shows results from estimating (10.6) for different investment horizons on data for a U.S. stock market index.*

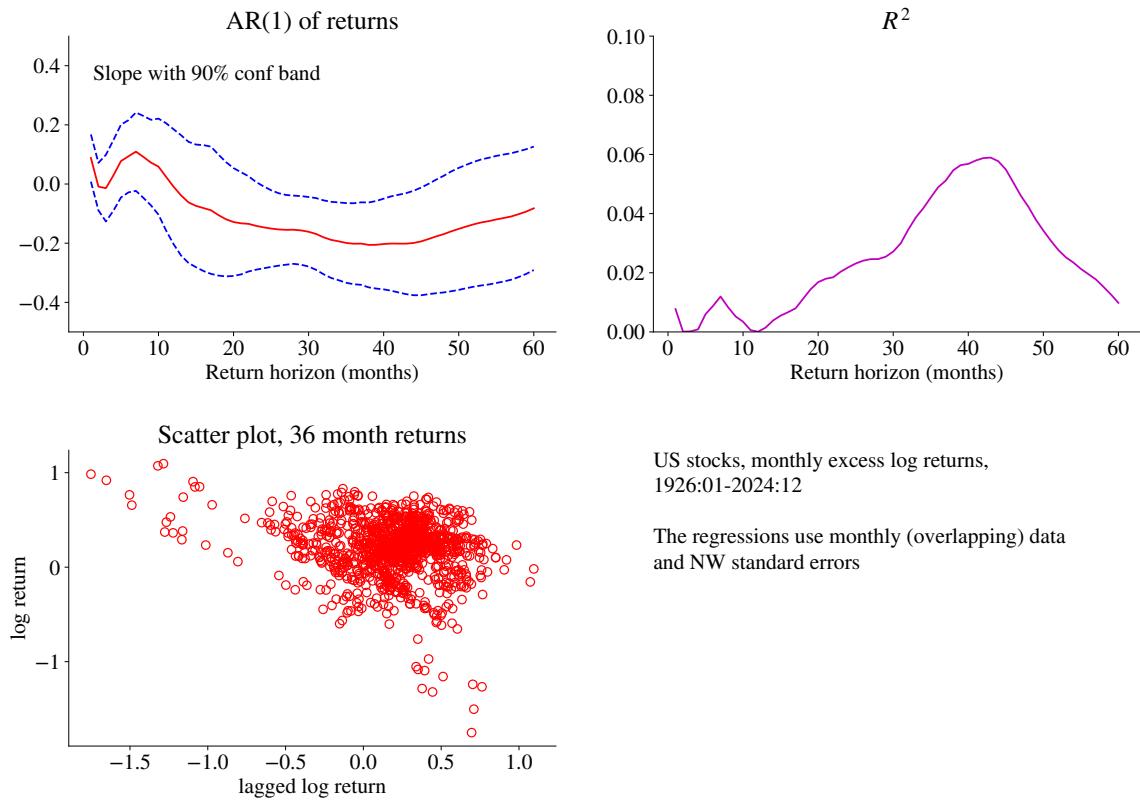


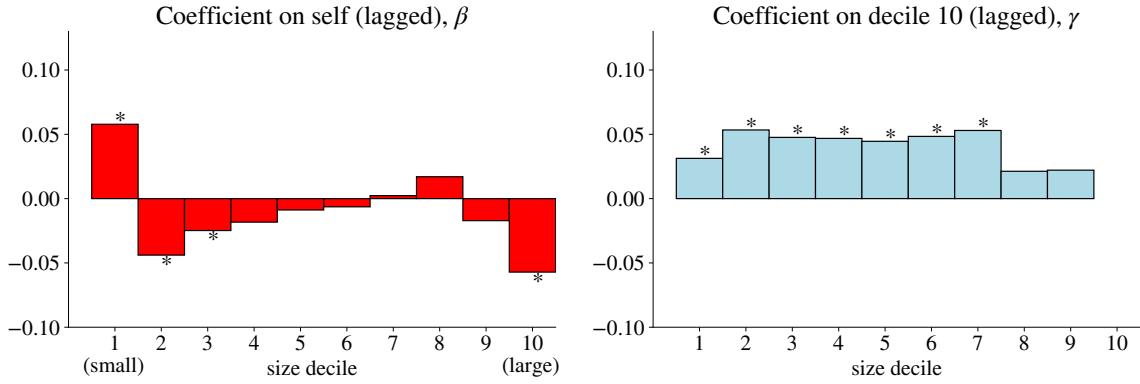
Figure 10.4: Predictability of long-run US stock returns

10.4 Out-of-Sample Forecasting Performance

10.4.1 In-Sample versus Out-of-Sample Forecasting

In-sample evidence on predictability can potentially be misleading because of (a) in-sample overfitting; and/or (b) structural breaks.

To gauge the out-of-sample predictability, forecasts are made using historical data and then updated as we get more information. To be precise, estimation is done using data up to and including $t - 1$ and a forecast is made for t . Then we use data up to and including t to make a forecast for $t + 1$ and so forth. This is called a recursive approach as the data sample is extended. See Figure 10.7. An alternative is to instead use a moving data window (ending in $t - 1$ and then in t, \dots) where really old data points are discarded. Yet another approach is downweight old data. In either case, the forecasting performance is often compared with a benchmark model, for instance, using the historical average as the prediction estimated on the same sample.



$$R_{i,t}^e = \alpha + \beta R_{i,t-1}^e + \gamma R_{10,t-1}^e + \epsilon_t$$

* indicates significance on the 10% level

US equity, daily excess returns, 1979:01-2024:12

Decile 1 contains the 10% smallest firms,
while decile 10 the 10% largest firms

Figure 10.5: Coefficients from multiple prediction regressions

One way to illustrate the relative forecast performance is the out-of-sample coefficient of determination (denoted R_{OS}^2 , not to be confused with a return)

$$R_{OS}^2 = 1 - \sum_{t=s}^T (R_t - \hat{R}_t)^2 / \sum_{t=s}^T (R_t - \tilde{R}_t)^2, \quad (10.7)$$

where s is the first period with an out-of-sample forecast, \hat{R}_t is the forecast based on the prediction model (estimated on data up to and including $t-1$) and \tilde{R}_t is the prediction from some benchmark model (typically also estimated on data up to and including $t-1$).

Example 10.8 (R_{OS}^2)

$$R_{OS}^2 = 1 - \frac{0.4}{0.5} = 0.2 \text{ (your model is better)}$$

$$R_{OS}^2 = 1 - \frac{0.5}{0.4} = -0.25 \text{ (your model is worse)}$$

To statistically *test* the relative predictive performance, define

$$g_t = (R_t - \hat{R}_t)^2 - (R_t - \tilde{R}_t)^2, \quad (10.8)$$

and test if the sample average, \bar{g} , differs from zero, a method described by Diebold and Mariano (1995). Instead of squared forecast errors, we could consider absolute values or an indicator of whether the sign is right. If there is little or no autocorrelation in g_t , then

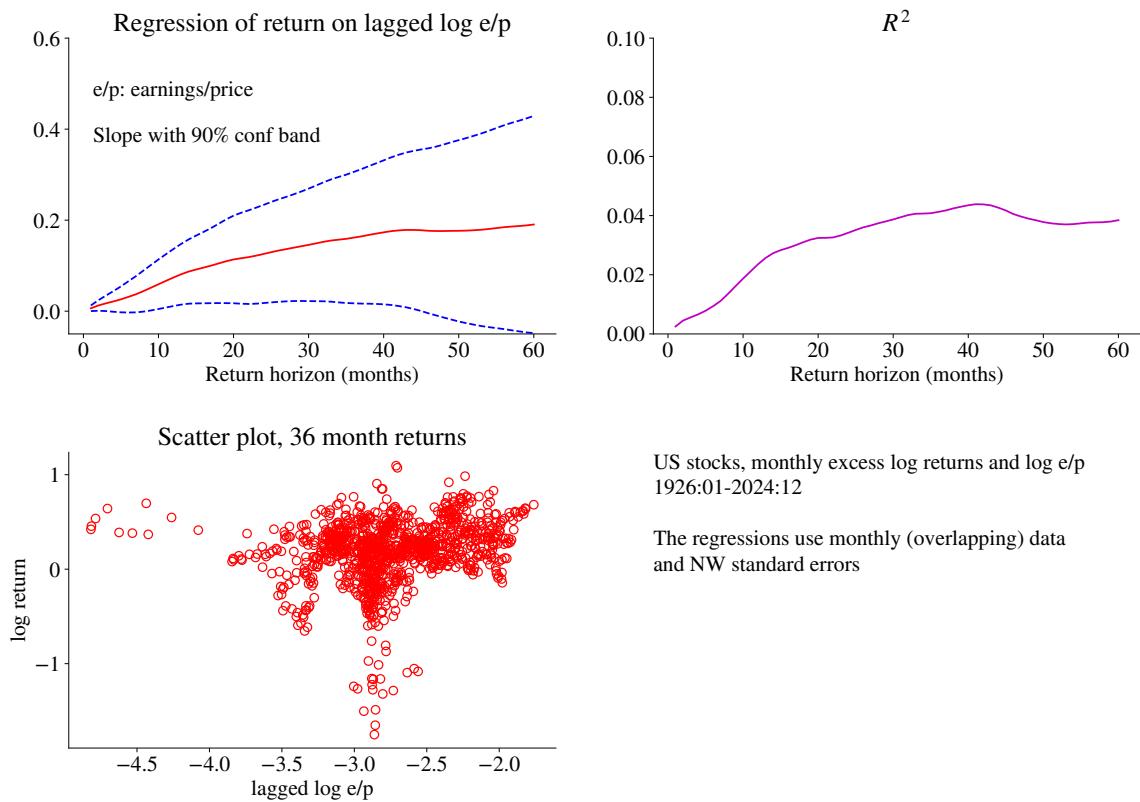


Figure 10.6: Predictability of long-run US stock returns

$\text{Var}(\bar{g}) = \text{Var}(g_t)/T$ so the t -stat is

$$\frac{\bar{g}}{\text{Std}(g_t)/\sqrt{T}}, \quad (10.9)$$

which could be compared with a $N(0, 1)$ distribution.

Empirical Example 10.9 *Figure 10.8 shows results based on daily data for different size deciles. It seems as if an AR(1) model is better than the historical average for small caps, but worse for large caps.*

Empirical Example 10.10 *Figure 10.9 shows how an e/p regression for a U.S. stock market index compares with the historical average—at different investment horizons. It seems as if it's consistently worse, which is similar to the findings of Goyal and Welch (2008).*

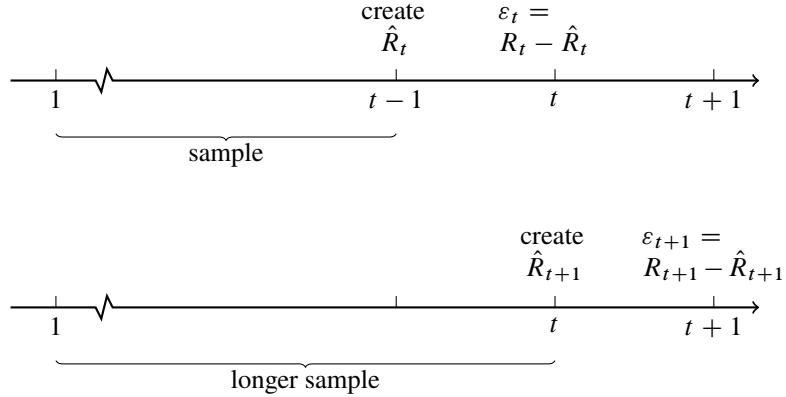


Figure 10.7: Out-of-sample forecasting

10.4.2 Trading Strategies

Another way to assess predictability and to illustrate its economic importance is to calculate the return of a *dynamic trading strategy*. In particular, the *alpha* (α) from a regression on the market excess return, $R_t^e = \alpha + \beta R_{mt}^e + \varepsilon_t$, is a useful measure. Neutral performance requires $\alpha = 0$, which can be tested with a t test.

Empirical Example 10.11 See Figure 10.10 for an empirical example based on a momentum strategy (bet on recent winners, bet against recent losers) on daily data for the 25 FF portfolios. The upper left figure shows that the strategy has high average returns and α . It also shows that frequent rebalancing is important for the performance. The lower left figure illustrates the magnitude of trading costs that the strategy can handle, while still generating a positive average excess return. The upper right figure instead investigates the importance of a formation lag: having a time gap between the sorting and portfolio formation. The results suggest a short gap, perhaps shorter than a week.

10.4.3 Technical Analysis

Technical analysis is typically a data mining exercise which looks for local trends or systematic non-linear patterns (see, for instance, Brock, Lakonishok, and LeBaron (1992)). The basic idea is that markets are not instantaneously efficient, so prices may exhibit delayed and predictable reactions to news. In practice, technical analysis amounts to analysing different transformations (for instance, a moving average) of prices—and to spot patterns. This section summarizes some simple models/trading rules.

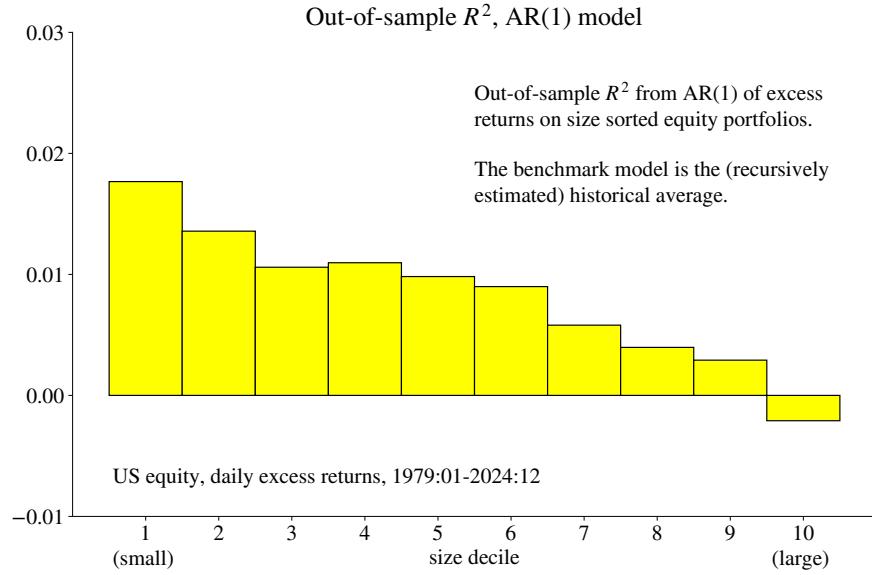


Figure 10.8: Short-run predictability of US stock returns, out-of-sample.

Many trading rules rely on some kind of local trend which can be thought of as positive autocorrelation in price movements (also called momentum).

A *moving average rule* involves buying when a short moving average exceeds a long moving average. The idea is that this signals a new upward trend. Let S be the lag order of a short moving average and L of a long moving average, with $S < L$ and let b be a bandwidth (perhaps 0.01). Then, a MA rule for period t could be

$$\begin{cases} \text{buy in } t \text{ if } MA_{t-1}(S) > MA_{t-1}(L)(1 + b) \\ \text{sell in } t \text{ if } MA_{t-1}(S) < MA_{t-1}(L)(1 - b) \\ \text{no change} \quad \quad \quad \text{otherwise} \end{cases}, \text{ where} \quad (10.10)$$

$$MA_{t-1}(x) = (p_{t-1} + \dots + p_{t-x})/x.$$

The difference between the two moving averages is called an *oscillator*

$$\text{oscillator}_t = MA_t(S) - MA_t(L), \quad (10.11)$$

(or sometimes, moving average convergence divergence, MACD) and the sign is often taken as a trading signal (this is the same as a moving average crossing, MAC). A version of the moving average oscillator is the *relative strength index*¹, which is the ratio of average

¹Not to be confused with relative strength, which typically refers to the ratio of two different asset prices

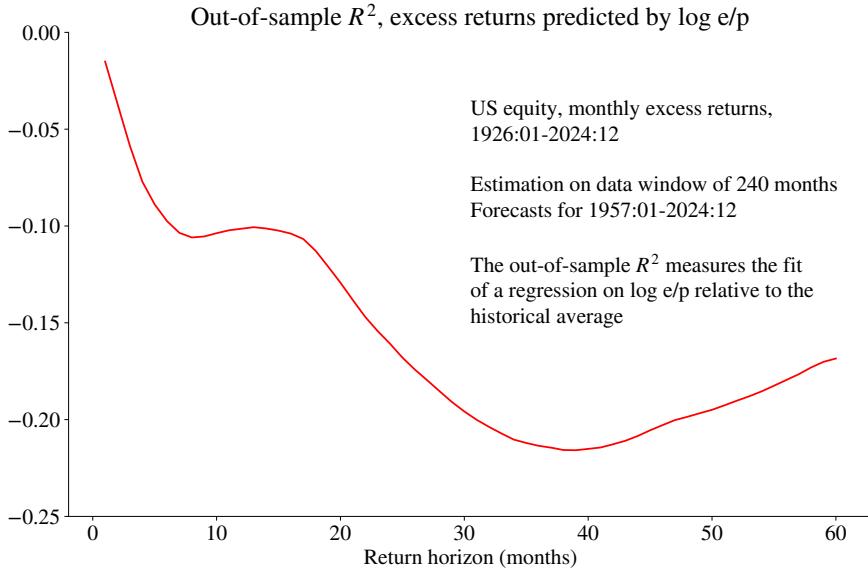


Figure 10.9: Predictability of long-run US stock returns, out-of-sample

price level (or returns) on “up” days to the average price (or returns) on “down” days—during the last z (14 perhaps) days. Yet another version is to compare the oscillator $_t$ to a moving average of the oscillator (also called a signal line).

The *trading range break-out rule* generally involves buying when the price rises above a previous peak (local maximum). The idea is that a previous peak is a *resistance level* in the sense that some investors are willing to sell when the price reaches that value (round numbers often play the role as resistance levels). Once this (artificial?) resistance level has been broken, the price can possibly rise substantially. On the downside, a *support level* plays the same role: some investors are willing to buy when the price reaches that value. To implement this, it is common to let the resistance/support levels be proxied by minimum and maximum values over a data window of length L . With a bandwidth b (perhaps 0.01), the rule for period t could be

$$\begin{bmatrix} \text{buy in } t \text{ if } P_t > M_{t-1}(1+b) \\ \text{sell in } t \text{ if } P_t < m_{t-1}(1-b) \\ \text{no change otherwise} \end{bmatrix}, \text{ where} \quad (10.12)$$

$$M_{t-1} = \max(p_{t-1}, \dots, p_{t-S})$$

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

(for instance, an equity compared to the market).

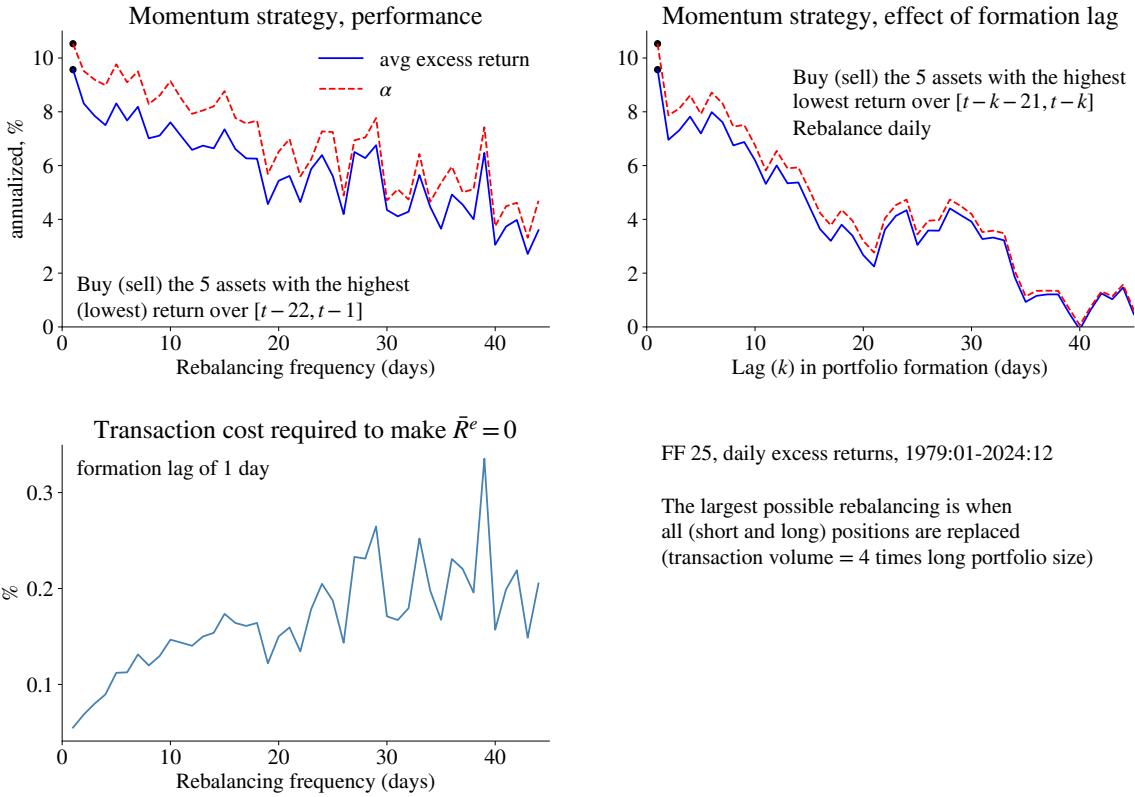


Figure 10.10: Predictability of US stock returns, momentum strategy

When the price is already trending up, then the trading range break-out rule may be replaced by a *channel rule*, which works as follows. First, draw a *trend line* through previous lows and a *channel line* through previous peaks. Extend these lines. If the price moves above the channel (band) defined by these lines, then buy. A version of this is to define the channel by a *Bollinger band*, which is ± 2 standard deviations from a moving data window around a moving average.

If we instead believe in mean reversion of the prices, then we can reverse the previous trading rules: we would typically sell when the price is high.

Empirical Example 10.12 *Figure 10.11 shows the idea of a reversal rule for S&P 500: buy when recent (a short MA) index values are outside the medium term trend (a long MA). The performance of implementing this rule over a long sample is shown in Table 10.2. The evidence suggest that the buy and sell signals do contain some information: average returns are high after buy signals, but that this may come at the cost of higher uncertainty (see Lo, Mamaysky, and Wang (2000) for a detailed study).*

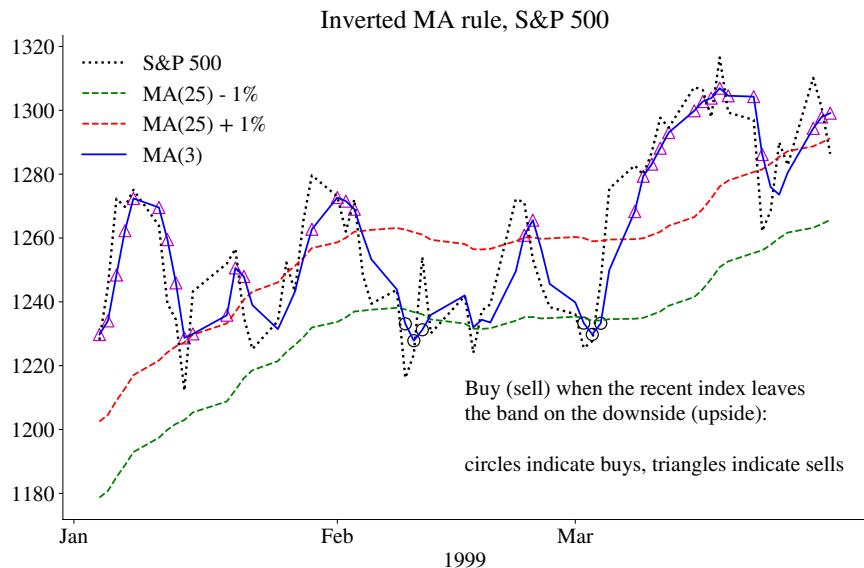


Figure 10.11: Example of a trading rule, illustration over short subsample

10.5 Security Analysts

Reference: Makridakis, Wheelwright, and Hyndman (1998) 10 and Elton, Gruber, Brown, and Goetzmann (2014) 27

To do: update this section with more recent evidence

10.5.1 Evidence on Analysts' Performance

Makridakis, Wheelwright, and Hyndman (1998) show that there is little evidence that the average stock analyst beats (on average) the market (or a passive index portfolio). In fact, less than half of the analysts beat the market. However, there are analysts which seem to outperform the market for some time, but the autocorrelation in over-performance is weak. The evidence from mutual funds is similar.

It should be remembered that many analysts also are sales persons: either of a stock (for instance, since the bank is underwriting an offering) or of trading services. It could well be that their objective function is quite different from minimizing the squared forecast errors. (The number of litigations in the US after the technology boom/bust should serve as a strong reminder of this.)

	Mean	Std
All days	7.2	18.0
After buy signal	16.2	27.3
After neutral signal	4.8	14.7
After sell signal	4.4	13.5
Strategy	9.1	27.5
Transaction cost	0.1	

Table 10.2: Excess returns (annualized, in %) from technical trading rule (Inverted MA rule). Daily S&P 500 data 1990:01-2024:12. The trading strategy involves (a) on every day: hold one unit of the index and short the risk-free; (b) on days after a buy signal: double the position in (a); (c) on days after a sell signal: short sell the position in (a), effectively having a zero investment. The transaction costs shows the cost (in %) of the trade volume that the strategy can pay and still perform as well as the static holding of (a).

10.5.2 Do Security Analysts Overreact?

The paper by Bondt and Thaler (1990) compares the (semi-annual) forecasts (one- and two-year time horizons) with actual changes in earnings per share (1976-1984) for several hundred companies. The paper has regressions like

$$\text{Actual earnings change} = \alpha + \beta(\text{forecasted earnings change}) + \text{residual},$$

and then studies the estimates of the α and β coefficients. With rational expectations (and a long enough sample), we should have $\alpha = 0$ (no constant bias in forecasts) and $\beta = 1$ (proportionality, for instance no exaggeration).

The main result is that $0 < \beta < 1$, so that the forecasted change tends to be too wild in a systematic way: a forecasted change of 1% is (on average) followed by a less than 1% actual change in the same direction. This means that analysts in this sample tended to be too extreme—to exaggerate both positive and negative news.

10.5.3 High-Frequency Trading Based on Recommendations from Stock Analysts

Barber, Lehavy, McNichols, and Trueman (2001) give a somewhat different picture. They focus on the profitability of a trading strategy based on analyst recommendations. They use a huge data set (some 360,000 recommendations, US stocks) for the period 1985–1996. They sort stocks in to five portfolios depending on the consensus (average) recommendation—and redo the sorting every day (if a new recommendation is published).

They find that such a daily trading strategy gives an annual 4% abnormal return on the portfolio of the most highly recommended stocks, and an annual -5% abnormal return on the least favourably recommended stocks.

This strategy requires a lot of trading (a turnover of 400% annually), so trading costs would typically reduce the abnormal return on the best portfolio to almost zero. A less frequent rebalancing (weekly, monthly) gives a very small abnormal return for the best stocks, but still a negative abnormal return for the worst stocks. [Chance and Hemler \(2001\)](#) obtain similar results when studying the investment advise by 30 professional “market timers.”

10.5.4 Economic Experts

Several papers, for instance, [Bondt \(1991\)](#) and [Söderlind \(2010\)](#), have studied whether economic experts can predict the broad stock markets. The results suggests that they cannot. For instance, [Söderlind \(2010\)](#) shows that the economic experts that participate in the semi-annual Livingston survey (mostly bank economists) (*ii*) forecast the S&P worse than the historical average (recursively estimated), and that their forecasts are strongly correlated with recent market data (which in itself, cannot predict future returns).

10.5.5 Analysts and Industries

[Boni and Womack \(2006\)](#) study data on some 170,000 recommendations for a very large number of U.S. companies for the period 1996–2002. Focusing on revisions of recommendations, the papers shows that analysts are better at ranking firms within an industry than ranking industries.

10.5.6 Insiders

Corporate insiders *used to* earn superior returns, mostly driven by selling off stocks before negative returns. (There is little/no systematic evidence of insiders gaining by buying before high returns.) Actually, investors who followed the insider’s registered transactions (in the U.S., these are made public six weeks after the reporting period), also used to earn some superior returns. It seems as if these patterns have more or less disappeared.

10.5.7 Mutual Funds

The general evidence on mutual funds is that they, on average, have zero alphas (or worse, after fees), and that there is little persistence in overperformance, at least among good funds (possible exceptions: hedge funds and private equity funds), while bad funds can stay bad for a long while.

Chapter 11

Performance Analysis

11.1 Performance Evaluation

11.1.1 The Idea behind Performance Evaluation

Traditional performance analysis seeks to answer the question: “Should we include an asset in our portfolio, assuming future returns will follow the same distribution as in a historical sample.”

Most performance measures rely on mean-variance (MV) analysis; however, the full MV portfolio optimization problem is not solved from scratch in these cases. Instead, the performance measures can be seen as different approximations of the MV problem, where the issue is whether we should invest in fund p or in fund q . (A mix of the two is not considered.)

Although the analysis is based on the MV model, it does not assume that all portfolios conform to Capital Asset Pricing Model’s (CAPM’s) beta representation or that the market portfolio is the optimal choice for every investor. Rather, CAPM is used as an approximation.

Mutual fund evaluations typically find (i) neutral performance on average (or less due to trading costs and fees); (ii) poorer performance among large funds; (iii) better performance in less liquid and possibly less efficient markets; and (iv) little persistence in performance, except for very bad funds.

Example 11.1 (*Steadman’s funds**) “*How can a fund be this bad?*” (NYT, 1991) (*the four Steadman funds rank among the six worst performers of the 244 stock funds tracked by Lipper Analytical Services for the 15 years that ended on Oct. 31. The Oceanographic Fund comes in at No. 243 and Steadman American Industry Fund, No. 244*); “*Steadman’s*

creature just won't die" (*Forbes*, 1999); "*Those awful Steadman's funds returning under a new name*" (*Baltimore Sun*, 2002).

Several popular performance measures are related to the CAPM regression

$$R_t^e = \alpha + \beta R_{mt}^e + \varepsilon_t, \quad (11.1)$$

where $E \varepsilon_t = 0$ and $\text{Cov}(R_{mt}^e, \varepsilon_t) = 0$. In many cases, R_{mt}^e represents the excess return on the market, but it could be some other benchmark return, for instance, for a segment of the market.

Example 11.2 (*Statistics for the examples of performance evaluations*) *The examples below use the following information about portfolios m (the market), p, and q*

	α	β	$\text{Std}(\varepsilon)$	μ^e	σ
m	0.00	1.00	0.00	10.00	18.00
p	1.00	0.90	14.00	10.00	21.41
q	5.00	1.30	3.00	18.00	23.59

Table 11.1: Basic facts about the market and two other portfolios, α , β , and $\text{Std}(\varepsilon)$ are from CAPM regression: $R_{it}^e = \alpha + \beta R_{mt}^e + \varepsilon_{it}$

11.1.2 Alpha

The intercept (α) in the regression (11.1) is often used as a performance measure. In CAPM, α measures the risk adjusted return. To see that, construct a portfolio with the weight β_i on the market portfolio (or some other benchmark) and $1 - \beta_i$ on the risk-free asset. The excess return on this portfolio is

$$R_p^e = \beta_i R_m^e, \quad (11.2)$$

since $R_p = \beta_i R_m + (1 - \beta_i) R_f$. This portfolio has the same systematic risk (sensitivity to the market) as asset i . As a practical matter, it is typically straightforward to create this portfolio by investing β_i in an index tracking fund and the remainder in the riskfree asset.

The α is then the difference in average excess returns of two portfolios with the same systematic risk

$$\alpha_i = E R_i^e - E R_p^e. \quad (11.3)$$

Empirical Example 11.3 Table 11.2 shows the various performance measures for two large mutual funds. The Vanguard fund seems to perform best.

	α	SR	$M_i^2 - M_m^2$	AR	Treynor	T^2
Market	0.00	0.40	0.00		7.19	0.00
Putnam	-0.28	0.37	-0.61	-0.07	6.83	-0.35
Vanguard	1.68	0.53	2.26	0.46	10.00	2.81

Table 11.2: Performance Measures of Putnam Asset Allocation: Growth A and Vanguard Wellington, weekly data 1999:01-2024:12 (annualized figures)

11.1.3 Sharpe Ratio and M^2

Suppose we want to determine whether fund p is better than fund q for allocating *all* our savings in. Again, we don't allow a mix of them. With MV preferences, the answer is that p is better if it has a higher Sharpe ratio—defined as

$$SR = \mu^e / \sigma. \quad (11.4)$$

Intuitively, for a given level of volatility, we obtain the highest expected return.

Example 11.4 (Performance measure) From Example 11.2 we get the performance measures in Table 11.3.

	SR	$M_i^2 - M_m^2$	AR	Treynor	T^2
m	0.56	0.00		10.00	0.00
p	0.47	-1.59	0.07	11.11	1.11
q	0.76	3.73	1.67	13.85	3.85

Table 11.3: Performance Measures

Remark 11.5 (Sortino ratio) The Sortino ratio is an alternative to the Sharpe ratio. It replaces σ_p with a measure of variation on the downside (typically, the square root of a semivariance).

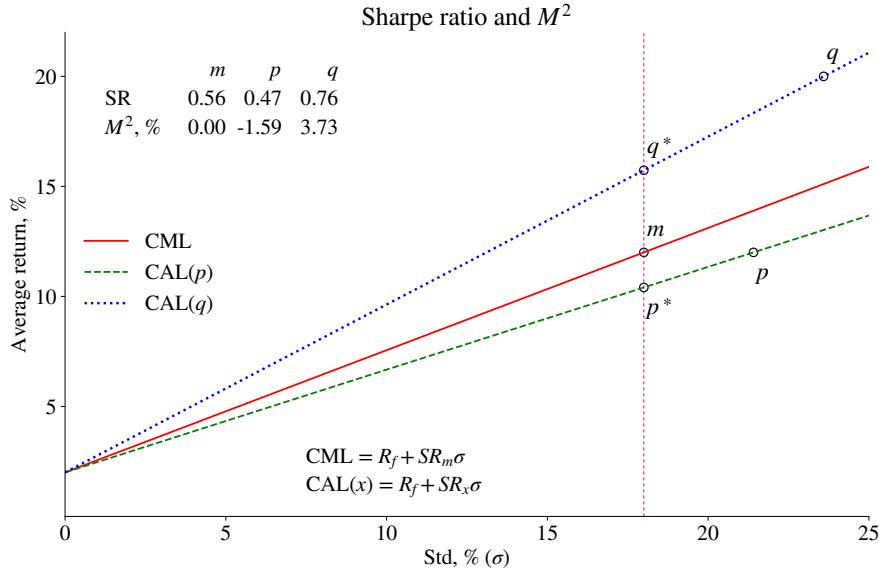


Figure 11.1: Sharpe ratio and M^2

The M^2 (“Modigliani and Modigliani”) measure is

$$M^2 = R_f + SR_p \sigma_m, \quad (11.5)$$

which is simple transformation of the Sharpe ratio. The difference of M^2 for portfolio p and the market (or another benchmark) m can also be written as a difference of two risk-adjusted expected returns

$$\begin{aligned} M_p^2 - M_m^2 &= (SR_p - SR_m) \sigma_m \\ &= \mu_{p^*}^e - \mu_m^e \end{aligned} \quad (11.6)$$

(or equivalently $\mu_{p^*}^e - \mu_m^e$). In this expression, $\mu_{p^*}^e$ is the expected excess return on a mix of portfolio p and the risk-free asset such that the volatility is the same as for the market return

$$R_{p^*} = a R_p + (1 - a) R_f, \text{ with } a = \sigma_m / \sigma_p. \quad (11.7)$$

The risk-adjustment here is thus to make the portfolios have the same volatility as the market. See Example 11.4 and Figure 11.1 for an illustration, which illustrate the relationship between the Sharpe ratio and M^2 , highlighting the differences in risk-adjusted average returns.

Proof of (11.6). Notice that $SR_p = SR_{p^*}$ and that $\sigma_{p^*} = \sigma_m$. The first line of (11.6)

can then be written $SR_{p^*}\sigma_{p^*} - SR_m\sigma_m$, which can be simplified as the 2nd line. \square

11.1.4 Appraisal and Information Ratios

If the question is “should I add fund p or fund q to my holding of the market portfolio?,” then the appraisal ratio provides an answer. The appraisal ratio is

$$AR = \alpha / \text{Std}(\varepsilon_t), \quad (11.8)$$

where α is the intercept and $\text{Std}(\varepsilon_t)$ is the standard deviation of the residual (“tracking error”) of the CAPM regression (11.1). If you think of (11.2) as the benchmark return, then AR is the average extra return per unit of extra standard deviation.

The motivation for AR indicating the best addition to the market portfolio is that the tangency portfolio based on *both* the market portfolio and portfolio p , has the following squared Sharpe ratio

$$SR_T^2 = AR^2 + SR_m^2. \quad (11.9)$$

(The proof is found below.) This is clearly increasing in AR (if positive), see Example 11.4 for an illustration.

The *information ratio*

$$IR_p = \frac{\text{E}(R_p - R_b)}{\text{Std}(R_p - R_b)}, \quad (11.10)$$

where R_b is some benchmark return. The information ratio is similar to both the Sharpe ratio and the appraisal ratio. The denominator in (11.10) can be thought of as the tracking error relative to the benchmark—and the numerator as the average active return (the gain from actively deviating from the benchmark). In fact, when the benchmark is as in (11.2), then the information ratio is the same as the appraisal ratio. Instead, when R_f is the benchmark, then the information ratio equals the Sharpe ratio.

Proof of (11.9). From the CAPM regression (11.1) we have

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

As usual, the square of the Sharpe ratio of the tangency portfolio is $\mu^{e'} \Sigma^{-1} \mu^e$. Combining, we get that the squared Sharpe ratio for the tangency portfolio (using both R_{it} and R_{mt}) is

$$\left(\frac{\mu_T^e}{\sigma_T} \right)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + \left(\frac{\mu_m^e}{\sigma_m} \right)^2.$$

\square

11.1.5 Treynor's Ratio and T^2

Suppose instead that the issue is whether we should add a *small* amount of fund p or fund q to an already well diversified portfolio (not necessarily the market portfolio). In this case, Treynor's ratio might be useful

$$TR = \mu^e / \beta. \quad (11.11)$$

The basic intuition is that, with a *diversified portfolio* and a *small investment*, idiosyncratic risk becomes negligible, whereas only systematic risk (β) remains significant.

The TR measure can be rephrased in terms of expected returns—and could then perhaps be called the T^2 measure

$$\begin{aligned} T^2 &= \mu_p^e / \beta_p - \mu_m^e \\ &= \mu_{p^*}^e - \mu_m^e \end{aligned} \quad (11.12)$$

(or equivalently, $\mu_{p^*}^e - \mu_m^e$). In this expression, $\mu_{p^*}^e$ is the expected excess return on a mix of portfolio p and the risk-free asset such that the beta is one (the same as for the market return)

$$R_{p^*} = aR_p + (1-a)R_f, \text{ with } a = 1/\beta_p, \quad (11.13)$$

so $\mu_{p^*}^e = \mu_p^e / \beta_p$. The risk-adjustment here is thus to make the portfolios have the same β as the market. See Example 11.4 and Figure 11.2 for an illustration which illustrate Treynor's Ratio and T^2 , highlighting the differences in risk-adjusted average returns.

11.1.6 Relationships among the Various Performance Measures

The different measures can give different answers when comparing portfolios, but they all share one thing: they are increasing in α . By using the expected values from the CAPM regression (11.2), $\mu_p^e = \alpha_p + \beta_p \mu_m^e$, simple rearrangements give

$$\begin{aligned} SR_p &= \frac{\alpha_p}{\sigma_p} + \text{Corr}(R_p, R_m)SR_m \\ AR_p &= \frac{\alpha_p}{\text{Std}(\varepsilon_{pt})} \\ TR_p &= \frac{\alpha_p}{\beta_p} + \mu_m^e. \end{aligned} \quad (11.14)$$

and M^2 is just a scaling of the Sharpe ratio. Notice that these expressions do not assume that CAPM is the right pricing model—we just use the definition of the intercept and slope

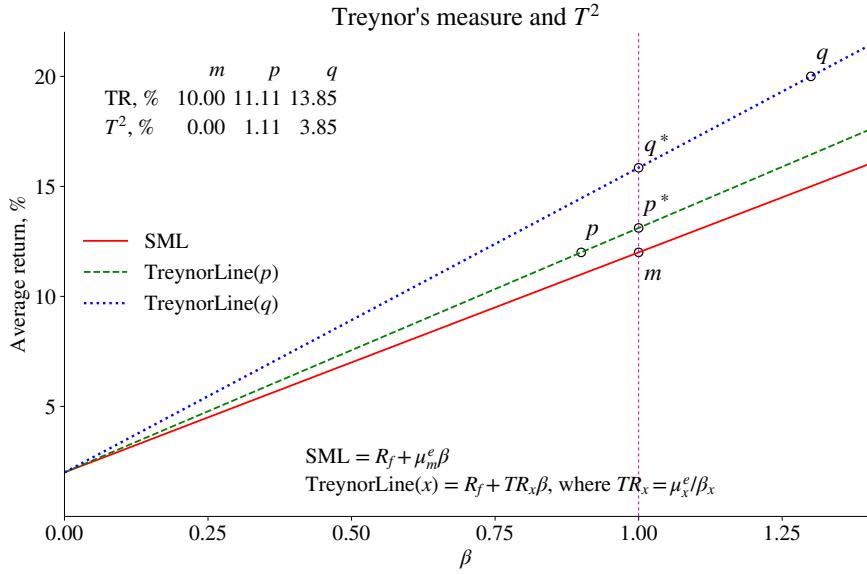


Figure 11.2: Treynor's ratio

in the CAPM regression.

Since alpha is the driving force in all these measurements, this further motivates its use as a performance measure in itself.

Proof of (11.14). Taking expectations of the CAPM regression (11.1) gives $\mu_p^e = \alpha_p + \beta_p \mu_m^e$, where $\beta_p = \text{Cov}(R_p, R_m) / \sigma_m^2$. The Sharpe ratio is therefore

$$SR_p = \frac{\mu_p^e}{\sigma_p} = \frac{\alpha_p}{\sigma_p} + \frac{\beta_p}{\sigma_p} \mu_m^e,$$

which can be written as in (11.14) since

$$\frac{\beta_p}{\sigma_p} \mu_m^e = \frac{\text{Cov}(R_p, R_m)}{\sigma_m \sigma_p} \frac{\mu_m^e}{\sigma_m}.$$

The AR_p in (11.14) is just a definition. The TR_p measure can be written

$$TR_p = \frac{\mu_p^e}{\beta_p} = \frac{\alpha_p}{\beta_p} + \mu_m^e,$$

where the second equality uses the expression for μ_p^e from above. \square

11.1.7 More Sophisticated Performance Measures

This section goes beyond CAPM to consider more sophisticated performance measures.

The logic of using α from a CAPM regression can be extended to a *multi-factor model*

where the factors are excess returns

$$R_t^e = \alpha + \beta_m R_{mt}^e + \beta_c R_{ct}^e + \dots + \varepsilon_t. \quad (11.15)$$

Once again α can be seen as a performance measure and the rest (excluding ε_t) as a portfolio that could be easily replicated.

If there are predictable movements in the market excess return, then it makes sense to add a “market timing” factor to the CAPM regression. For instance, [Treynor and Mazuy \(1966\)](#) argue that market timing is analogous to having a beta that varies linearly with the market excess return

$$\beta = b + c R_{mt}^e. \quad (11.16)$$

Using this in a traditional market model (CAPM) regression, $R_t^e = a_i + \beta R_{mt}^e + \varepsilon_t$, gives

$$R_t^e = a + b R_{mt}^e + c(R_{mt}^e)^2 + \varepsilon_t, \quad (11.17)$$

where c captures the ability to “time” the market. That is, if the investor systematically exits the market prior to periods of low returns and vice versa, then the slope coefficient c is positive. The interpretation is not clear cut, however. If we still regard the market portfolio as the benchmark, then $a + c(R_{mt}^e)^2$ could be counted as performance. In contrast, if we think that this sort of market timing is straightforward to implement, that is, if the benchmark is the market plus market timing, then only a should be counted as performance.

A recent way to merge the ideas of market timing and multi-factor models is to allow the coefficients to be time-varying according to some predetermined information (“state”) variable, z_{t-1} . To illustrate this, suppose z_{t-1} is a single variable, so the time-varying (or “conditional”) CAPM regression is

$$\begin{aligned} R_t^e &= (\theta_1 + \theta_2 z_{t-1}) + (\theta_3 + \theta_4 z_{t-1}) R_{mt}^e + \varepsilon_t \\ &= \theta_1 + \theta_2 z_{t-1} + \theta_3 R_{mt}^e + \theta_4 z_{t-1} R_{mt}^e + \varepsilon_t. \end{aligned} \quad (11.18)$$

Similar to the market timing regression, there are two possible interpretations of the results: if we still regard the market portfolio as the benchmark, then the other three terms should be counted as performance. In contrast, if the benchmark is a dynamic strategy in the market portfolio (where z_{t-1} is allowed to affect the choice market portfolio/risk-free asset), then only the first two terms are performance. In either case, the performance is time-varying.

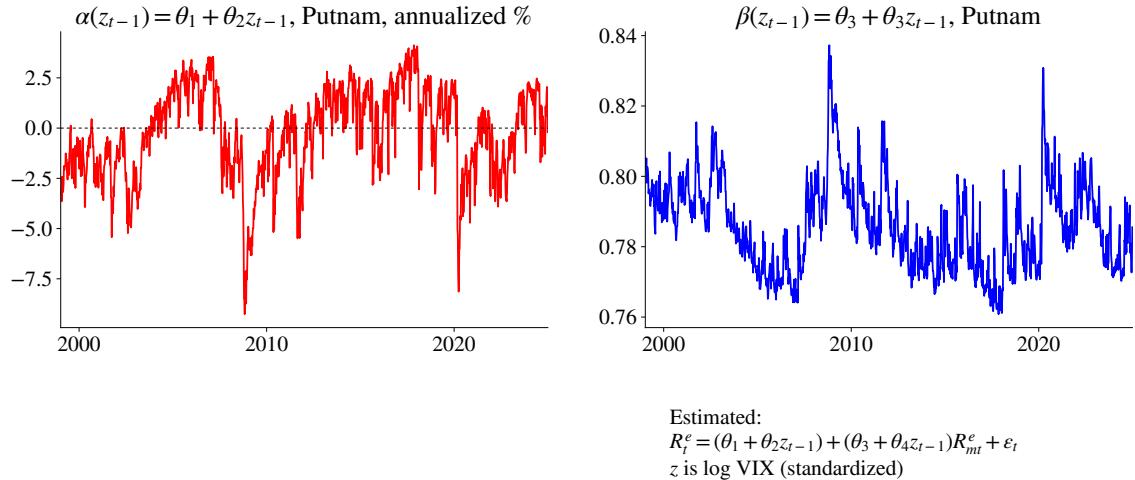


Figure 11.3: Conditional CAPM regression

Empirical Example 11.6 (Conditional CAPM regression) Figure 11.3 illustrates the results from estimating (11.18) for the Putnam fund, using the logarithm of VIX (standardized to have zero mean and unit variance) as the state variable. The results indicate modest swings in the effective β , but considerable movements in the effective α .

11.2 Holdings-Based Performance Measurement

As a complement to the purely return-based performance measurements discussed, it may be of interest to study how the portfolio weights change (if that information is available). This highlights *how* the performance has been achieved.

The Grinblatt-Titman measure (see Grinblatt and Titman (1993)) in period t is

$$GT_t = \sum_{i=1}^n (w_{i,t-1} - w_{i,t-2}) R_{it}, \quad (11.19)$$

where $w_{i,t-1}$ is the weight on asset i in the portfolio chosen (at the end of) in period $t-1$ and R_{it} is the return of that asset between (the end of) period $t-1$ and (end of) t . A positive value of GT_t indicates that the fund manager has moved into assets that turned out to give positive returns. Researchers commonly report the time-series average of GT_t .

11.3 Performance Attribution

The performance of an investment fund often depends on decisions taken on several levels of the organisation. To get a better understanding of where the performance was generated, a performance attribution calculation can be helpful. It uses information on portfolio weights (often in-house information) to decompose overall performance according to a number of criteria (typically related to different levels of decision making).

For instance, we could decompose the return into the effects of (a) allocation to asset classes (equities, bonds, bills); and (b) security choice within each asset class. Alternatively, for a pure equity portfolio, it could be the effects of (a) allocation to industries; and (b) security choice within each industry.

Consider portfolios p and b (for benchmark) from the same set of assets. Let n be the number of asset classes (or industries). Returns are

$$R_p = \sum_{i=1}^n w_i R_{pi} \text{ and } R_b = \sum_{i=1}^n v_i R_{bi}, \quad (11.20)$$

where w_i is the weight on asset class i (for instance, T-bonds) in portfolio p , and v_i is the corresponding weight in the benchmark b . Analogously, R_{pi} is the return that the portfolio earns on asset class i , and R_{bi} is the return the benchmark earns. In practice, the benchmark returns are typically taken from well established indices.

Form the difference and rearrange ($\pm w_i R_{bi}$) to get

$$R_p - R_b = \underbrace{\sum_{i=1}^n (w_i - v_i) R_{bi}}_{\text{allocation effect}} + \underbrace{\sum_{i=1}^n w_i (R_{pi} - R_{bi})}_{\text{selection effect}}. \quad (11.21)$$

The first term is the *allocation effect*, that is, the importance of allocation across asset classes, measured using the benchmark return of that asset class. If decisions on allocation to different asset classes are taken by senior management (or a board), then this is the contribution of that level. Instead, the second term is the *selection effect*, that is, the importance of selecting the individual securities within an asset class, as it depends on difference in returns that the fund and the benchmark earns in a given asset class. This contribution is more likely to come from the the trading desk.

Remark 11.7 (Alternative expression for the allocation effect*) The allocation effect is sometimes defined as $\sum_{i=1}^n (w_i - v_i) (R_{bi} - R_b)$, where R_b is the benchmark return. This is clearly the same as in (11.21) since $\sum_{i=1}^n (w_i - v_i) R_b = R_b \sum_{i=1}^n (w_i - v_i) = 0$.

11.3.1 What Drives Differences in Performance across Funds?

Much research shows that the asset allocation (choice between markets or large market segments) is more important for mutual fund returns than the asset selection (choice of individual assets within a market segment), see for instance, Ibbotson and Kaplan (2000). For other investors, including hedge funds, leverage also plays a role.

11.4 Style Analysis

Style analysis (Sharpe (1992)) is a way to use econometric tools to find out the portfolio composition from a series of the returns, at least in broad terms. This is clearly a bit cruder than having access to the actual portfolio weights (as discussed above), especially since the estimation requires some data points before the portfolio change can be detected.

The key idea is to identify several return indices (typically 5 to 10) believed to account for the majority of the portfolio's returns, followed by running regressions to find the portfolio "weights." It is essentially a multi-factor regression without a intercept and where the coefficients are constrained to sum to unity (and, optionally, to be positive)

$$R_{pt}^e = \sum_{j=1}^K b_j R_{jt}^e + \varepsilon_{pt}, \text{ with}$$

$$\sum_{j=1}^K b_j = 1 \text{ and } b_j \geq 0 \text{ for all } j.$$
(11.22)

Clearly, the restrictions could be changed to $U_j \leq b_j \leq L_j$, which could allow for some short positions.

The coefficients are typically estimated by minimizing the sum of squared residuals. In case the only restriction is that the coefficient should sum to one, then this can be solved with basic linear algebra (see the Remark below). With restrictions on the individual coefficient (for instance, no short sales), this is a non-linear least squares problem, but there are very efficient numerical methods for such problems.

Remark 11.8 (*Restricted OLS**) If we want to impose the restrictions $Rb = q$ on OLS where R is an $L \times K$ matrix and q is an $L \times 1$ vector, then the closed form solution is

$$\hat{b} = b_{OLS} - S_{xx}^{-1} R' (R S_{xx}^{-1} R')^{-1} (R b_{OLS} - q),$$

where $S_{xx} = \Sigma_{t=1}^T x_t x_t'$ (that is, $X'X$ if row t of X contains x_t') and b_{OLS} is the

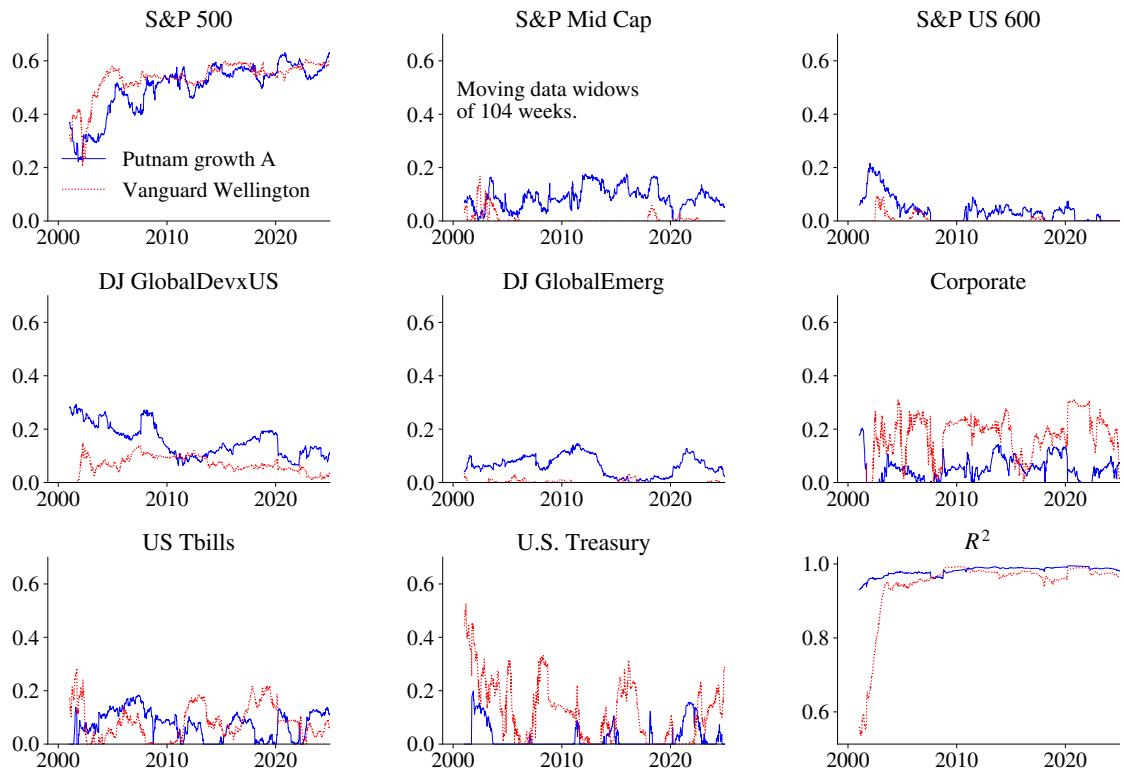


Figure 11.4: Example of style analysis, rolling data window

unrestricted OLS estimate. For instance, $R = \mathbf{1}'_K$ and $q = 1$ gives the style analysis solution (11.22) except that short sales are allowed.

A pseudo- R^2 (the squared correlation of the fitted and actual values) is sometimes used to gauge how well the regression captures the returns of the portfolio.

Empirical Example 11.9 See Figure 11.4 for an example of style analysis for the two mutual funds studied earlier. The results indicate that the coefficients move considerably over time (based on estimations from rolling data windows) and that the R^2 is above 95% except for the first few years.

Chapter 12

Investment for the Long Run

12.1 Time Diversification

This section discusses the notion of “time diversification,” which essentially amounts to claiming that equity is safer for long run investors than for short run investors. The argument comes in two flavours: (1) Sharpe ratios increase with the investment horizon and (2) the probability that equity returns outperform bond returns increases with the horizon.

This chapter will compare these findings with results from mean-variance (MV) analysis.

Empirical Example 12.1 *Figure 12.1 shows how, for the U.S. equity market index, the Sharpe ratio and the probability of outperforming a safe asset differ across investment horizons.*

12.1.1 Long-Run Return as a Sum of Short-Run Returns

This section explains how a long-run return can be expressed in terms of multiple short-run returns. In particular, we use logarithmic returns, since they can easily be cumulated over time.

The gross (buy-and-hold) return on a q -period investment made in period 0 can be written

$$1 + Z(q) = \prod_{t=1}^q (1 + R_t), \quad (12.1)$$

where R_t is the net portfolio return in period t . Taking logs (and using lower case letters to denote them), we have the log q -period return

$$z(q) = \sum_{t=1}^q r_t, \quad (12.2)$$

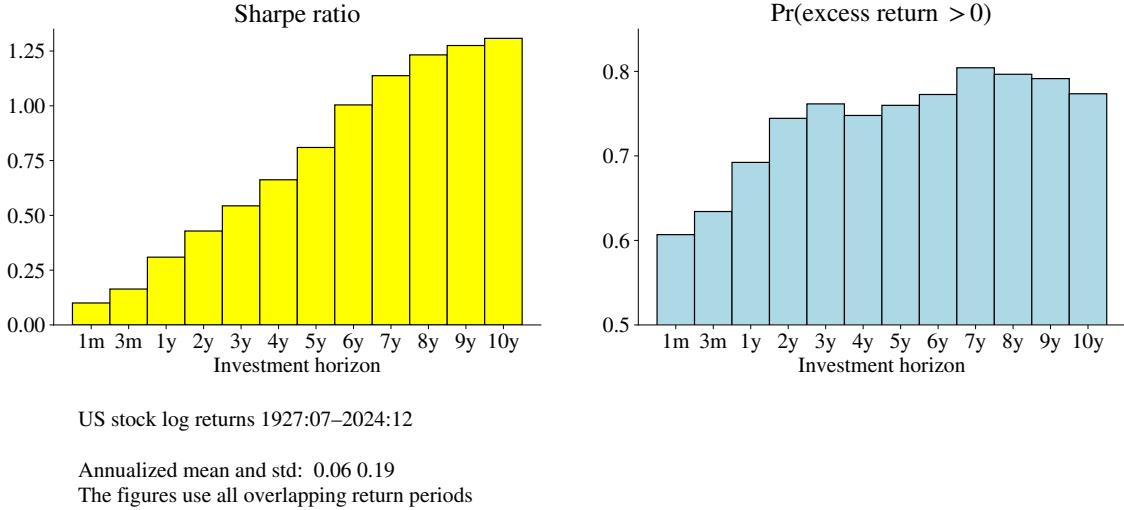


Figure 12.1: Empirical evidence on SR and probability of excess return > 0

where $z(q) = \ln(1 + Z(q))$ and where the log one-period return is $r_t = \ln(1 + R_t)$. Notice that if R is small, then $\ln(1 + R) \approx R$. We use r_t^e to denote the excess long return, $r_t^e = r_t - r_f$, where $r_f = \ln(1 + R_f)$, and similarly for $z^e(q)$.

Remark 12.2 (*Approximating q-period returns**) *It is sometimes convenient to approximate the q-period net return $Z(q)$ as*

$$Z(q) \approx \sum_{t=1}^q R_t.$$

This approximation works well unless there are numerous periods or extreme one-period returns. For instance, if $R_1 = 0.9$ and $R_2 = -0.9$ (indeed very extreme returns), then the two-period net return is $Z(2) = (1 + 0.9)(1 - 0.9) - 1 = -0.81$, while the approximation gives $Z(2) \approx R_1 + R_2 = 0$. The difference is dramatic. If the two net returns instead are $R_1 = 0.09$ and $R_2 = -0.09$, then $Z(2) = (1 + 0.09)(1 - 0.09) - 1 = -0.01$ and the approximation is still zero: this difference is much smaller.

Remark 12.3 (*Geometric mean returns**) *The average log return, $\bar{r} = \sum_{t=1}^q r_t/q$, is closely related to the geometric mean return. To see that, notice that a geometric mean gross return is $1 + \tilde{R} = [\prod_{t=1}^q (1 + R_t)]^{1/q}$, which equals $\exp(\bar{r})$. For values close to 0, $\tilde{R} \approx \bar{r}$.*

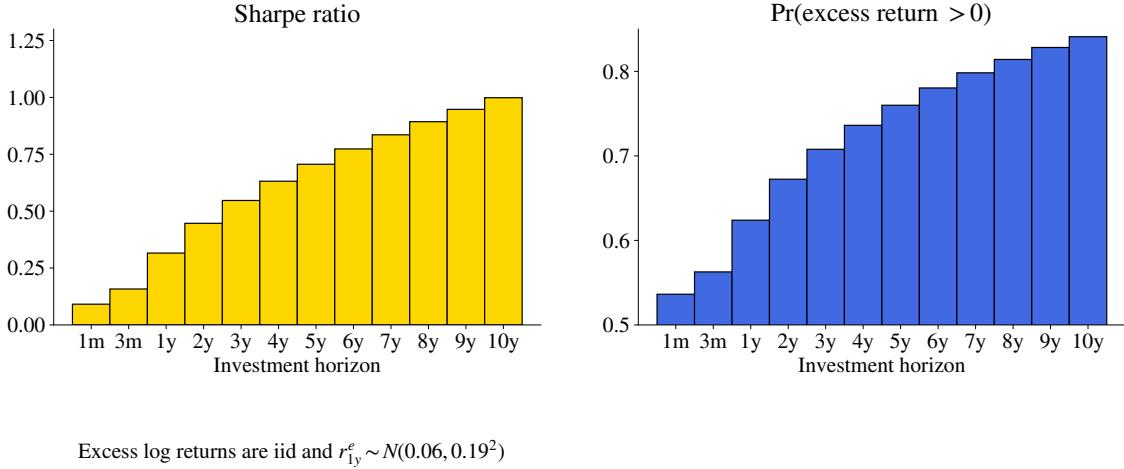


Figure 12.2: SR and probability of excess return > 0 , iid returns

12.1.2 Increasing Sharpe Ratios

This section demonstrates that with iid (independently and identically distributed) one-period returns, both the average and variance of a multi-period return grow linearly with the investment horizon. Consequently, the Sharpe ratio, defined as the expected excess return divided by the standard deviation, increases with the square root of horizon. This needs to be considered when comparing Sharpe ratios across investment horizons.

As before, let $z(q)$ be the log return on a q -period investment. *If log returns are iid*, the Sharpe ratio of $z(q)$ is

$$SR(z(q)) = \sqrt{q} SR(r), \quad (12.3)$$

where $SR(r)$ is the Sharpe ratio of the *one-period* log return. Clearly, this scales with the horizon, q . See Figure 12.2 for an illustration.

Proof of (12.3). The q -period log return is as in (12.2). If one-period excess log returns are iid with mean μ^e and variance σ^2 , then the mean and variance of the q -period excess log returns are $E z^e(q) = q\mu^e$ and $\text{Var}(z(q)) = q\sigma^2$. \square

12.1.3 Probability of Outperforming a Risk-Free Asset

Under the assumption of normally distributed returns, the increasing Sharpe ratios imply higher probabilities of out-performing a risk-free asset.

In particular, assume that the log one-period returns are jointly normally distributed,

which carries over to the q -period excess log return, $z^e(q)$. Then, we have

$$\Pr(z^e(q) > 0) = \Phi[SR(z(q))], \quad (12.4)$$

where $\Phi()$ is the cumulative distribution function of a standard normal variable, $N(0, 1)$. Again, see Figure 12.2 for an illustration.

Together with the results in (12.3) this suggests that the empirical evidence in Figure 12.1 could potentially be explained by iid returns.

Proof of (12.4). By standard manipulations we have that if $x \sim N(\mu, \sigma^2)$, then

$$\Pr(x \leq 0) = \Pr((x - \mu)/\sigma \leq -\mu/\sigma) = \Phi(-\mu/\sigma),$$

since $(x - \mu)/\sigma$ is an $N(0, 1)$ variable. Clearly, $\Pr(x > 0) = 1 - \Pr(x \leq 0)$. Use the fact that $\Phi(z) + \Phi(-z) = 1$ (since the standard normal distribution is symmetric around zero) and substitute $z^e(q)$ for x (and notice that μ/σ then corresponds to a Sharpe ratio) to get (12.4). \square

12.1.4 Why These Arguments Are Not Enough

Although increasing Sharpe ratios (at longer investment horizons) suggest a higher probability of out-performing a risk-free asset, that does not necessarily imply that the risky asset is safer for a long-run investor. We also have to take into account the size of the loss—in case the asset underperforms. With a longer horizon (and therefore higher dispersion), *really* bad outcomes are more likely: the expected loss, conditional of having one, is increasing with the investment horizon. See Figure 12.3 for an illustration.

Remark 12.4 (*Expected excess return conditional on a negative one**) If $x \sim N(\mu, \sigma^2)$, then $E(x|x \leq b) = \mu - \sigma\phi(b_0)/\Phi(b_0)$ where $b_0 = (b - \mu)/\sigma$ and where $\phi()$ and $\Phi()$ are the pdf and cdf of a $N(0, 1)$ variable respectively. To apply this, use $b = 0$ so $b_0 = -\mu/\sigma$. This gives $E(x|x \leq 0) = \mu - \sigma\phi(-\mu/\sigma)/\Phi(-\mu/\sigma)$.

To further explore how the investment horizon affects the portfolio weights, it is necessary to clarify the investor's preferences, specifically how risks and opportunities are compared.

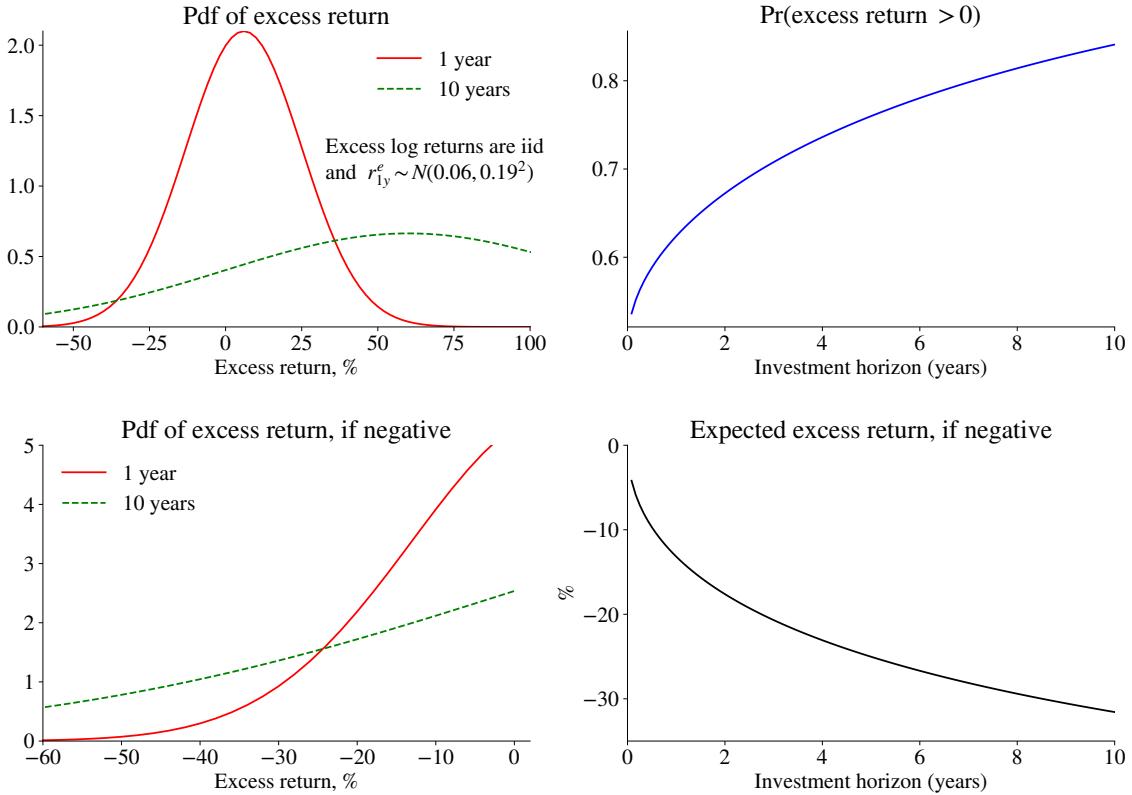


Figure 12.3: Time diversification, normally distributed returns

12.2 Mean-Variance Portfolio Choice

12.2.1 Approximating the Log Portfolio Return

Logarithmic portfolio returns are convenient in a dynamic setting since they are additive across time. However, they have a drawback on the portfolio formation stage: the logarithmic portfolio return is a *non-linear* function of the logarithmic returns of the assets. Therefore, we will use an approximation.

If there is only one risky asset with return Z and risk-free asset with return Z_f , then the portfolio return is $Z_p = vZ + (1 - v)Z_f$. For this case we approximate the log portfolio return as

$$z_p = \ln[v e^z + (1 - v)e^{z_f}] \quad (12.5)$$

$$\approx z_f + v(z - z_f) + v\sigma_z^2/2 - v^2\sigma_z^2/2, \quad (12.6)$$

where σ_z^2 is the variance of z over the relevant investment horizon (see Campbell and

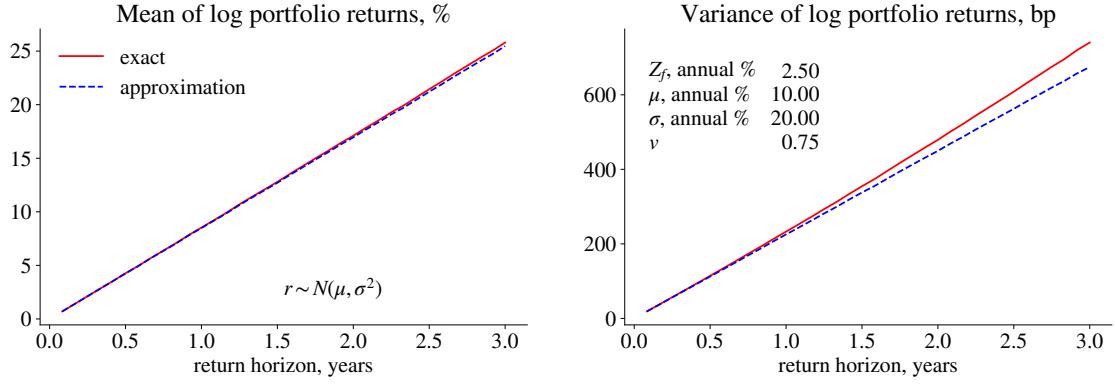


Figure 12.4: Mean and variance of log portfolio return for different return horizons, exact and according the approximation (12.6).

Viceira (2002)). As usual, all moments represent the beliefs of the investor, conditional on the information available at the time of investment. For convenience, we suppress the indicator q for the investment horizon.

The approximation error from (12.5) is straightforward to calculate, and it will differ across return levels. What is more important, however, is how the mean and variance differ between the exact calculation and the approximation. To assess that we need to assume a distribution of the returns. Figure 12.4 provides an illustration, where the basic assumption is that the log return on the risky asset is normally distributed, with a mean and standard deviation that are broadly in line with the U.S. equity market index. The figure suggests that both the mean and variance scale fairly linearly with the return horizon, and that the approximation works well up to 2–3 years. After that, we see a deviation, where the approximation underestimates the variance somewhat so a manual adjustment of the risk aversion parameter k might be sensible.

Proof of (12.6). The portfolio return $Z_p = vZ + (1 - v)Z_f$ can be used to write

$$(1 + Z_p)/(1 + Z_f) = 1 + v[(1 + Z)/(1 + Z_f) - 1].$$

The logarithm is

$$z_p - z_f = \ln\{1 + v[\exp(z - z_f) - 1]\}.$$

The function $f(x) = \ln\{1 + v[\exp(x) - 1]\}$, where $x = z - z_f$, has the following derivatives (evaluated at $x = 0$): $df(x)/dx = v$ and $d^2f(x)/dx^2 = v(1 - v)$, and notice that $f(0) = 0$. A second order Taylor approximation of the log portfolio return around $z - z_f = 0$ is then

$$z_p - z_f \approx v(z - z_f) + v(1 - v)(z - z_f)^2/2.$$

In a continuous time model, the square would equal its expectation, σ_z^2 , so this further approximation is used to give (12.6). \square

12.2.2 Mean-Variance Optimization

We assume that the investor solves a traditional mean-variance problem, but expressed in terms of log returns

$$\max_v \mathbb{E} z_p(q) - \frac{k}{2} \text{Var}(z_p(q)), \quad (12.7)$$

where $z_p(q)$ is the q -period log return of a portfolio of a risky and a risk-free asset. Notice that we further assume that the investor picks a portfolio and then stays with it, that is, we rule out rebalancing during the investment horizon. (A later chapter will relax that.)

Using the approximation (12.6), the optimal weight on the risky asset is

$$v = \frac{\mu_z^e(q)}{(1+k)\sigma_z^2(q)} + \frac{1}{2(1+k)}. \quad (12.8)$$

This is, of course, very similar to the traditional MV results based on net returns. In particular, the key driver is the mean excess return divided by the variance and the risk aversion.

Example 12.5 (of (12.8)) With $(\mu_z^e, \sigma_z^2, k) = (0.008, 0.05^2, 5)$ we get $v \approx 0.62$, but with $(\mu_z^e, \sigma_z^2) = (0.016, 0.0594^2)$ we get $v \approx 0.84$.

Proof of (12.8). Using (12.6), (12.7) is approximately the same as

$$\max_v z_f + v\mu_z^e + v\sigma_z^2/2 - v^2\sigma_z^2/2 - kv^2\sigma_z^2/2,$$

where z_f is the risk-free rate over the investment horizon and (μ_z^e, σ_z^2) are mean and variance of the excess log return of the risky asset. (The indicator q for the investment horizon is suppressed.) The first order condition is $\mu_z^e + \sigma_z^2/2 - v(1+k)\sigma_z^2 = 0$, which gives (12.8). \square

12.2.3 Mean-Variance Optimization with iid Logarithmic Returns

When log returns are iid, then both the mean and the variance scale with the investment horizon

$$\mu_z^e(q) = q \mathbb{E} r^e \text{ and } \sigma_z^2(q) = q \text{Var}(r), \quad (12.9)$$

where $\mathbb{E} r^e$ is the expected excess 1-period return and $\text{Var}(r)$ its variance.

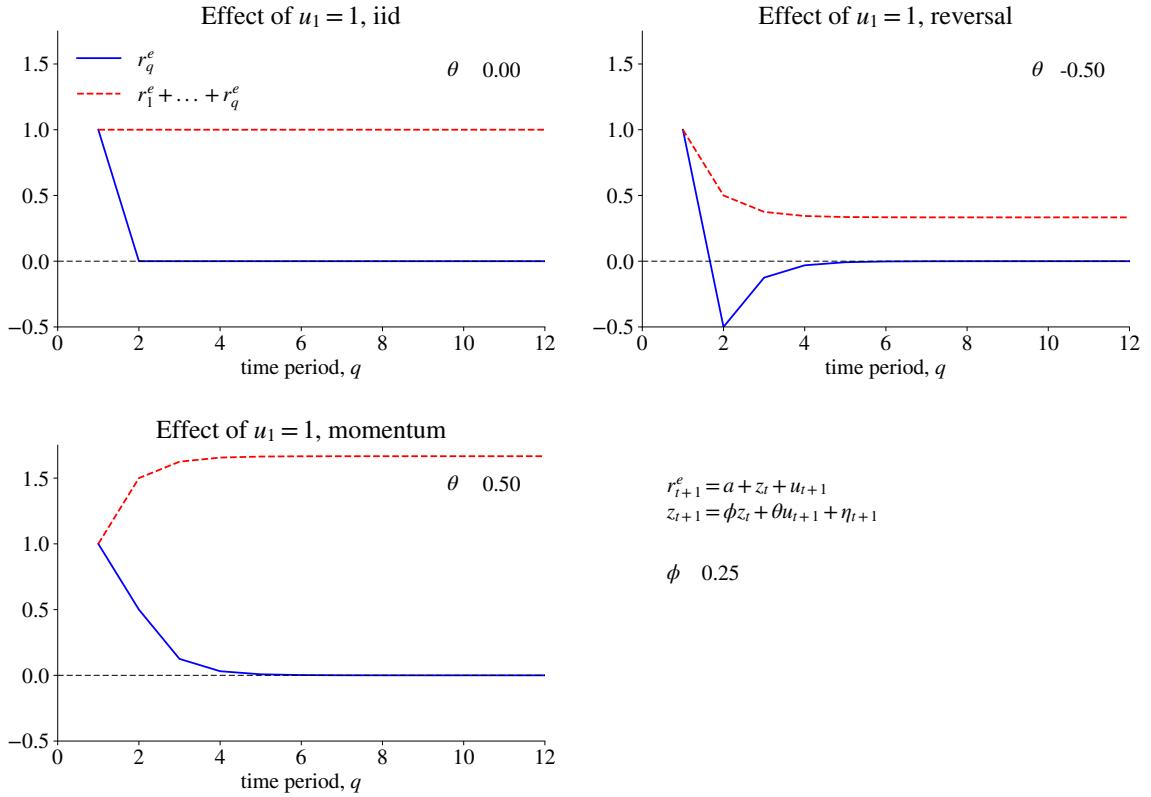


Figure 12.5: Impulse responses to an u_1 shock in the time series model (12.10)

This implies that, with iid log returns, the portfolio weight on the risky asset (12.8) is the same for all investment horizons, q . This is in stark contrast to what the increasing Sharpe ratios and probability of beating a risk-free asset would suggest. The key reason is that the mean-variance preferences consider also the magnitude of a loss, not just the probability of one.

12.2.4 A Time Series Model for Autocorrelated Logarithmic Returns

To discuss the case of autocorrelated returns, we use a simple time series model for 1-period log returns which allows for predictability

$$\begin{bmatrix} r_{t+1}^e \\ z_{t+1} \end{bmatrix} = \begin{bmatrix} a \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \begin{bmatrix} r_t^e \\ z_t \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \theta & 1 \end{bmatrix} \begin{bmatrix} u_{t+1} \\ \eta_{t+1} \end{bmatrix} \quad (12.10)$$

where u_{t+1} and η_{t+1} are iid and uncorrelated shocks, with standard deviations σ_u and σ_η , respectively. The z_t variable can be thought of as a state variable which affects future

average returns. The θ parameter controls how return shocks u_{t+1} spill over to future average returns, which will turn out to be a key feature. Formally, (12.10) is a state space model with correlated shocks, but it is written on VAR(1) form. For our purposes, the model has the advantage that it can generate both long-run momentum and reversals. (In contrast, an AR(1) with a negative coefficient has an oscillating forecast for future values.)

With $\theta = 0$ and $\sigma_\eta = 0$ (no spillover from u to z and no shocks to z), the return process is iid, which is the case discussed above. In contrast, with $\theta < 0$ a positive return shock in $t + 1$ will tend to be followed by negative returns, that is, a long-run reversal. See Figure 12.5 for an illustration. The reversal lasts for one period in case there is no autoregression in z_t ($\phi = 0$), but over a sequence of periods when there is positive autocorrelation in z_t ($\phi > 0$). The figure illustrates the latter case. In particular, notice how the response of long-run returns (the sum of 1-period returns $r_1^e + \dots + r_q^e$), is muted by the reversal. This will imply that the volatility of long-run returns is low. In contrast, $\theta > 0$ will cause movements in the same direction (momentum) and increase thus volatility, again see Figure 12.5.

Since the shocks are iid and uncorrelated, the expectation and variance of r_{t+1}^e , *conditional on the information available* at the time of investment in t , are straightforward to calculate (see the Appendix). Figure 12.6, which is roughly calibrated to monthly U.S. equity data although the autocorrelations are exaggerated to make a point, provides an illustration. In particular, it shows how the variance of q -period returns, conditional on the information available at the time of investment ($t = 0$ in the figure), scales with the investment horizon q when returns are iid, but slower when the model exhibits mean reversion ($\theta < 0$) and vice versa.

12.2.5 Mean-Variance Optimization with Autocorrelated Logarithmic Returns

Autocorrelation can affect both the expectations and the uncertainty of future returns. However, the analysis here will focus on how the uncertainty depends on the investment horizon, disregarding the “market timing” issue. That is, we assume a neutral *initial* state, $z_t = 0$, but allow for future shocks to the state.

In general, positive autocorrelation will make the sum of returns, $z(q)$, have a variance that scales quicker than the return horizon q as shocks “build up” over time. The opposite holds for a negative autocorrelation.

Figure 12.6 suggests that (12.10) can replicate the iid case (12.9). But it also shows that with long-run reversal, uncertainty increases slower than the investment horizon, so

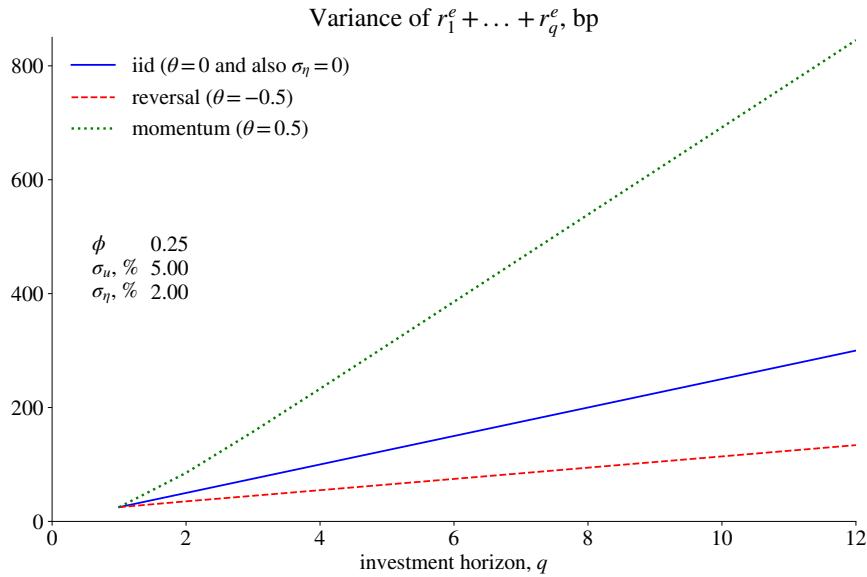


Figure 12.6: Variances of q -period return in the time series model (12.10)

equity is safer for a long-run investor. See Figure 12.7 for an illustration of the optimal portfolio weight on the risky asset, based on the same rough calibration as before as well as the same exaggeration of the autocorrelation. The key result is that the optimal weight on the risky asset increases with the investment horizon—if the returns show reversals (are negatively autocorrelated).

In summary, this analysis suggests that iid log returns are *not* sufficient to make equity relatively safer for a long-run investor—if we use a mean-variance approach to portfolio choice. Rather, it requires long-run reversals. Empirical evidence suggests that there might be some reversals, but not very much, questioning the notion of equity being safe in the long run.

Note, however, that the analysis in this chapter relies on the assumption that the investor makes *one* portfolio choice, irrespective of investment horizon. That is, no rebalancing. A later chapter will look at that issue in more detail as well as discuss the optimal response to differences in the initial state.

12.2.6 Utility Based Portfolio Choice

To study whether the conclusions from the MV approach are robust, this section considers utility based portfolio choice. For instance, a *logarithmic utility function* means setting $k = 0$ in (12.7)–(12.8).

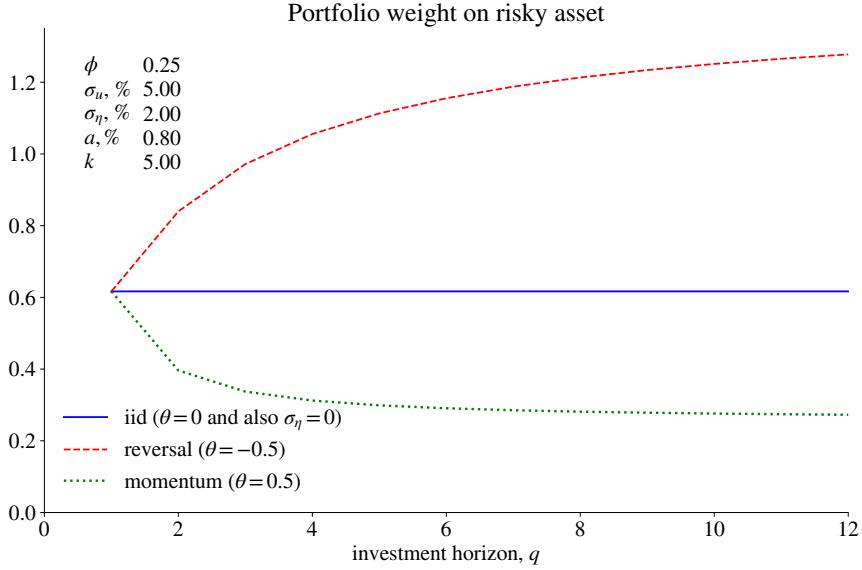


Figure 12.7: Portfolio weight on risky asset based on the time series model (12.10)

For a CRRA utility function and normally distributed log portfolio returns, we know (from an earlier chapter) that maximizing $E(1+Z_p)^{1-\gamma}/(1-\gamma)$ is equivalent to maximizing

$$E z_p - (\gamma - 1) \text{Var}(z_p)/2, \quad (12.11)$$

which is once again of the same form as (12.7)–(12.8), but with $k = \gamma - 1$. (See the chapter on utility theory for a proof.)

Both these examples lend support to the conclusion from the MV approach: to make equity safer for a long-run investor, returns must show reversals, so the asset price level is mean-reverting.

12.3 Appendix – The Conditional Variances*

First, simulate the system by setting $u_t = 1$ and all other shock to zero. Trace out the dynamic effects on r_t, r_{t+1}, \dots . Let the first element in ψ_0 be the effect on r_t , the first element in ψ_1 be the effect on r_{t+1} . Now, instead set $\eta_t = 1$ and trace of the dynamic effects to fill the second elements in the ψ vectors. We thus have the MA representation

of a return in period t

$$r_t = \psi'_0 \varepsilon_t + \psi'_1 \varepsilon_{t-1} + \psi'_2 \varepsilon_{t-2} + \dots, \text{ where } \varepsilon_t = \begin{bmatrix} u_t \\ \eta_t \end{bmatrix}.$$

Second, notice that the innovations of future returns, compared to the information available in t , can be written

$$r_{t+1} = \begin{bmatrix} \mathbf{0}' \varepsilon_{t+3} \\ \mathbf{0}' \varepsilon_{t+2} \\ \psi'_0 \varepsilon_{t+1} \end{bmatrix}, r_{t+2} = \begin{bmatrix} \mathbf{0}' \varepsilon_{t+3} \\ \psi'_0 \varepsilon_{t+2} \\ \psi'_1 \varepsilon_{t+1} \end{bmatrix}, r_{t+3} = \begin{bmatrix} \psi'_0 \varepsilon_{t+3} \\ \psi'_1 \varepsilon_{t+2} \\ \psi'_2 \varepsilon_{t+1} \end{bmatrix}.$$

Note that shocks in t or earlier are known in t , so they do not enter the expressions. Also, future shocks cannot affect current returns, which explains the zeros.

Third, since ε_{t+1} is uncorrelated across time, variances and covariances are straightforward to calculate, for instance,

$$\text{Cov}_t(r_{t+2}, r_{t+3}) = \psi'_0 \Omega \psi_1 + \psi'_1 \Omega \psi_2, \text{ where } \Omega = \text{Var}(\varepsilon_t).$$

Fourth, once we have the variance-covariance matrix of $(r_{t+1}, r_{t+2}, r_{t+3})$, the variance of $r_{t+1} + r_{t+2} + r_{t+3}$ can be calculated as the sum of all the elements.

Chapter 13

Dynamic Portfolio Choice

This chapter discusses portfolio choice of a long-run investor who can rebalance in each period. This means that the investor may form a portfolio that also hedges (predictable) future changes of the investment opportunity set.

13.1 Logarithmic Utility

Let the objective in period t be to maximize the expected log wealth in some future period

$$\max E_t \ln W_{t+q} = \max(\ln W_t + E_t r_{p,t+1} + E_t r_{p,t+2} + \dots + E_t r_{p,t+q}), \quad (13.1)$$

where r_{pt} is the log portfolio return, $r_{pt} = \ln(1 + R_{pt})$ with R_{pt} being the net portfolio return. The investor can rebalance the portfolio weights every period.

Remark 13.1 (*The Kelly criterion**) *The portfolio that solves (13.1) is said to be “growth optimal” as it maximizes the expected growth of the portfolio, also known as the Kelly criterion. It can be noted that this portfolio also maximizes the geometric mean return. To see this, recall from an earlier chapter that the geometric mean return is an increasing function of the average log return. Maximizing one of them means maximizing the other.*

Since the returns in the different periods enter separably in the utility function, the best an investor can do in period t is to choose a portfolio that maximizes $E_t r_{p,t+1}$. That is, to choose the one-period growth-optimal portfolio. This *myopic* approach is thus the optimal *dynamic* portfolio choice. Notice that the investment horizon q does not matter: short-run and long-run investors choose the same portfolio. This is specific to the logarithmic utility function.

However, the portfolio choice may change over time (t), if the distribution of the returns changes; that is, when returns are *not iid*, but this is unrelated to the investment horizon.

To solve the optimization problem, we approximate the *log* portfolio return, $r_p = \ln(1 + R_p)$, as in Campbell and Viceira (2002). (An earlier chapter includes a proof and an application.)

Remark 13.2 (*Approximate log portfolio return*) *The log portfolio return, $\ln(1+)R_p = \ln(1 + v'R + (1 - v'\mathbf{1})R_f)$, is approximately*

$$r_{pt} \approx r_{ft} + v'(r_t - r_{ft}) + v' \text{diag}(\Sigma)/2 - v'\Sigma v/2, \quad (13.2)$$

where Σ is the $n \times n$ variance-covariance matrix of r_t and $\text{diag}(\Sigma)$ is the n -vector of the variances (that is, the diagonal elements of Σ). With a single risky asset, this can be simplified as

$$r_{pt} \approx r_{ft} + v(r_t - r_{ft}) + v\sigma^2/2 - v^2\sigma^2/2, \quad (13.3)$$

where σ^2 is variance of r_t .

Using the approximation (13.2) and maximizing $E_t r_{p,t+1}$ gives the optimal n -vector of portfolio weights as

$$v = \Sigma^{-1}(\mu^e + \text{diag}(\Sigma)/2), \quad (13.4)$$

where μ^e is the vector of excess log returns of the risky assets, Σ is their variance-covariance matrix and $\text{diag}(\Sigma)$ picks out the diagonal of Σ , that is, the vector of variances. (The proof is at the end of the section.) The weight on the risk-free asset is the remainder, $1 - v'\mathbf{1}$. The case of a single risky asset was solved in an earlier chapter, yielding $v = \mu^e/\sigma^2 + 1/2$.

Clearly, the portfolio weights v change over time if the expected excess returns and/or the variance-covariance matrix change; that is, when returns are not iid. We could think of this as a *managed portfolio*.

Proof of (13.4). From (13.2) we have that the objective function can be written $r_f + v'\mu^e + v' \text{diag}(\Sigma)/2 - v'\Sigma v/2$, so the first order conditions are $\mu^e + \text{diag}(\Sigma)/2 - \Sigma v = \mathbf{0}_{n \times 1}$, which gives (13.4). \square

Example 13.3 (*One risky asset*) Suppose there is one risky asset with $\sigma = 5\%$, and the expected excess returns are different the two “scenarios” A and B: $\mu_A^e = 0.8\%$ or $\mu_B^e = 0.2\%$. Then (13.4) gives $v_A = 3.7$ and $v_B = 1.3$ in the two scenarios.

Example 13.4 (Three risky assets) Suppose we have three assets with the variance-covariance matrix (which is the same in both states)

$$\Sigma = \begin{bmatrix} 83 & 17 & 29 \\ 17 & 32 & 2 \\ 29 & 2 & 50 \end{bmatrix} bp,$$

and the means (in scenario A and B, respectively)

$$\mu_A^e = \begin{bmatrix} 0.8 \\ 0.9 \\ 0.3 \end{bmatrix} \% \text{ and } \mu_B^e = \begin{bmatrix} 0.4 \\ 0.45 \\ 0.15 \end{bmatrix} %,$$

In this case, the portfolio weights in the two states are

$$v_A \approx \begin{bmatrix} 0.65 \\ 2.93 \\ 0.60 \end{bmatrix} \text{ and } v_B \approx \begin{bmatrix} 0.49 \\ 1.62 \\ 0.45 \end{bmatrix}.$$

Empirical Example 13.5 Figure 13.1 illustrates mean returns and standard deviations, estimated by exponentially weighted moving averages. Figure 13.2 shows how the optimal portfolio weights change. It is clear that the portfolio weights can be fairly extreme and also change a lot—perhaps too much to be realistic.

13.2 CRRA Utility

The previous section has shown that logarithmic utility leads to myopic behaviour where the optimal portfolio depends only on beliefs about the next-period return. This clearly simplifies the choice, but it is unclear if logarithmic utility is a good representation of preferences. We therefore extend the analysis to the general constant relative risk aversion (CRRA) case.

As a benchmark, recall that an earlier chapter has established that if the log portfolio return, $r_p = \ln(1 + R_p)$, is normally distributed, then maximizing $E(1 + R_p)^{1-\gamma}/(1 - \gamma)$ is equivalent to maximizing

$$E r_p + (1 - \gamma) \text{Var}(r_p)/2. \quad (13.5)$$

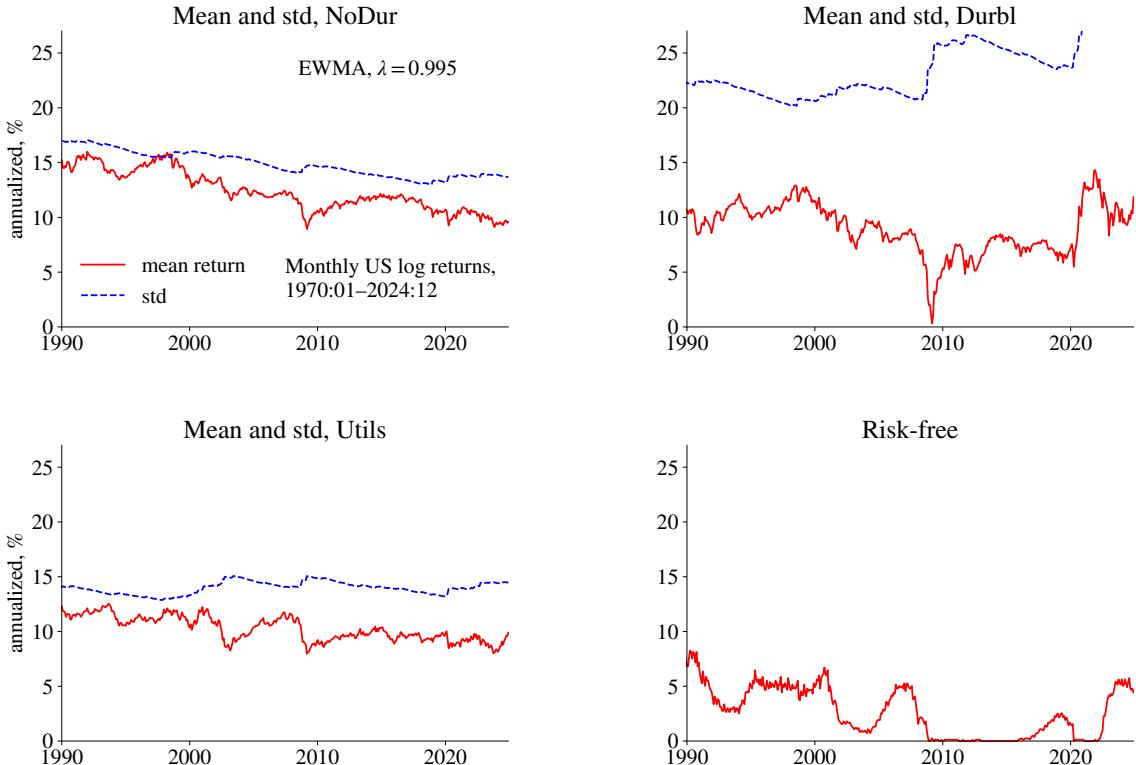


Figure 13.1: Dynamically updated estimates, 5 U.S. industries

Note that this is a one-period (myopic) optimum, since the utility function only depends on the return in the next periods. A dynamic optimum is discussed later on.

Using the approximation (13.2) gives optimal portfolio weights for the one-period (myopic) case as

$$v = \Sigma^{-1}(\mu^e + \text{diag}(\Sigma)/2)/\gamma. \quad (13.6)$$

These are the same weights as from the log utility case, but now divided by the risk aversion γ .

Proof of (13.6). From (13.2) we have that the objective function can be written $r_f + v'\mu^e + v'\text{diag}(\Sigma)/2 - v'\Sigma v/2 + (1-\gamma)v'\Sigma v/2$, so the first order conditions are $\mu^e + \text{diag}(\Sigma)/2 - \gamma\Sigma v = \mathbf{0}_{n \times 1}$, which gives (13.6). \square

Example 13.6 (One risky asset) Using the same figures as in Example 13.3 and $\gamma = 6$ gives $v_A = 0.62$ and $v_B = 0.22$.

Example 13.7 (Three risky assets) Using the same figures as in Example 13.4 and $\gamma = 6$

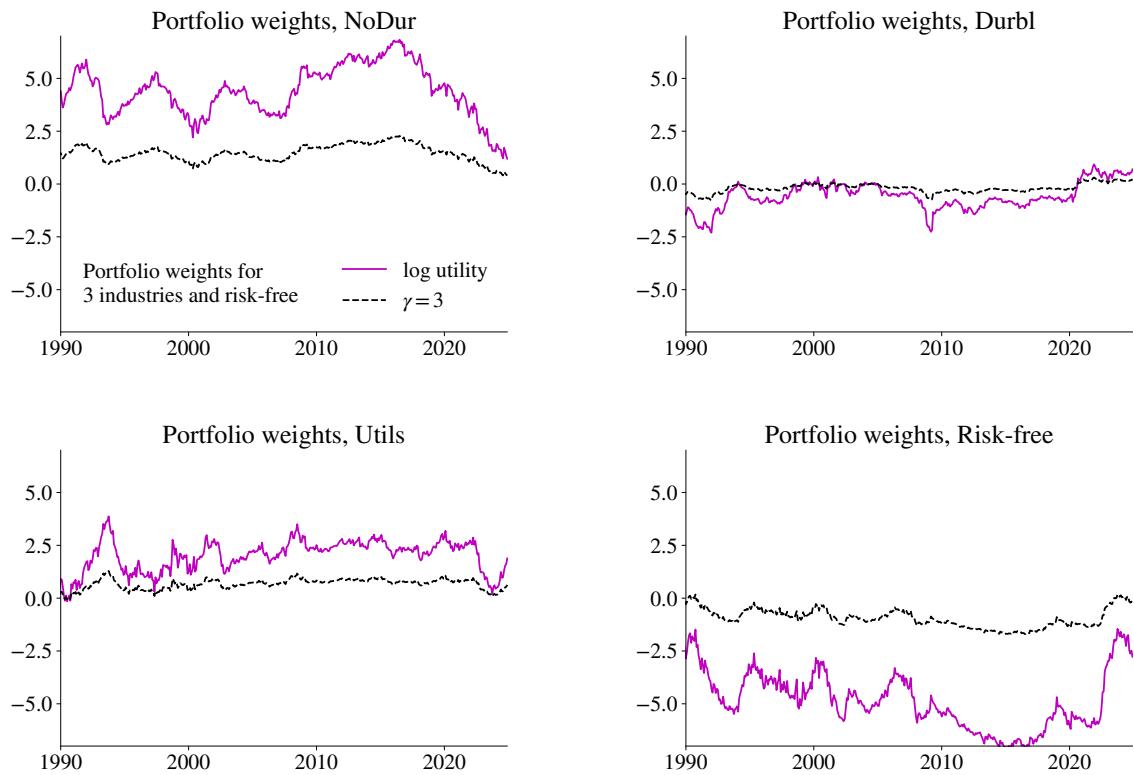


Figure 13.2: Dynamically updated portfolio weights, T-bill and 5 U.S. industries

gives

$$v_A \approx \begin{bmatrix} 0.11 \\ 0.49 \\ 0.10 \end{bmatrix} \text{ and } v_B \approx \begin{bmatrix} 0.08 \\ 0.27 \\ 0.07 \end{bmatrix}.$$

Empirical Example 13.8 See Figure 13.2 for a comparison of the solutions from log utility and from CRRA with $\gamma = 3$. The changes are more muted with the higher risk aversion.

13.3 Intertemporal Hedging

The combination of a CRRA utility function (with $\gamma \neq 1$) and non-iid returns makes the portfolio choice more challenging. If there is a link between returns in different periods, then a long-run investor might want to take this into account as it provides “diversification” across periods. This is *intertemporal hedging*. For instance, if some asset return (in $t + 1$) is negatively correlated with the investment outlook between $t + 1$ and $t + 2$, then that

asset is (in t) seen as providing a hedge.

We illustrate this below by using a simple model, although there are a few steps involved in solving it. (See Merton (1973a) and Campbell and Viceira (1999) for more elaborate approaches.) We also compare (in several figures) with myopic and static (multi-period) investment to highlight the differences.

13.3.1 An Autocorrelated Return Process

We use a simple time series model which encompasses both the iid case as well as long-run reversal or momentum

$$\begin{bmatrix} r_{t+1}^e \\ z_{t+1} \end{bmatrix} = \begin{bmatrix} a \\ \mathbf{0}_n \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{n \times n} & I_n \\ \mathbf{0}_{n \times n} & \phi \end{bmatrix} \begin{bmatrix} r_t^e \\ z_t \end{bmatrix} + \begin{bmatrix} I_n & \mathbf{0}_{n \times n} \\ \theta & I_n \end{bmatrix} \begin{bmatrix} u_{t+1} \\ \eta_{t+1} \end{bmatrix}. \quad (13.7)$$

where r^e and z (and thus also u and η) are n -vectors. We assume that u_{t+1} is uncorrelated with η_{t+1} , but there may be correlations within each vector. The respective variance-covariance matrices are Σ_{uu} and $\Sigma_{\eta\eta}$. Note that a is an n -vector and that ϕ and θ are both $n \times n$ matrices. (An earlier chapter used a similar model for the case of $n = 1$.)

13.3.2 Myopic Portfolio Choice

A *myopic investor* maximizes (13.5), which gives the same solution for the portfolio weights as in (13.6), but with

$$\mu^e = a + z_t \text{ and } \Sigma = \Sigma_{uu}. \quad (13.8)$$

Notice that both the expectation and variance are conditional on the information available at the time of the portfolio choice (t).

Figures 13.3–13.4 indicate the myopic portfolio weights with dots, mostly to make a comparison with the other cases discussed in more detail below.

Proof of (13.8). Notice that $r_{t+1}^e = a + z_t + u_{t+1}$. This immediately gives $E_t r_{t+1}^e = a + z_t$ and $\text{Var}_{t+1}(r_{t+1}^e) = \Sigma_{uu}$. \square

13.3.3 A Two-Period Investor (No Rebalancing)

We now consider a two-period investor who does not rebalance. This investor also maximizes (13.5), but the expectation and variance are for a 2-period return, so the $E r_p$

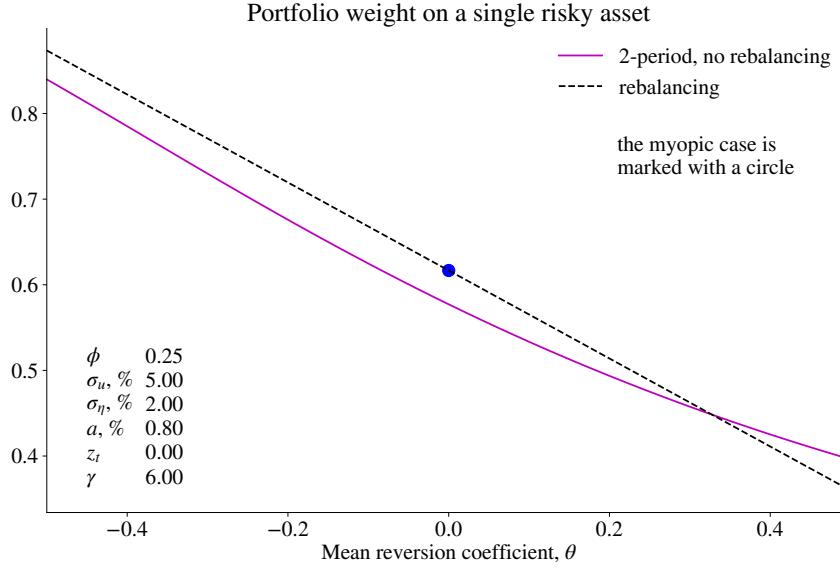


Figure 13.3: Weight on a single risky asset, two-period investor with CRRA utility for three cases: myopic, 2-period optimization without rebalancing and also with rebalancing. The return process is a scalar version of (13.7).

and $\text{Var}(r_p)$ terms should (respectively) be interpreted as

$$\mathbb{E}_t r_{p,t+1}^e + \mathbb{E}_t r_{p,t+2}^e \text{ and} \quad (13.9)$$

$$\text{Var}_t(r_{p,t+1}) + \text{Var}_t(r_{p,t+2}) + 2 \text{Cov}_t(r_{p,t+1}, r_{p,t+2}). \quad (13.10)$$

(The second line defines $\text{Var}_t(r_{p,t+1} + r_{p,t+2})$.) This involves, among other things, the covariance of the returns in the two periods, which is different from the myopic case. Intuitively, assets with reversals (negative autocorrelation) are less risky.

Again, we get the same solution for the portfolio weights as in (13.6), but with

$$\mu^e = 2a + (I + \phi)z_t \quad (13.11)$$

$$\Sigma = 2\Sigma_{uu}(I + \theta') + \theta\Sigma_{uu}\theta' + \Sigma_{\eta\eta}. \quad (13.12)$$

(See below for a proof.) When returns are iid ($\theta = \mathbf{0}, \Sigma_{\eta\eta} = \mathbf{0}$), then these expressions are two times those for the myopic case (13.8).

See Figure 13.3 for an illustration of how the portfolio weight on a single risky asset depends on the degree of reversal (θ , on the horizontal axis). In particular, with strong reversal ($\theta < 0$), equity is relatively safe for a long-run investor, which increases the portfolio weight. This is driven by the $\Sigma_{uu}\theta'$ term in (13.12), which captures the

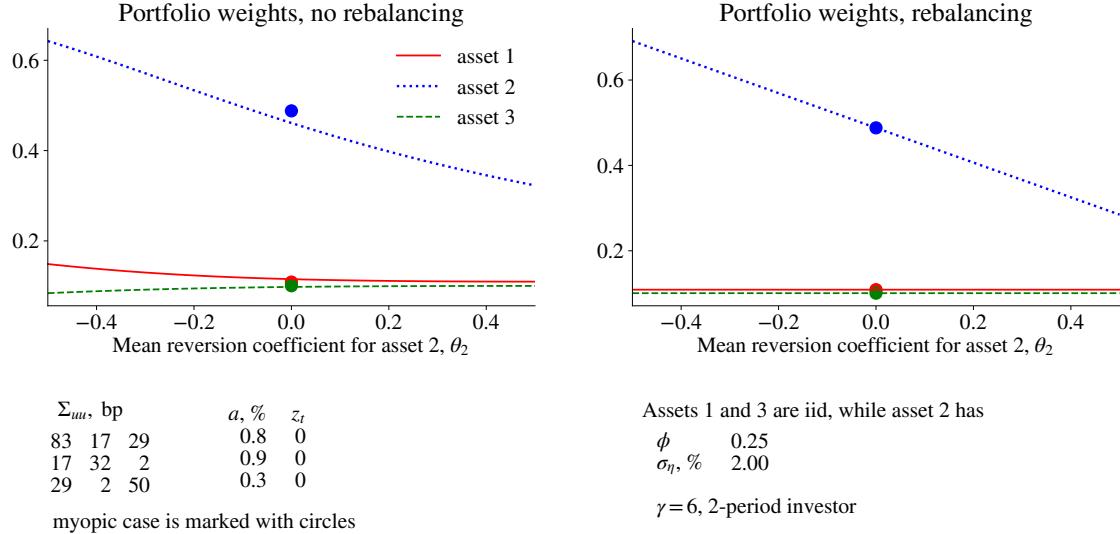


Figure 13.4: Weights on three risky asset, two-period investor with CRRA utility for three cases: myopic, 2-period optimization without rebalancing and also with rebalancing. The return process is a (13.7).

covariance of the returns in different periods. With reversal, this covariance is negative. (At $\theta = 0$ in the figure, the 2-period investor holds less risky assets than the myopic investor just because the volatility of z_{t+1} , $\Sigma_{\eta\eta} > 0$, adds uncertainty.) Notice that this figure assumes a neutral state ($z_t = 0$) in order to focus on the effect of risk.

Also, see Figure 13.4 for an illustration of a case with several risky assets (also assuming $z_z = 0$). It shows that the portfolio weight on the asset with non-iid returns (asset 2 in the figure) reacts strongly to variation in the reversal/momentum (here θ_2). Again, reversal makes the asset safer. There are some spillover effects on the other assets because of correlations.

Proof of (13.11)–(13.12). By combining the equations in (13.7) we can also write $r_{t+1}^e = a + z_t + u_{t+1}$ and $r_{t+2}^e = a + \phi z_t + \theta u_{t+1} + \eta_{t+1} + u_{t+2}$. The conditional moments are therefore (a) $E_t r_{t+1}^e = a + z_t$ and $\text{Var}_{t+1}(r_{t+1}^e) = \Sigma_{uu}$; (b) $E_t r_{t+2}^e = a + \phi z_t$, $\text{Var}_t(r_{t+2}^e) = \theta \Sigma_{uu} \theta' + \Sigma_{uu} + \Sigma_{\eta\eta}$; and (c) $\text{Cov}_t(r_{t+1}^e, r_{t+2}^e) = \Sigma_{uu} \theta'$. Combining gives (13.11)–(13.12). \square

13.3.4 Two-Period Investor (with Rebalancing)

We now consider the portfolio choice in t of an investor who will be able to rebalance in $t + 1$. She solves a similar problem to that of an investor who cannot rebalance, where

(13.9)–(13.10) define the relevant terms. However, in this case, $E_t r_{p,t+2}$ and $\text{Var}_t(r_{p,t+2})$ does not depend on the portfolio choice made in t (rather those in $t + 1$). In addition, $\text{Cov}_t(r_{p,t+1}, r_{p,t+2})$ depends on both the portfolio choices in t (through $r_{p,t+1}$) and in $t + 1$ (through $r_{p,t+2}$).

In principle, the covariance term is

$$\text{Cov}_t(r_{p,t+1}, r_{p,t+2}) = v'_t \text{Cov}_t(r_{t+1}^e, r_{t+2}^e v_{t+1}), \quad (13.13)$$

where v_t is the portfolio choice in t and v_{t+1} is the portfolio choice in $t + 1$. The latter is a 1-period choice made in $t + 1$, since that is the last time this 2-period investor makes an investment. Clearly, those weights are not known in t , so we apply an *approximation* by replacing v_{t+1} by its expected value obtained from the myopic case.

The optimal portfolio weights are as in (13.6), using

$$\mu^e = a + z_t + (1 - \gamma) \Sigma_{uu} \theta' E_t v_{t+1} \text{ and } \Sigma = \Sigma_{uu}. \quad (13.14)$$

(See below for a proof.) This version of “ μ^e ” captures both the expected return and the covariance term. The expression for $E_t v_{t+1}$ is easily calculated as

$$E_t v_{t+1} = \Sigma_{uu}^{-1} (a + \phi z_t + \text{diag}(\Sigma_{uu})/2)/\gamma, \quad (13.15)$$

which follows directly from (13.6) and (13.8).

The solution (13.14)–(13.15) equals the myopic portfolio in two cases: (1) when $\gamma = 1$ (log utility); and/or when (2) $\theta = \mathbf{0}$ (no reversal or momentum).

See Figure 13.3 for an illustration of the case of a scalar risky return. The general pattern is similar to the case without rebalancing. This holds also for the multi-asset case in Figure 13.4.

Proof of (13.14). Since $r_{t+1}^e = a + z_t + u_{t+1}$ and $r_{t+2}^e = a + \phi z_t + \theta u_{t+1} + \eta_{t+1} + u_{t+2}$, the covariance is

$$\text{Cov}_t(r_{p,t+1}, r_{p,t+2}) \approx v'_{t+1} \Sigma_{uu} \theta' E_t v_{t+1}.$$

We also immediately get $E_t r_{t+1}^e = a + z_t$ and $\text{Var}_{t+1}(r_{t+1}^e) = \Sigma_{uu}$. Combining gives (13.14). \square

13.3.5 Summary

The analysis in this section has shown that the optimal portfolio choice for a long-run (here, two-period) investor may differ substantially from that of a one-period investor if

returns are non-iid. In particular, assets with long-run reversals are “safe” for a long run investor. This holds irrespective of whether the investor rebalances or not, although the mechanisms differ somewhat.

Chapter 14

Foreign Exchange

14.1 Investing in Foreign Currency

14.1.1 The Return from Holding Currency

Investing in a foreign currency typically means that you buy that currency, lend on the foreign money market and eventually buy back domestic currency. To define the return, let S_t be today's price, measured in domestic currency, of one unit of foreign currency, referred to as the “asset.” Also, let R_{ft}^* be the foreign risk-free rate between $t - 1$ and t . The *return*, measured in domestic currency, is then

$$R_t = (1 + R_{ft}^*) \frac{S_t}{S_{t-1}} - 1. \quad (14.1)$$

Remark 14.1 (**Details of the currency return R_t*) In $t - 1$, invest S_{t-1} (of domestic currency) to buy one unit of foreign currency and lend it on the foreign money market. After one period you have $1 + R_f^*$ units of foreign currency, which buys $(1 + R_f^*)S_t$ units of domestic currency (this is the payoff). The gross return is payoff/investment, which is $(1 + R_f^*)S_t / S_{t-1}$. See Figure 14.1.

The return of the foreign investment *in excess of the domestic risk-free rate* is then

$$R_t^e = (1 + R_{ft}^*) \frac{S_t}{S_{t-1}} - (1 + R_{ft}). \quad (14.2)$$

Clearly, an appreciation of the foreign currency (or a depreciation of the domestic currency), along with a high foreign and low domestic risk-free rate, positively impacts the return. See Figure 14.2.

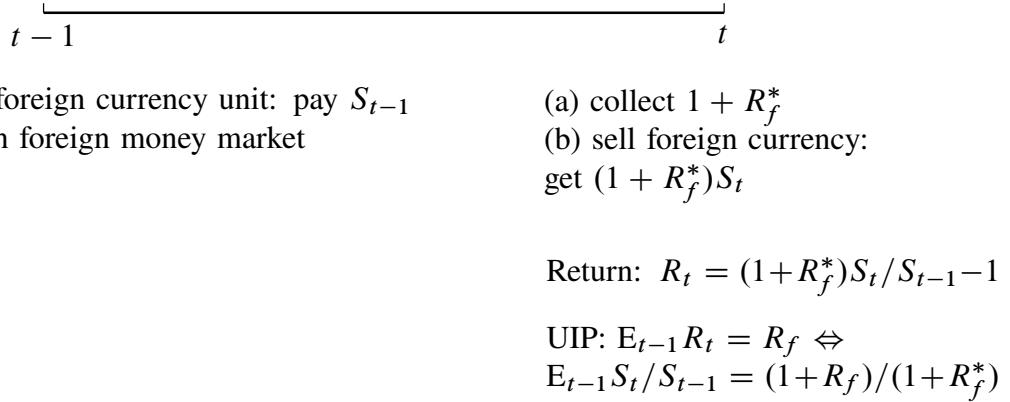


Figure 14.1: Return on currency investment

Example 14.2 With $(S_{t-1}, S_t, R_f^*, R_f) = (1.20, 1.25, 0.06, 0.04)$

$$R_t^e = (1 + 0.06) \frac{1.25}{1.20} - (1 + 0.04) = 0.064.$$

With $S_t = 1.20$ the excess return is $0.06 - 0.04 = 0.02$. Instead with $S_t = 1.177$ the excess return is close to zero

$$R_t^e = (1 + 0.06) \frac{1.177}{1.20} - (1 + 0.04) \approx 0.$$

Remark 14.3 (*Indirect exchange rate quotation) If you instead work with exchange rate quotes that use the number of foreign currency units needed to buy one domestic currency unit, \tilde{S} , then replace S by $1/\tilde{S}$ in the previous equations.

In practice, *risk-free returns are from zero-coupon bonds* (bills). We can thus rewrite $1 + R_{ft}$ in terms of an interest as

$$1 + R_{ft} = (1 + Y_{t-1})^m, \quad (14.3)$$

where Y_{t-1} is an effective *interest rate* determined in $t - 1$ and m is the fraction of a year between date $t - 1$ and t , for instance, $m = 1/12$ for a month. Notice that the interest rate is dated $t - 1$, since we know already then how much we earn on the bond between $t - 1$ and t . For the foreign market, we have $1 + R_{ft}^* = (1 + Y_{t-1}^*)^m$. Using this in (14.2) gives the excess return on the foreign investment as

$$R_t^e = (1 + Y_{t-1}^*)^m \frac{S_t}{S_{t-1}} - (1 + Y_{t-1})^m. \quad (14.4)$$

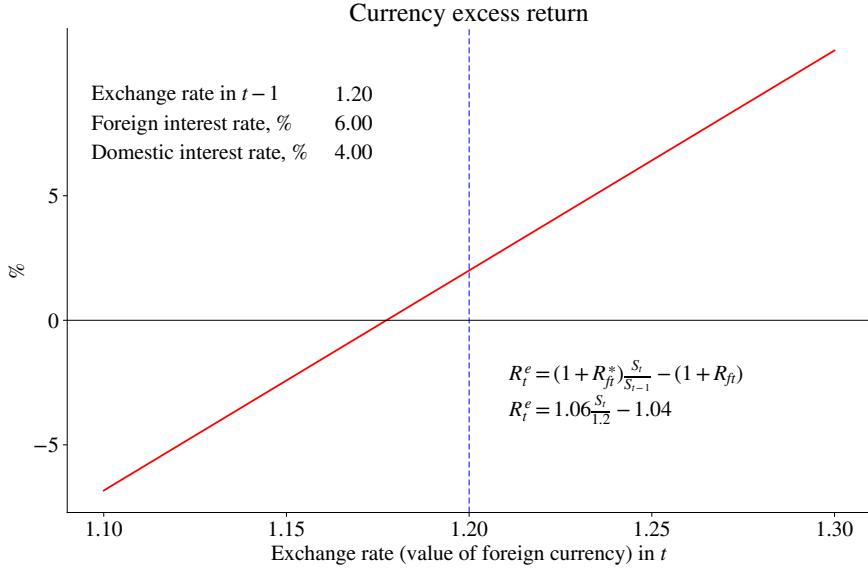


Figure 14.2: Illustration of currency excess returns as a function of the realized exchange rate

14.1.2 Covered Interest Rate Parity

To avoid arbitrage opportunities, the forward price in $t - 1$ for delivery of one unit of foreign asset in t must obey

$$F_{t-1} = \frac{1 + R_{ft}}{1 + R_{ft}^*} S_{t-1}. \quad (14.5)$$

This is an application of the spot-forward parity, which for the FX market is often called covered interest rate parity (CIP). In practice, there are (albeit small) deviations from CIP, depending on which interest rates (secured or unsecured?) that are used. The *forward premium*, $F_{t-1}/S_{t-1} - 1$, reflects the interest rate difference: a higher value means that the domestic interest rate is higher than the foreign. See Sercu (2009) for more details.

Example 14.4 Using the same numbers as in Example 14.2 we get

$$F_{t-1} = \frac{1 + 0.04}{1 + 0.06} \times 1.20 \approx 1.177.$$

Proof (of (14.5)) Replace the risky strategy in (14.2) by “locking in” the FX rate with a forward contract (replace S_t by F_{t-1}) to get $(1 + R_{ft}^*) \frac{F_t}{S_{t-1}} - (1 + R_{ft})$. This risk-free return must be zero, or else arbitrageurs step in. Rearrange as (14.5). \square

Remark 14.5 (Alternative expression of (14.2)) Use CIP to rewrite the excess return (14.2) as $R_t^e = (1 + R_{ft}) \frac{S_t}{F_{t-1}} - (1 + R_{ft})$. This is sometimes approximated by $S_t/F_{t-1} - 1$.

14.1.3 Uncovered Interest Rate Parity

The uncovered interest rate parity (UIP) says that the expected exchange rate ($E_{t-1} S_t$) must be such that the expected excess return from the currency speculation in (14.2) is zero

$$E_{t-1} R_t^e = 0. \quad (14.6)$$

This means that investing on the foreign money market (and then changing back to the domestic currency) has the *same expected returns* as investing on the domestic money market—in spite of having different risks. Notice that this is very different from CIP, which only rules out arbitrage opportunities and says nothing about expectations or risk. A somewhat more flexible form of UIP would add a *constant* risk premium to (14.6).

A zero expected excess return in (14.2) means that we must have

$$\frac{E_{t-1} S_t}{S_{t-1}} = \frac{1 + R_{ft}}{1 + R_{ft}^*}. \quad (14.7)$$

UIP thus says that the foreign currency is expected to appreciate ($E_{t-1} S_t / S_{t-1} > 1$) if the foreign interest rate is lower than the domestic. In this way, the foreign investment gains from the (expected) exchange rate movement, but loses from the interest rate—leaving the (expected) return the same as in the domestic market.

Example 14.6 (UIP) Using the same number as in Example 14.2, UIP says that

$$E_{t-1} S_t = 1.20 \times \frac{1 + 0.04}{1 + 0.06} = 1.177,$$

so the domestic currency is expected to appreciate.

Remark 14.7 (*UIP in terms of interest rates or forwards) Using the definition of the risk-free rate in (14.3) and CIP (14.5), we can rewrite UIP (14.6) as $E_{t-1} S_t = F_{t-1}$.

Empirical evidence is mixed but often reveals considerable deviations from UIP, potentially due to either (a) significant shifts in risk premia over time or (b) systematic disparities between expectations and historical exchange rate movements, including large surprises or even non-rational expectations. In the latter case, UIP might hold although ex post data says little about the market expectations.

Empirical Example 14.8 Tables 14.1–14.2 show regressions of exchange rate depreciations $S_t / S_{t-1} - 1$ on the lagged forward premium $F_{t-1} / S_{t-1} - 1$) currencies. The slope coefficients are mostly far from one (1), and even negative in several cases.

	AUD	CAD	EUR	JPY	NZD
forward premium	-1.55 (-1.24)	-1.50 (-0.93)	-2.39 (-1.83)	0.50 (0.53)	-0.27 (-0.21)
constant	-0.00 (-1.06)	-0.00 (-0.12)	0.00 (0.90)	-0.00 (-0.90)	-0.00 (-0.19)
R^2	0.01	0.00	0.01	0.00	0.00
obs	371	371	371	371	371

Table 14.1: Regressing 1-month depreciation $S/S_{t-1} - 1$ on forward premium 1 month earlier $F_{t-1}/S_{t-1} - 1$, 1994:01-2024:12. Numbers in parentheses are t-stats.

	NOK	SEK	CHF	GBP
forward premium	-0.60 (-0.55)	-1.35 (-1.30)	-0.84 (-0.67)	-0.79 (-0.54)
constant	-0.00 (-0.81)	-0.00 (-0.28)	0.00 (0.98)	-0.00 (-0.53)
R^2	0.00	0.00	0.00	0.00
obs	371	371	371	371

Table 14.2: Regressing the 1-month depreciation ($S/S_{t-1} - 1$) on the forward premium 1 month earlier ($F_{t-1}/S_{t-1} - 1$), 1994:01-2024:12. Numbers in parentheses are t-stats.

14.1.4 Carry Trade

A common FX strategy is to borrow a low interest rate currency (CHF and JPY?), buy a high interest rate currency (AUD?) and lend on its money market. This is called a *carry trade*. This is the same as selling (buying) currencies with low (high) forward premia.

This strategy has a positive return if the high (low) interest rate currency depreciates (appreciates) less than suggested by UIP, but clearly also carries the risk of the opposite happening. Empirical evidence suggests that carry trades have generated positive average returns, but are exposed to (intermittent) dramatic losses.

The excess return of a carry trade is given by R_t^e in (14.2). However, a carry trade need not borrow the domestic currency. For instance, a US investor could borrow JPY and lend AUD. Clearly, this strategy would benefit from an appreciation of the AUD and a depreciation of the JPY, as well from a high AUD interest rate and low JPY interest rate.

Empirical Example 14.9 Figures 14.3–14.4 illustrate the performance of a monthly carry trade implemented on 10 key currencies. The strategy performs much better than an

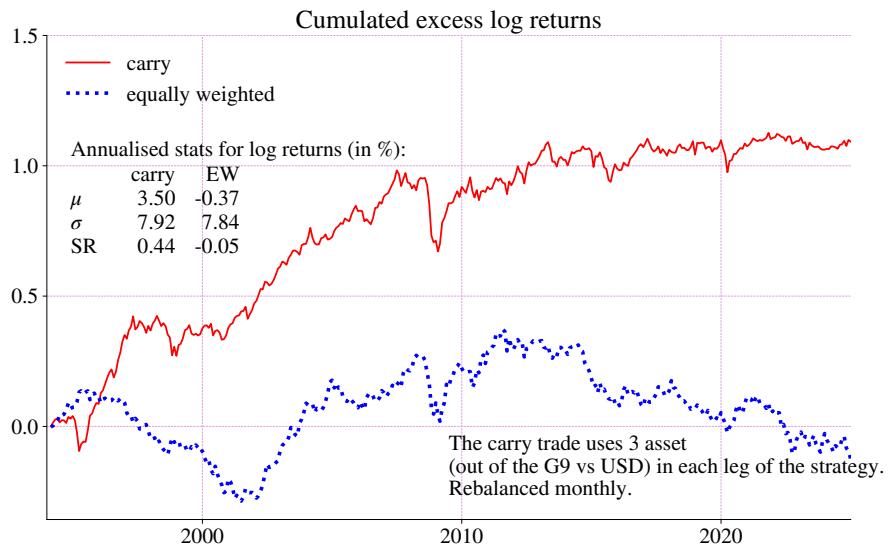


Figure 14.3: Return on currency investment

equally weighted investment in all currencies (financed by borrowing USD), but suffers in periods of high uncertainty (as measured by VIX).

14.2 Exchange Rate Quotation*

14.2.1 Direct and Indirect Quotation

An exchange rate is the price of one currency in terms of another currency. There are clearly *two ways of quoting* this price: the price (measured in domestic currency) of one unit of foreign currency (“direct quotation”), or the price (measured in foreign currency) of one unit of domestic currency (“indirect quotation”). A reasonably established set of quotations and symbols exists in the interbank market, but in other settings, either type of quotation is possible; one should always verify.

Example 14.10 As an example, Datastream/Refinitiv defaults to reporting “how many USD you pay for one AUD”, but “how many CAD you pay for one USD”.

Example 14.11 For a Swiss investor in 2014, a direct quotation meant that EUR 1 cost CHF 1.2 (“EUR 1 = CHF 1.2”), and an indirect quotation that CHF 1 cost EUR 0.8333 (sometimes written as “CHF 1 = EUR 0.8333”).

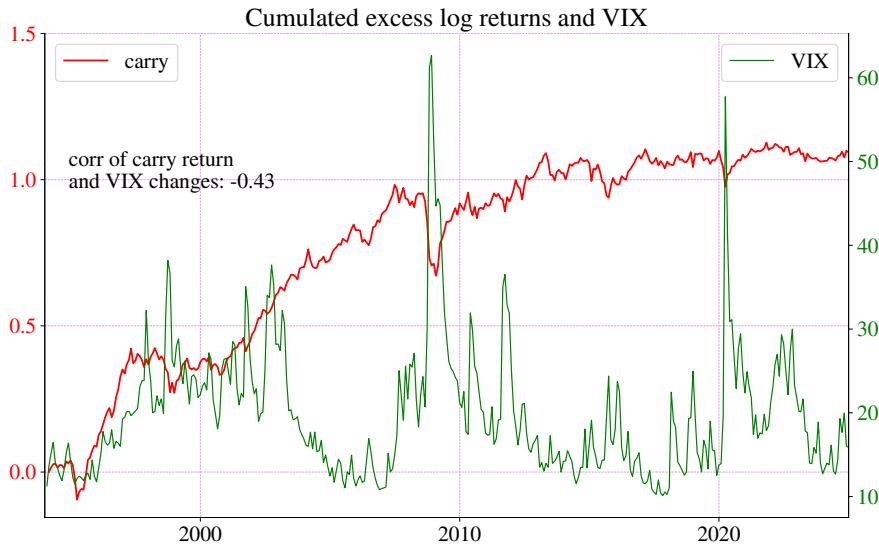


Figure 14.4: Return on currency investment, see Figure 14.3 for details

Remark 14.12 (*The meaning of CHF/EUR*) These lecture notes follow the convention that CHF/EUR (or $S^{CHF/EUR}$) denotes how many CHF you have to pay for each EUR, for instance, $S^{CHF/EUR} = 1.2$. Clearly, $S^{x/y} = 1/S^{y/x}$, for instance, $S^{EUR/CHF} = 0.8333$. (In contrast, the interbank FX market often use $EURCHF$ to denote the same thing, that is, how many CHF you pay for one EUR.)

Remark 14.13 (**Currency codes, according to ISO 4217*) USD , EUR , JPY , GBP , AUD , CAD , CHF , CNY (Chinese yuan), SEK (Swedish krona), MXN (Mexican peso).

14.2.2 Cross Rates*

Exchange rate across “smaller” currencies are often established indirectly and are therefore called “cross rates”: as a combination of two trades. For instance, suppose you own CHF and want to buy CAD (Canadian dollars). It may well be that this involves two trades: use the CHF to buy USD and then use the USD to buy CAD. (Even 15 years after the collapse of the Bretton-Woods system in the early 1970s almost all currency trades went via the USD. Since then there are more direct trades, but trade via the USD still dominates.)

Example 14.14 (*The implicit trade in a cross rate*) (a) Buy one USD, costs 0.95 CHF; (b) use the one USD to buy 1.25 CAD; (c) in total you have paid 0.95 CHF and got 1.25 CAD. Therefore, the implied price (in CHF) per AUD is $0.95/1.25 \approx 0.76$. (You can memorize this as “ $CHF/USD \times USD/CAD = CHF/CAD$ ”) See Figure 14.5 for an illustration.

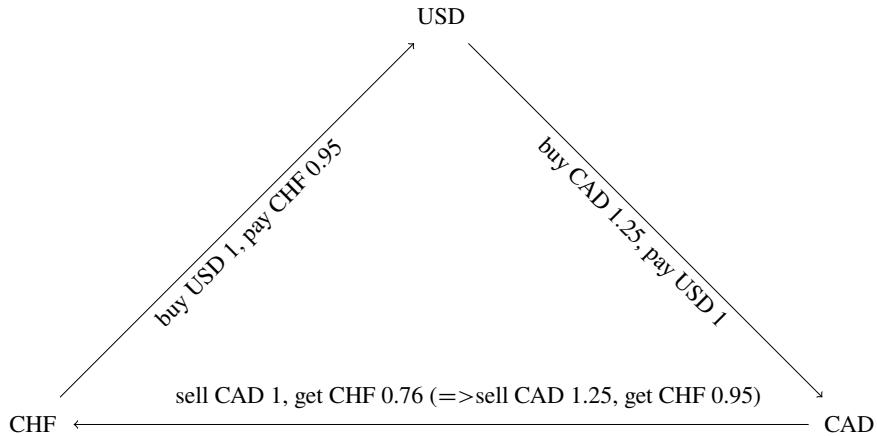


Figure 14.5: Cross-rates

Remark 14.15 (**The implicit trade in a cross rate, using $S^{x/y}$ notation*) In the previous example, $S^{CHF/USD} = 0.95$ and $S^{USD/CAD} = 1/1.25$, so $S^{CHF/USD} = S^{CHF/USD} S^{USD/CAD} = 0.95/1.25 \approx 0.76$. In general, cross rates mean that $S^{x/y} = S^{x/z} S^{z/y}$.

If there is a way to trade without going through another currency (and there typically is), then the price on this market should be very close to the cross rate. If not, there would be an arbitrage opportunity.

14.2.3 Log Rates*

A fair amount of exchange rate analysis is done in terms of log rates. For that reason, this section summarizes how the previous expressions look like in logarithmic terms.

Remark 14.16 (Log FX returns, (14.1)–(14.2)) Let r_t be the log return, $\ln(1 + R_t)$. From (14.1), it can be written $r_t = r_{ft}^* + \Delta s_t$, where r_{ft}^* is the log foreign gross risk-free rate, $\ln(1 + R_{ft}^*)$, and Δs_t is the relative change of the exchange rate, $\ln S_t / S_{t-1}$. Subtract $r_{ft} = \ln(1 + R_{ft})$ to get the excess log return $r_t^e = \Delta s_t + r_{ft}^* - r_{ft}$, which is the log version of (14.2).

Remark 14.17 (Log FX returns, (14.4)) Equation (14.3) can be used to rewrite the excess log return in Remark 14.16 as $r_t^e = \Delta s_t + m(y_{t-1}^* - y_{t-1})$, where $y = \ln(1 + Y)$. This is the log version of (14.4).

Remark 14.18 (*Log FX returns*) Take logs of (14.5), rearrange and use in the excess log return in Remark 14.16 to get $r_t^e = \Delta s_t - (f_{t-1} - s_{t-1}) = s_t - f_{t-1}$, which is the log version of the result in Remark 14.5. Also, the interest rate differential can be written $r_{ft}^* - r_{ft} = s_{t-1} - f_{t-1}$.

14.3 Currency Risk in Foreign Investments

We now consider an investment in a *risky foreign asset*. The definition of the return is similar to (14.1), except that we replace the safe foreign return (R_{ft}^*) with a risky foreign return (R_t^*). This means that the foreign investment contributes to the uncertainty about the total return via both the uncertainty in R_t^* and its covariance with the exchange rate movements.

The gross return (measured in domestic currency) of this investment is

$$1 + R_t = (1 + R_t^*) \frac{S_t}{S_{t-1}}. \quad (14.8)$$

Take logs to get the log return

$$r_t = r_t^* + \Delta s_t, \quad (14.9)$$

where r_t^* is the log foreign return, $\ln(1 + R_t^*)$, and Δs_t is the change of the log exchange rate, $\ln(S_t/S_{t-1})$. Notice that our investor gains if the (a) foreign asset (equity?) increases in value ($r_t^* > 0$) and (b) if the foreign currency increases in value (appreciates) relative to the domestic currency ($\Delta s_t > 0$). See Figure 14.6 for an empirical illustration.

Example 14.19 (*Investing abroad*) Consider a US investor buying British equity in period $t-1$: 5.5 GBP per British share \times 1.6 USD per GBP = 8.8 USD, and selling in t : 5.1 GBP per British share \times 1.9 USD per GBP = 9.69 USD. The gross return for the US investor (in USD) is $1 + R = (1 - 0.073) \times (1 + 0.188) = 1.10$. Taking logs gives $\ln(1 + R) = 0.096$.

From (14.9) the mean and variance of the log return are

$$\mathbb{E} r_t = \mathbb{E} r_t^* + \mathbb{E} \Delta s_t \quad (14.10)$$

$$\text{Var}(r_t) = \text{Var}(r_t^*) + \text{Var}(\Delta s_t) + 2 \text{Cov}(r_t^*, \Delta s_t). \quad (14.11)$$

Notice that a negative covariance (the foreign local return is high at the same time as the foreign currency depreciates) may reduce the variance of the return measured in domestic currency. See Elton, Gruber, Brown, and Goetzmann (2014) 12 for more details.

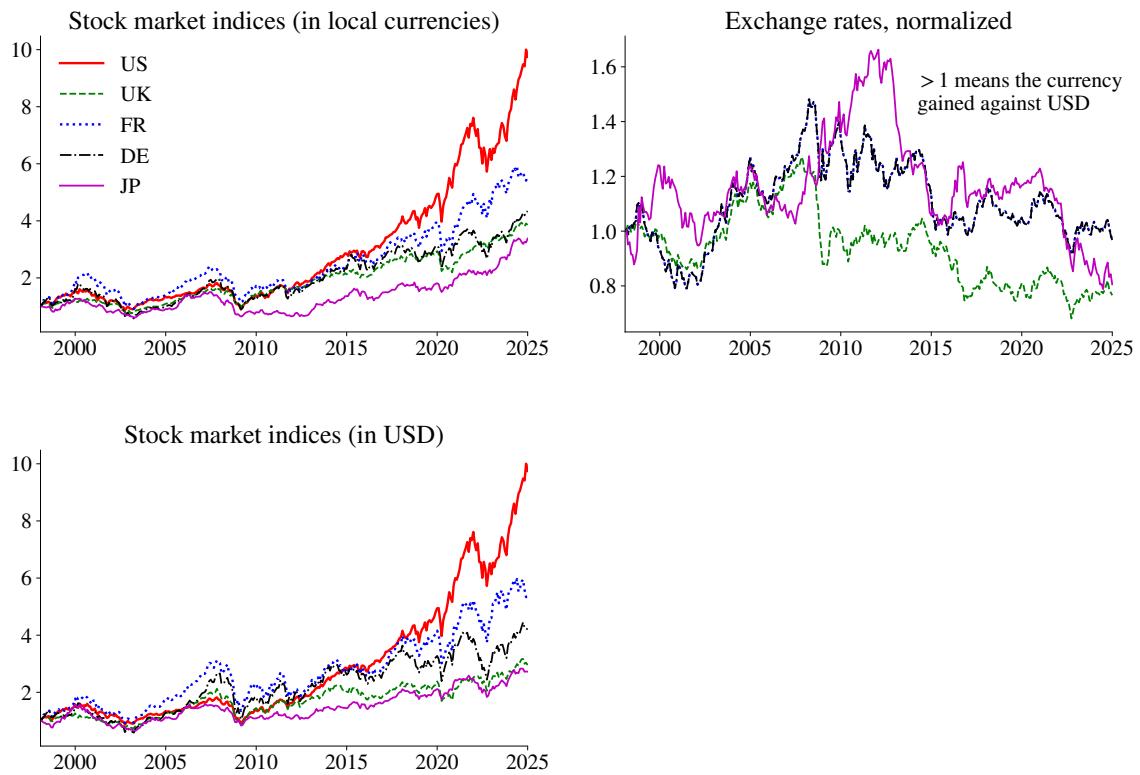


Figure 14.6: International stock market indices and exchange rates

Empirical Example 14.20 See Tables 14.3–14.4 and Figure 14.6 for an empirical illustration.

14.4 Hedging Exchange Rate Movements

International equity or bond investments often involve considerable exchange rate risk. It may be useful to hedge that risk. For instance, the investment strategy may be based on industry analysis (“pick promising pharma companies across the globe”), while the currency exposure is just unwanted risk which requires a different type of analysis—and exchange rates are notoriously difficult to predict. Unless the covariance is very negative (as discussed above) this may motivate hedging the currency exposure.

The most common ways of hedging the exchange rate risk involve forward and option contracts (mostly for short horizons) or swap contracts (longer horizons). Alternatively, a partial hedge is achieved by financing the investment by borrowing on the foreign market. In that way only the profit, not the entire investment, is exposed to exchange rate risk.

	Local currency	Exchange rate	in USD
US	8.5	0.0	8.5
UK	5.0	-1.0	4.0
FR	6.4	-0.1	6.2
DE	5.5	-0.1	5.3
JP	4.5	-0.8	3.7

Table 14.3: Contribution to the average (annualized, in %) log return for a US investor investing in different equity markets, 1998:01-2024:12

	Local currency	Exchange rate	2*Cov	in USD
US	2.5	0.0	0.0	2.5
UK	1.8	0.7	0.2	2.7
FR	3.1	0.9	0.3	4.3
DE	4.2	0.9	0.3	5.4
JP	3.0	1.1	-1.3	2.7

Table 14.4: Contribution to the variance (annualized, in %) of the log returns for a US investor investing in different equity markets, 1998:01-2024:12

To illustrate how a forward contract might help, suppose we could lock in the period t exchange rate by entering a forward contract in $t - 1$. If so, the return of the foreign investment (but measured in domestic currency) changes from (14.8) to

$$1 + R_t^{hedged} = (1 + R_t^*) \frac{F_{t-1}}{S_{t-1}}, \quad (14.12)$$

where the currency risk is eliminated.

The practical problem with (14.12), as mentioned before, is that the foreign return, R_t^* , typically is not known in $t - 1$, so we do not know how many units of currency to hedge via forward contracts. One possibility is to only hedge the investment, in which case the right hand side changes to $F_{t-1}/S_{t-1} + R_t^* S_t / S_{t-1}$ so only the foreign return is exposed to currency risk.

Remark 14.21 (*[\(14.12\)](#) when the foreign return is risk-free) Use the forward-spot parity [\(14.5\)](#) to substitute for the forward price in [\(14.12\)](#) to see that the hedged return then equals the domestic risk-free rate.

14.5 Explaining Exchange Rates

Economic models of exchange rates can be thought of as trying to understand the “fundamental” value of currencies (similar to valuing a company according to the discounted sum of future dividends). This section briefly summarizes some of these ideas. It should be noticed, however, that most models of exchange rates only have explanatory power over longer horizons (5–10 years or longer).

14.5.1 Purchasing Power Parity and the Real Exchange Rate?

The basic idea is that a product should *cost the same at home and abroad* (when measured in a common currency). If this is not the case, then (goods) arbitrage will take place, driving down demand for the currency of the more expensive country which leads to an depreciation of its currency.

The strong assumption about goods arbitrage can be relaxed by instead assuming that goods may differ across countries, but that the import/export demand is somewhat price elastic. The *real exchange rate* (the relative price of foreign and domestic goods, measured in the domestic currency) is often used as an indicator of the competitiveness of a country. If the domestic price is too high, then export will decrease and import will increase, leading to a trade deficit. This puts pressure on the exchange rate in the same way as discussed above. The mechanism is thus that the real exchange rate puts pressure on the (nominal) exchange rate.

Empirical tests strongly refutes this set of theories for price and exchange rate *levels*, but may work reasonably well for changes over the long run (10+ years). In particular, it points at the important link between inflation (which drives up prices) and depreciations, which is a well established fact over longer runs. In the short run, the causality seems to be the reverse: (nominal) exchange rate movements cause movements in the real exchange rate (competitiveness).

It is observed that price levels, when measured in a common currency, are higher in wealthier countries. Once we adjust for that, we get a better measure of over/under valuation of the currency.

14.5.2 Interest Rates?

The exchange rate often appreciates when the central bank raises the interest rate. This typically happens very quickly. One possible explanation is financial flows: if international

investors want to benefit from the higher interest rates, then they first need to buy the currency. However, if we were to believe in UIP then an investor needs to buy the currency before it has appreciated fully. Otherwise, the higher interest rate will be offset with a future depreciation. In short, the interest rate hike causes an immediate appreciation, followed by a slow depreciation.

Empirical tests suggests that high interest rate currencies can continue to appreciate for several years (this forms the basis for carry trades), but that they typically eventually suffer a sudden depreciation.

14.5.3 Transactions? (Business Cycles and Financial Flows)

The business cycle theory for exchange rates goes back to first principles to ask the question: why do we hold a currency (cash or cash-like assets)? After all, cash is typically not a good savings instrument (cash is eroded by inflation and there are typically better investment vehicles). Some cash is held because some people want to avoid banks (distrust of bank, fear of taxation and other legal issues), which seems to be an important driver of demand for large denomination bills. This may historically have had an effect on exchange rates, but less so today.

Instead, the key use of a currency is that it facilitates transactions, which suggests that both business cycle conditions (which drive the transaction volumes for goods and services) and financial flows are the most important factors behind exchange rates.

Empirical tests of these models find that also they have some explanatory power over longer horizons.

Chapter 15

Forwards and Futures

15.1 Derivatives

Remark 15.1 (*On the notation*) The notation is kept short. The current period is assumed to be $t = 0$ and the derivative expires in $t = m$, which means m years later. Time subscripts and indicators of time to maturity are typically suppressed, unless strictly needed in the context. For instance, instead of $F_0(m)$ we often use F denote the forward price (contracted in $t = 0$, expiring in $t = m$) and similarly for interest rates (y instead of $y_0(m)$). Also, instead of S_0 we use S , but we keep the subscript on S_m .

Derivatives are assets whose payoff depend on some underlying asset (for instance, the stock of a company). The most common derivatives are futures contracts (including forward contracts) and options. However, options sometimes depend not directly on the underlying asset, but on the price of a futures contract for the underlying. See Figure 15.1.

Derivatives are in zero net supply, so a contract must be issued (a short position) by someone for an investor to be able to buy it (long position). For that reason, gains and losses on derivatives markets sum to zero.

15.2 Present Value

The present value of Z units paid m periods (years) into the future is

$$PV(Z) = (1 + Y)^{-m} Z, \text{ or} \quad (15.1)$$

$$= e^{-my} Z, \quad (15.2)$$

where Y is effective spot interest rate on an m -period loan, and y is the continuously compounded m -period interest rate ($y = \ln(1 + Y)$). As usual, the *interest rates are*

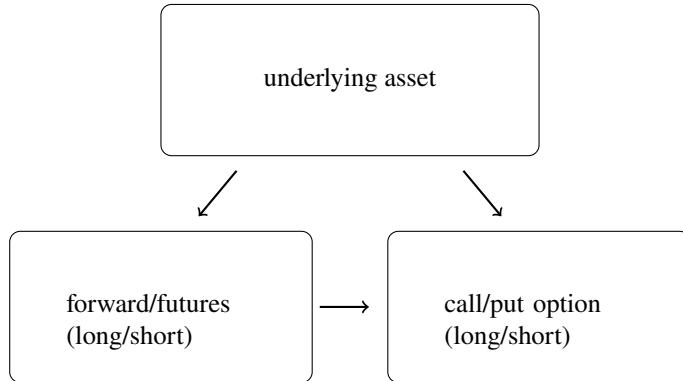


Figure 15.1: Derivatives on an underlying asset

expressed on an annual basis; hence, m should measure time in years. For instance, $m = 1/4$ means a quarter of a year (3 months).

Example 15.2 (Present value) With $y = 0.05$ and $m = 3/4$ we have the present value $e^{-3/4 \times 0.05} Z \approx 0.963Z$.

15.3 Forward Contracts

15.3.1 Definition of a Forward Contract

A forward contract specifies, among other details, the expiration date, which asset should be delivered, and the agreed payment for it, referred to as the forward price F . See Figure 15.2 for an illustration.

The profit (payoff) of a forward contract *at expiration* is straightforward to calculate. Let S_m be the price (on the spot market) of the underlying asset at expiration (in m). Then, for the *buyer* of a forward contract the

$$\text{payoff of a forward contract} = S_m - F. \quad (15.3)$$

The reason is that, at expiration, the owner of the forward contract pays F to get the asset which is worth S_m . See Figure 15.3 for an illustration of the payoff (at expiration) as a function of the underlying price, S_m . (The payoff function will look more interesting for options.) Similarly, the payoff for the *seller* (or issuer) of a forward contract is $F - S_m$ (she buys the asset on spot market for S_m , gets F for asset according to the contract). This sums to *zero*, irrespective of the value of the underlying asset.

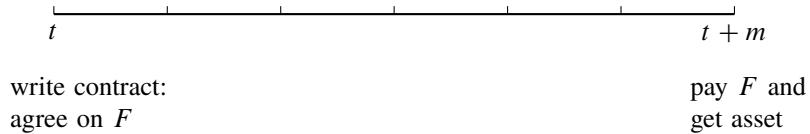


Figure 15.2: Timing convention of forward contract

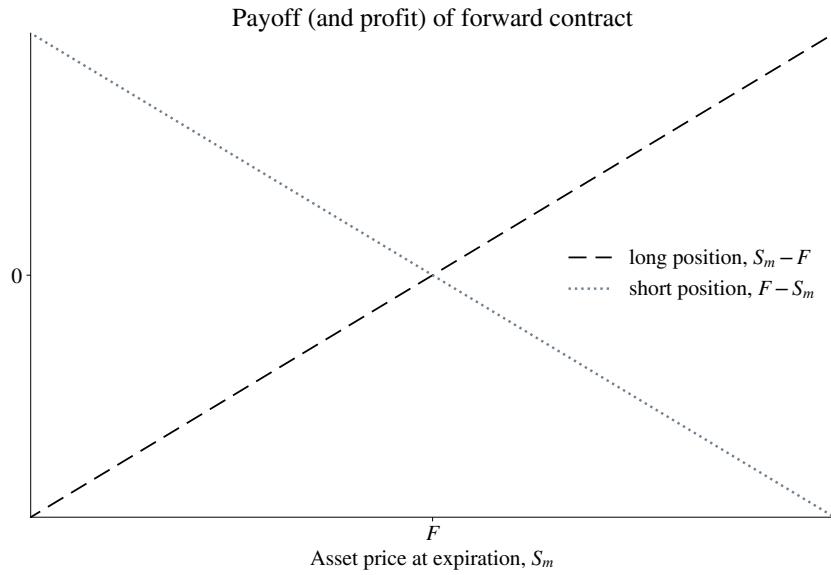


Figure 15.3: Profit (payoff) of forward contract at expiration

15.3.2 Forward-Spot Parity

A forward contract entails both a right (to get the underlying asset at expiration) and an obligation (to pay the forward price at expiration), so it is perhaps not obvious what the value of it is. However, in the absence of trading costs, a no-arbitrage argument shows that the following proposition must hold.

Proposition 15.3 (*Forward-spot parity, no dividends*) *The present value of the forward price, F , contracted in $t = 0$ (but to be paid in m) on an asset without dividends equals the spot price:*

$$e^{-my} F = S, \text{ so} \quad (15.4)$$

$$F = e^{my} S, \quad (15.5)$$

where S is the spot price in $t = 0$ and y is m -period spot interest rate.

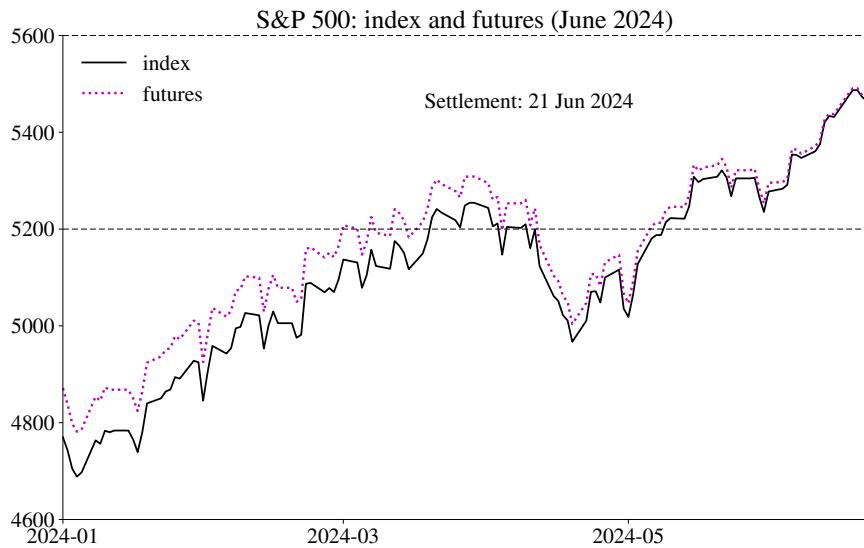


Figure 15.4: S&P 500 index level and futures

(If you prefer effective interest rates, then (15.5) reads $F = (1 + Y)^m S.$)

With a positive interest rate, the forward price is higher than today's underlying price. The intuition is that the forward contract is like buying the underlying asset on credit: $e^{-my} F$ can be thought of as a *prepaid forward contract*. It is worth the same as the underlying, if there are no dividends (to be discussed below).

The forward-spot parity is the same as that a “covered strategy” should have the same return as the risk-free rate: buy the underlying asset now (S) and issue a forward contract, and get the forward price (F) at expiration. This is a risk-free strategy with a gross return of $F/S = e^{my}$. See Hull (2022) 5 and 8–9 and McDonald (2014) 6–8 for more details.

Example 15.4 (*Forward-spot parity*) With $y = 0.05$, $m = 3/4$ and $S = 100$ we have the forward price $e^{3/4 \times 0.05} 100 \approx 103.82$.

Proof (of Proposition 15.3) Portfolio A: enter a forward contract, with a present value of $e^{-my} F$. Portfolio B: buy one unit of the asset at the price S . Both portfolios give one asset at expiration, so they must have the same costs today. \square

Example 15.5 (*Arbitrage when Proposition 15.3 does not hold*) Assume the same parameters as in Example 15.4, except that $F = 105$. Today: issue a forward contract, borrow $e^{-3/4 \times 0.05} 105 \approx 101.14$ and buy the underlying asset for 100. You have made a risk-free profit of 1.14. (At expiration, hand over the underlying and collect the forward price—which is just enough to repay the loan.)

Proposition 15.6 (*Forward-spot parity, continuous dividends*) When the dividend is paid continuously as the rate δ (of the price of the underlying asset), then

$$e^{-my} F = S e^{-m\delta}, \text{ so} \quad (15.6)$$

$$F = S e^{m(y-\delta)} \quad (15.7)$$

Notice that the dividends decrease the forward price. The intuition is that the forward contract does not give the right to these dividends so its present value is the underlying asset value stripped of the present value of the dividends.

Proof (*of Proposition 15.6) Portfolio A: enter a forward contract, with a present value of $e^{-my} F$. Portfolio B: buy $e^{-m\delta}$ units of the asset at the price $e^{-m\delta} S$, and then collect dividends and reinvest them in the asset. Both portfolios give one asset at expiration, so they must have the same costs today. \square

Example 15.7 (*Forward-spot parity*) With $y = 0.05$, $m = 0.75$ and $S = 100$ we have the forward price $F = e^{0.75 \times 0.05} 100 \approx 103.82$. Instead with a continuous dividend rate of $\delta = 0.01$, we get $F = e^{0.75 \times (0.05 - 0.01)} 100 \approx 103.04$.

Notice that the forward prices converges to the underlying price at expiration of the futures. Before that it can deviate because of delayed payment (+) and no part in dividend payments (-).

Empirical Example 15.8 Figure 15.4 show the underlying price and the futures price on S&P 500 developed over six months. (A futures price is typically very close to a forward price, as discussed below.)

Proposition 15.9 (**Forward-spot parity, discrete dividends*) Suppose the underlying asset pays the dividend d_i at m_i ($i = 1, \dots, n$) periods into the future (but before the expiration date of the forward contract). To do the proper discounting, let $y(m_i)$ be today's m_i -period interest rate. If the dividends are known already today, then the forward price satisfies

$$e^{-my(m)} F = S - \sum_{i=1}^n e^{-m_i y(m_i)} d_i, \text{ so} \quad (15.8)$$

$$F = e^{my(m)} S - e^{my(m)} \sum_{i=1}^n e^{-m_i y(m_i)} d_i. \quad (15.9)$$

Proof (*of Proposition 15.9) Portfolio A: enter a forward contract, with a present value of $e^{-my} F$. Portfolio B: buy one unit of the asset at the price S and sell the rights to the known dividends at the present value of the dividends. Both portfolios give one asset at expiration, so they must have the same costs today. \square

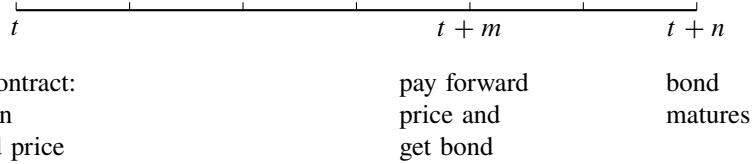


Figure 15.5: Timing convention of forward contract on a bond

15.3.3 Application: The Forward Price of a Bond

Consider a forward contract (expiring in m) on a zero coupon bond that matures in n (assuming $n > m$). See Figure 15.5 for an illustration.

By the forward spot parity (15.5) and the definition of a present value (15.3), today's forward price is

$$\begin{aligned} F &= e^{my(m)} B(n) \\ &= B(n)/B(m), \end{aligned} \tag{15.10}$$

where $B(n)$ is the price of an n -period bond today and $B(m) = e^{-my(m)}$ is the price of an m -period bond (with a face value of 1). The forward price is just the price of a long-maturity bond relative to that of a short-maturity bond.

Example 15.10 (*Forward price of a bond*) Let $(m, n, B(m), B(n)) = (5, 7, 0.779, 0.657)$. Then, $F = 0.657/0.779 \approx 0.843$.

15.3.4 Application: The Forward Price of Foreign Currency

Let S be the price (measured in domestic currency) of one unit of foreign currency. Investing in foreign currency effectively means investing in a foreign interest bearing instrument which earns the continuous interest rate (“dividend”) y^* . Use $\delta = y^*$ in (15.7)

$$F = S e^{m(y-y^*)}. \tag{15.11}$$

This is called the *covered interest rate parity* (CIP).

Example 15.11 (*CIP*) With $S = 1.20, m = 1, y = 0.0665$ and $y^* = 0.05$ we have

$$F = 1.20e^{0.0165} = 1.22.$$

Buying one unit of foreign currency costs 1.20 and after one year we have $e^{0.05} = 1.0513$ units of foreign currency, which are (when converted with $F = 1.22$) worth $1.0513 \times 1.22 = 1.2826$ in domestic currency. Since we invested 1.20, the gross return is $1.2826/1.20 = 1.0688$, which equals $e^{0.0665}$.

15.3.5 The Return on Holding a Forward Contract until Expiration

Suppose you enter a forward contract in period 0 and hold it until it expires in period m . You do not pay anything up front in, but you have pledged to pay F in period m , which has a present value of $e^{-my} F$. You could put this amount on a bank (money market) account and consider it your investment. The payoff is clearly the value of the underlying asset at expiration: S_m . The gross return is therefore

$$1 + R = \frac{S_m}{F} e^{my}. \quad (15.12)$$

For an asset with continuous (or no) dividends, the forward-spot parity (15.7) then shows that the gross return is just $S_m e^{m\delta}/S$, which is the same as holding the underlying asset (and collecting the dividends, if any).

15.3.6 The Value of an Old Forward Contract*

Consider a forward contract that expires in $t + m$, although the contract was written at some earlier point in time ($\tau < t$) and specified a forward price of F_τ (time subscripts are needed for the analysis here). The value of this contract in t is

$$W_t = e^{-my}(F_t - F_\tau), \quad (15.13)$$

where F_t is today's forward price on the same underlying asset (and same expiration date). For an underlying asset without dividends, this equals $S_t - e^{-my} F_\tau$. This value, W_t , is what someone would pay in order to buy the old forward contract. The intuition is that an owner of an old (τ) forward contract can short sell a new forward contract (t) and thereby cancel all risk—and stand to win $F_t - F_\tau$ at expiration. The present value of this is (15.13). Clearly, for a new contract ($t = \tau$), the value is zero.

Proof (15.13) An investor sells (issues) a forward contract in t . At expiration, this will give $F_t - S_{t+m}$, where S_{t+m} is the price of the underlying asset at expiration. If she buys an old forward contract (paying W_t today), the payoff of that is $S_{t+m} - F_\tau$ at expiration. Hence, the total portfolio has the payoff $F_t - F_\tau$, which is risk-free so it must earn the

risk-free rate: $(F_t - F_\tau)/W_t = e^{my}$. Rearrange to get (15.13). \square

15.4 Forwards versus Futures

A forward contract is typically a private agreement between two investors—and can therefore be tailor-made. A futures contract is similar to a forward contract (write contract, get something later at a pre-determined price), but is typically traded on an exchange—and is therefore standardized (amount, maturity, settlement process). As for the settlement, it is either in cash (paying the value of the underlying asset) or physical (delivering the underlying asset). The latter is not used for synthetic/complex assets like equity indices since it would involve considerable transaction costs.

Another important difference is that a forward contract is settled at expiration, whereas a futures contract is settled daily (*marking-to-market*). This essentially means that gains and losses (due to price changes) are transferred between issuer and owner daily—but kept at an interest bearing account at the exchange. The counterparties have to post *initial margins*—and the marking-to-market then adds to/subtracts from the margin accounts. If the amount decreases below a certain level (maintenance margin), then a *margin call* is issued to the investor—informing him/her to add cash to the margin account. See Example 15.13.

The margin requirements for an investor is governed by his/her overall portfolio (for instance, it is smaller if the portfolio includes negatively correlated positions) and is set by statistical measurements of the portfolio risk (see the *SPAN* system applied at CME and other exchanges).

If interest rates change randomly over time (and they do), the rate at which the money on the margin account is invested at will be different from the rate when the futures was issued. This risk of this happening is reflected in the futures price.

Instead and more theoretically, if the interest rate path were non-stochastic (and there was no counterparty risk), then the forward and futures prices would be the same. See the proposition below. In practice, the difference between forward and futures prices is typically small.

Proposition 15.12 (*Forward vs. futures prices, non-stochastic interest rates*) *The forward and futures prices would be the same (a) if there were no counterparty risk; (b) and if the interest rate only changed in a non-stochastic way.*

Proof (of Proposition 15.12) To simplify the notation, let $t = 0$ and $m = 2$. Also, let r_s continuously compounded rate at which you accumulate interest on the margin account between days s and $s + 1$ ($r = y/365$) and f_s be the futures price on day s . *Strategy A*: have e^{-r_1} long futures contracts on (the end of) day 0, pre-commit to increase it to 1 on day 1 and keep all settlements on the margin account. This gives

Day (s)	Settlement	Futures Position (EOD)	Margin Account (EOD)
0		e^{-r_1}	0
1	$e^{-r_1} (f_1 - f_0) = A$	1	A
2	$f_2 - f_1 = B$	0	$e^{r_1} A + B$,

where EOD means end of day. The end-value of strategy A is therefore $f_2 - f_0$, which equals $S_2 - f_0$ since the value at expiration is the value of the underlying asset. *Strategy B*: be long one forward contracts, which gives a payoff on day 2 of $S_2 - F_0$. Both strategies take on exactly the same risk, so the prices must be the same: $f_0 = F_0$. (The proof relies on knowing r_1 already on day 0.) \square

Example 15.13 (*Margin account*) Margin account of a buyer (holder) of a futures contract (here the maintenance margin = $0.75 \times$ initial margin where the initial margin might be some 3-12% of the notional value of the contract) could be as follows (assuming a zero interest rate):

Day	Futures price	Daily gain	Posting of margin	Margin account
0	100		4	4
1	99	-1		3
2	97	-2	2	3
3	99	2		5

On day 2, the investor received a margin call to add cash to the account—to make sure that the maintenance margin (here 3) is kept. Notice that the overall profit is the difference of what has been put into the margin account ($4 + 2$) and the final balance (5), that is, -1 . This is also the cumulative daily gain ($-1 - 2 + 2 = -1$). With marking to market this is all that happens: no payment of the futures price and no delivery of the underlying asset. However, it is equivalent to what happen without marking to market, since at expiration, the gain is $99 - 100 = -1$ (futures = underlying at expiration).

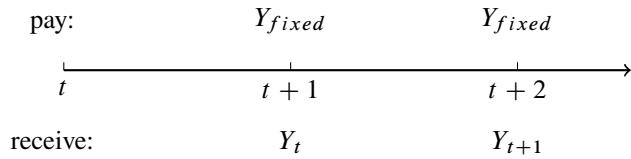


Figure 15.6: 2-year fixed-for-floating interest rate swap

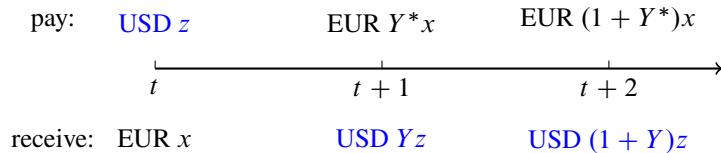


Figure 15.7: 2-year fixed-for-fixed currency swap

15.5 Swap Contracts

Swap contracts involve the exchange of two payment streams over a predefined period.

For instance, in a *fixed-for-floating interest rate swap* as illustrated in Figure 15.6, counterparty A pays a fixed interest rate at the end of each period (in the figure simplified to be each year) to counterparty B , while B the pays a floating rate (defined by referencing to an existing asset or index). In this case, this is very similar to a portfolio of forward contracts: the fixed rate is the forward price and the underlying assets are the values of the floating rates (for each respective quarter). Therefore, the pricing of the fixed leg of the swap contract could be derived from forward contracts (and vice versa).

An *FX swap* is typically just a spot buy of currency and a contracted agreement to sell it back (for a fixed price F) in a predetermined future period. This is basically a spot transaction combined with a forward contract. It can also be thought of as an exchange of loans: one counterparty lends one currency to another counterparty, who in turn lends another currency.

As an another example, Figure 15.7 illustrates a *fixed-for-fixed currency swap*. It is essentially two loans, but in different currencies: counterparty A borrows EUR (and pays interest on that), and lends USD (and receives interest on that). Counterparty B does the opposite. It is called fixed-for-fixed since both interest rates are fixed.

Chapter 16

Interest Rate Calculations

16.1 Zero Coupon Bonds

16.1.1 Zero Coupon Bond Basics

Remark 16.1 (*On the notation*) These notes often use B and Y instead of $B_t(m)$ and $Y_t(m)$, unless the indicator for the trading date (t) and/or time to maturity (m) are important in the specific context.

Consider a zero coupon bond (also called a discount or bullet bond) that costs $B_t(m)$ in t and pays the face value in $t + m$ (we will often use the short hand notation B). The time to maturity (also called tenor), m , is measured in years (for instance, $m = 1/2$ means half a year). See Figure 16.1 for an illustration.

The gross return (payoff divided by price) from investing in this bond is $1/B$, as the face value is here normalized to unity. The relation between the *bond price* B and the

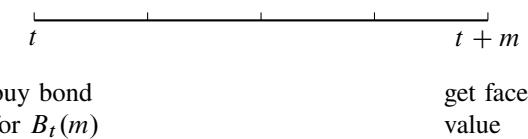


Figure 16.1: Timing convention of zero coupon bond

effective (spot) interest rate $Y(m)$ is

$$\frac{1}{B} = (1 + Y)^m, \quad (16.1)$$

$$B = (1 + Y)^{-m}, \quad (16.2)$$

$$Y = B^{-1/m} - 1. \quad (16.3)$$

Equation (16.1) states that the interest rate Y is an *annualized rate of return* derived from investing B and receiving the face value (here normalized to 1) m years later. Similarly, (16.2) says that the bond price is the present value of the face value (one). Equation (16.3) solves for the interest rate in terms of the bond price.

Example 16.2 (*Effective rates*) Consider a six-month bill so $m = 0.5$. Suppose $B = 0.95$. From (16.1) we then have that

$$\frac{1}{0.95} = (1 + Y)^{0.5}, \text{ so } Y \approx 0.108.$$

Remark 16.3 (*A face value of 100*) In case the face value is X (say, 100) instead of 1, then the bond price will be X times higher than with a face value of 1. The left hand side of (16.1) will be X/B and give the same interest rate. In practice, bond quotes are typically expressed in percentages (like 97, often leaving out the % sign) of the face value, whereas the discussion here effectively uses the fraction of the face value (like 0.97).

The relation between the interest rate and the price depends on the time to maturity (m): prices on long-maturity bonds are more sensitive to interest rate changes than prices on short-maturity bonds. The relationship is also slightly convex. These features will be important when we discuss hedging bond portfolios. See Figure 16.2 for an illustration.

We also have the following relation between the bond price and the *continuously compounded interest rate* (y)

$$\frac{1}{B} = \exp(my), \quad (16.4)$$

$$B = \exp(-my), \quad (16.5)$$

$$y = -(\ln B)/m. \quad (16.6)$$

Example 16.4 (*Continuously compounded rate*) Using the numbers as in Example 16.2, (16.4) gives

$$\frac{1}{0.95} = \exp(0.5y), \text{ so } y \approx 0.103.$$

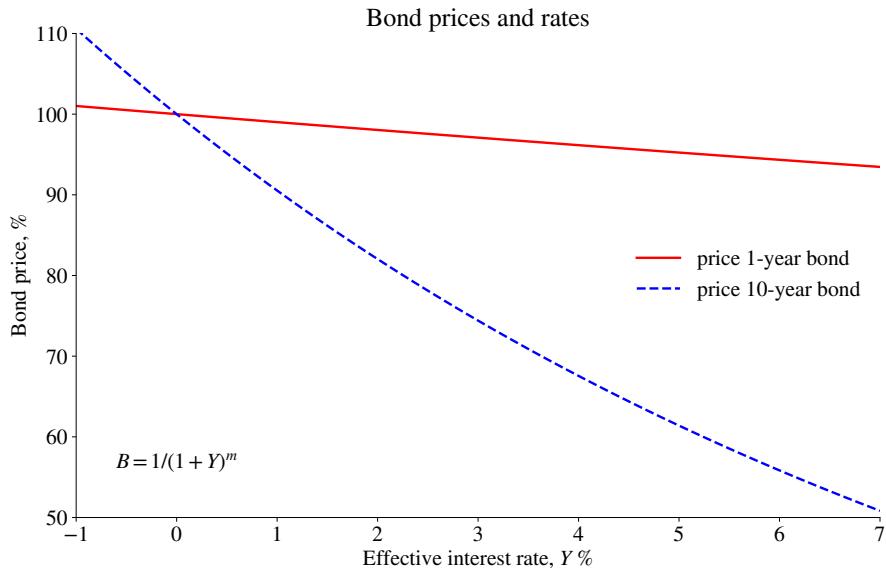


Figure 16.2: Interest rate vs. bond price

Some fixed income instruments (in particular, interbank loans) are quoted in terms of a *simple interest rate* (\tilde{Y})

$$\frac{1}{B} = 1 + m\tilde{Y} \quad (16.7)$$

$$B = \frac{1}{1 + m\tilde{Y}}, \quad (16.8)$$

$$\tilde{Y} = \frac{1/B - 1}{m}. \quad (16.9)$$

Example 16.5 (*Simple rates*) Consider a six-month bill so $m = 0.5$. Suppose $B = 0.95$. From (16.8) we then have that

$$0.95 = \frac{1}{1 + 0.5\tilde{Y}}, \text{ so } \tilde{Y} \approx 0.105.$$

Remark 16.6 (*The transformation from one type of interest rate to another**) We have

$$\begin{aligned} Y &= \exp(y) - 1 \text{ and } Y = (1 + m\tilde{Y})^{1/m} - 1 \\ y &= \ln(1 + Y) \text{ and } y = \ln(1 + m\tilde{Y}) / m, \\ \tilde{Y} &= [(1 + Y)^m - 1] / m \text{ and } \tilde{Y} = [\exp(my) - 1] / m. \end{aligned}$$

The different interest rates (effective, continuously compounded, and simple) are typically quite similar, except at very high rates. See Figure 16.3 for an illustration.

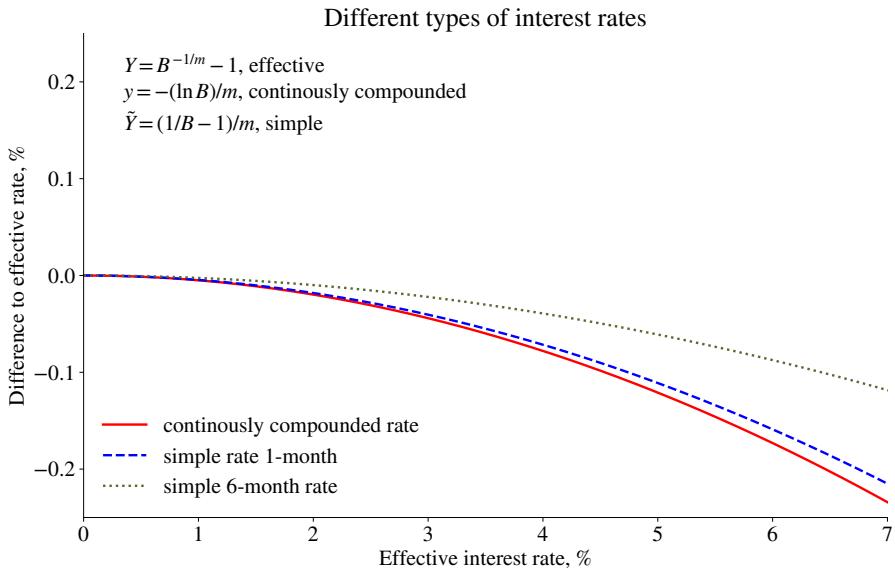


Figure 16.3: Different types of interest rates

Example 16.7 (Different interest rates) For $m = 1/2$, $Y = 0.108$, $y = 0.103$ and $\tilde{Y} = 0.106$

$$1.053 \approx (1 + 0.108)^{0.5} \approx \exp(0.5 \times 0.103) \approx 1 + 0.5 \times 0.105.$$

16.1.2 The Return from Holding a Zero Coupon Bond

The log *return from holding a zero coupon bond* from t to $t + s$ is clearly the relative change in the bond price

$$r_{t+s} = \ln \frac{B_{t+s}(m-s)}{B_t(m)}, \quad (16.10)$$

where the subscripts indicate the trading date and the values in parentheses the time to maturity, both previously suppressed. Notice that the bond's time to maturity decreases with time: in this case from m to $m - s$.

Equation (16.10) defines a return over s periods and it is *not* expressed on a “per year” basis, as interest rates are. In simplified notation (dropping the indicator of time to maturity), the right hand side is simply $\ln(B_{t+s}/B_t)$. Clearly, at maturity (when $s = m$) the bond price is 1, so (16.10) becomes $\ln(1/B_t(m)) = my$, which is the log *return from holding a zero coupon bond until maturity*.

Example 16.8 (Bond return) If the bond price decreases from 0.95 to 0.86, then (16.10)

gives the log return

$$\ln \frac{0.86}{0.95} = -0.1.$$

Substituting for the bond prices in (16.10), and using a simplified notation (by dropping the indicator of the maturity) gives

$$r_{t+s} = -m(y_{t+s} - y_t) + s y_{t+s}. \quad (16.11)$$

We use this expression to study some special cases to highlight key properties of bond returns.

Remark 16.9 (*[\(16.11\)](#) in more precise notation)...is

$$r_{t+s} = -m[y_{t+s}(m-s) - y_t(m)] + s y_{t+s}(m-s),$$

where $y_{t+s}(m-s)$ is the interest rate determined (traded) on date $t+s$ for an $m-s$ year loan.

The first special case of (16.11) considers a very short holding period (s is very small). The second term is then virtually zero, so we can write

$$r_{t+s} \approx -m(y_{t+s} - y_t) \text{ when } s \approx 0. \quad (16.12)$$

This value is clearly negative if the interest rate change is positive—and even more so if the time to maturity (m) is long. See Figure 16.4 for an illustration. Also, see Elton, Gruber, Brown, and Goetzmann (2014) 21–22 and Hull (2022) 4 for more detailed discussions.

Example 16.10 (Bond returns vs interest rate changes) Suppose that, over a split second (so the time to maturity is virtually unchanged), the interest rates for all maturities increase from 0.5% to 1.5%. Using (16.4) gives the following bond prices

	1-year bond	10-year bond
at 0.5%	$e^{-1 \times 0.005} = 0.995$	$e^{-10 \times 0.005} = 0.951$
at 1.5%	$e^{-1 \times 0.015} = 0.985$	$e^{-10 \times 0.015} = 0.861$
Change in logs (%)	-1%	-10%

Using (16.12) directly gives the same: $-1 \times 0.01 = -0.01$ and $-10 \times 0.01 = -0.1$.

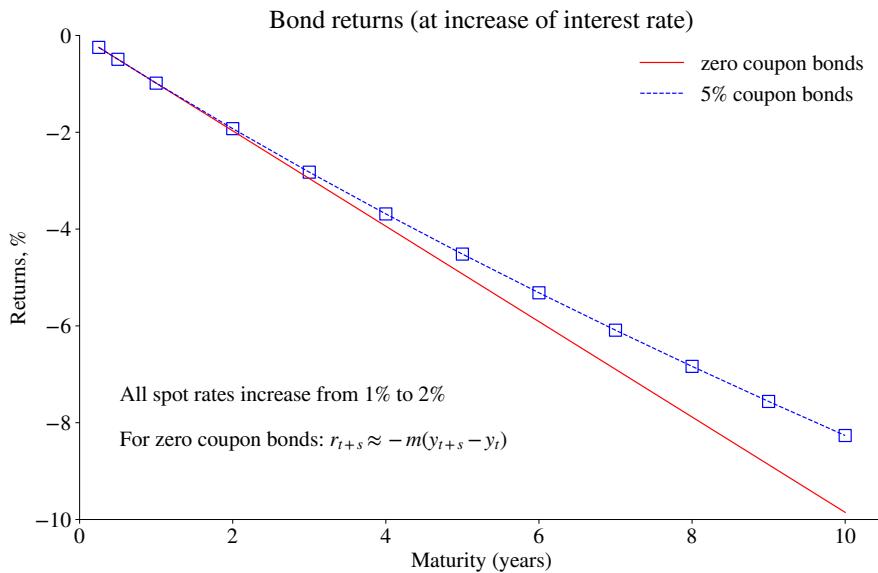


Figure 16.4: Returns after interest rate changes

The second special case is *an unchanged flat yield curve*. In this case, all interest rates in (16.11) are the same (and here denoted y), so we get

$$r_{t+s} = sy, \quad (16.13)$$

which is just the holding period times the interest rate. The reason is that the bond starts out as a m -maturity bond, but becomes an $(m - s)$ -maturity bond—and the latter has a higher price (if $y > 0$). See Figure 16.5.

16.2 Forward Rates

16.2.1 Definition of Forward Rates

A forward contract on a bond allows an investor to lock in an interest rate for a future investment period. Consider entering a forward contract in t : it specifies (a) the amount the investor has to pay at $t + m$ (the forward price, F), and (b) which discount bond that will be delivered, in particular, one that matures at $t + n$, where $n > m$. See Figure 16.6 for an illustration.

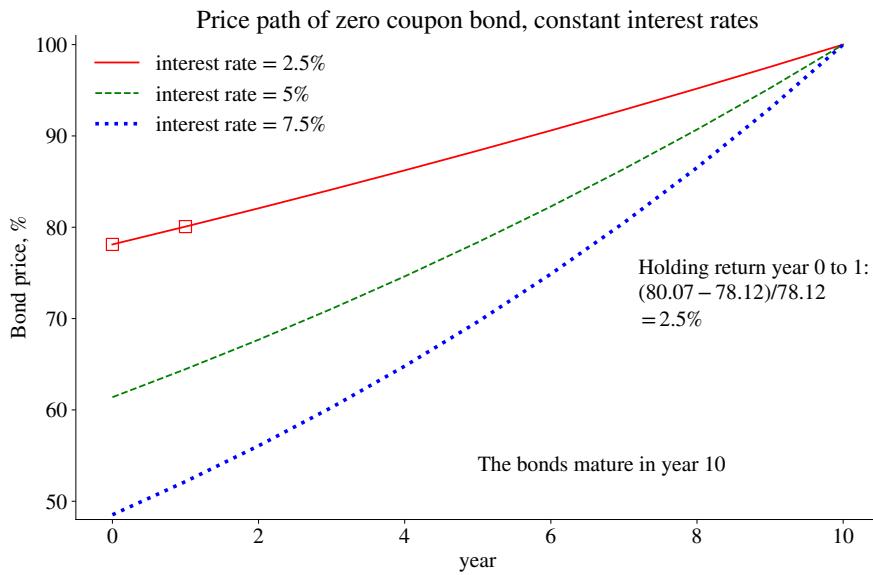


Figure 16.5: The price of a zero coupon bond maturing in year 10

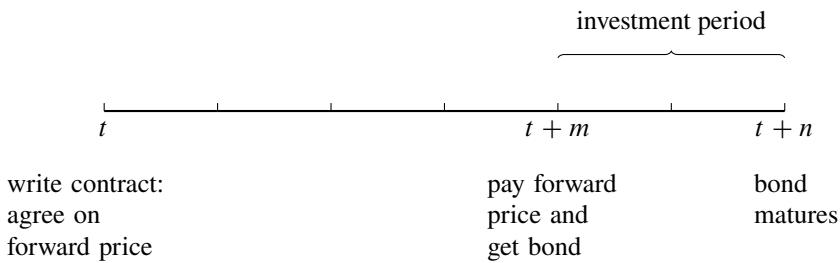


Figure 16.6: Timing convention of forward contract

16.2.2 Implied Forward Rates

The forward-spot parity establishes the relationship between the forward price, the spot rate and the current bond price

$$F = [1 + Y(m)]^m B(n). \quad (16.14)$$

Purchasing a forward contract represents a commitment to an investment from $t + m$ to $t + n$, which spans $n - m$ years. The gross return, which is known already in t , is $1/F$. A per-year effective rate of return, referred to as the *forward rate* (Γ), is defined

analogously to an interest rate

$$\frac{1}{F} = (1 + \Gamma)^{n-m}. \quad (16.15)$$

By using the relation between bond prices and yields (16.1), the forward rate can be written

$$\Gamma = \frac{[1 + Y(n)]^{n/(n-m)}}{[1 + Y(m)]^{m/(n-m)}} - 1. \quad (16.16)$$

Note that all values in this expression are determined in t . This expression demonstrates that the forward rate depends on both interest rates and, consequently, the general shape of the yield curve. Actually, the forward rate can be interpreted as the “marginal cost” of extending the loan’s duration. See Figure 16.7 for an illustration.

Example 16.11 (Forward rate) Let $m = 0.5$ (six months) and $n = 0.75$ (nine months), and suppose that $Y(0.5) = 0.04$ and $Y(0.75) = 0.05$. Then (16.16) gives

$$\Gamma = \frac{(1 + 0.05)^{0.75/0.25}}{(1 + 0.04)^{0.5/0.25}} - 1 \approx 0.07.$$

See Figure 16.7 for an illustration.

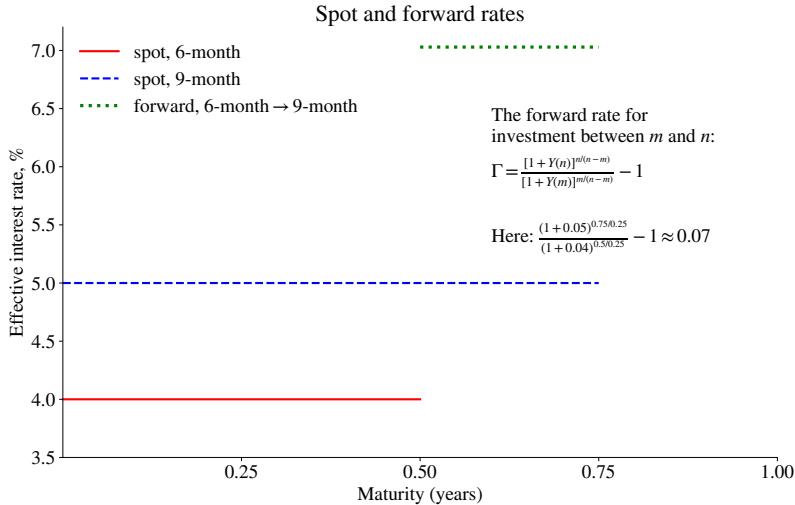


Figure 16.7: Spot and forward rates

Remark 16.12 (Forward Rate Agreement (FRA)) An FRA is an over-the-counter contract that secures an interest rate during a future period in exchange for a floating rate. The

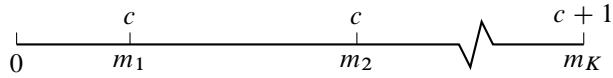


Figure 16.8: Timing convention of coupon bond

FRA does not involve any lending or borrowing; rather, it provides compensation for deviations between the future interest rate and the agreed forward rate. An FRA is similar to a one-period, and typically short-term, interest rate swap.

Remark 16.13 (*Alternative way of deriving the forward rate**) Rearrange (16.16) as

$$[1 + Y(m)]^m (1 + \Gamma)^{n-m} = [1 + Y(n)]^n.$$

This says that compounding $1 + Y(m)$ over m years and then $1 + \Gamma$ for $n - m$ years should give the same amount as compounding the long rate, $1 + Y(n)$, over n years.

16.3 Coupon Bonds

Remark 16.14 (*On the notation*) These notes often use P instead of $P_t(c, m_1, \dots, m_K)$ to denote the price of a coupon bond unless the indicator for the trading date (t), coupon rate (c) and time until coupon payments m_1, \dots, m_K are important in the specific context.

16.3.1 Coupon Bond Basics

Consider a bond that pays coupons, c , on K occasions (in $t + m_1, t + m_2, \dots, t + m_K$), and the face (or par) value, normalized to 1, at maturity ($t + m_K$). As before, m_k is measured in years. See Figure 16.8 for an illustration.

A coupon bond can be considered a portfolio of zero coupon bonds: c of them maturing in $t + m_1$, another c in $t + m_2, \dots$, and $c + 1$ in $t + m_K$. The price of the coupon bond (P) must, therefore, equal the price of the portfolio

$$P = \sum_{k=1}^{K-1} B(m_k)c + (c + 1)B(m_K) \quad (16.17)$$

where $B(m_k)$ is the price of a zero coupon bond maturing m_k years later. This is illustrated in Figure 16.9. Using the relation between (zero coupon) bond prices and spot interest



Figure 16.9: Using zero-coupon bonds to value a coupon bond

rates in (16.1), the bond price can also be written

$$P = \sum_{k=1}^K \frac{c}{[1 + Y(m_k)]^{m_k}} + \frac{1}{[1 + Y(m_K)]^{m_K}}. \quad (16.18)$$

This shows that coupon bond price is just the present value of the cash flow from coupons and the face value, but where the discounting is made by the different spot interest rates. In these calculations, P represents the full (invoice) price of the bond, which can differ from the quoted price (also called “clean price”) by an accrued interest rate term. See the appendix on market conventions for a discussion. Also, see McDonald (2014) 9 and Fabozzi (2004) for more detailed discussions.

The same valuation principle can be applied to more complicated cash flow processes, such as a portfolio of bonds. Suppose the bond portfolio pays the cash flow $c f_k$ in m_k years from now, as illustrated in Figure 16.10. This cash flow includes both coupon payments and face values. The pricing expressions (16.17)–(16.18) can then be generalised to

$$P = \sum_{k=1}^K B(m_k) c f_k \quad (16.19)$$

$$= \sum_{k=1}^K \frac{c f_k}{[1 + Y(m_k)]^{m_k}}. \quad (16.20)$$

Clearly, setting $c f_k = c$ for $k \leq K - 1$ and $c f_K = c + 1$ gives (16.17) and (16.18).

Remark 16.15 ((16.20) with continuously compounded rates*) $P = \sum_{k=1}^K c f_k / \exp[m_k y(m_k)]$.

Remark 16.16 (*Floating Rate Notes*) FRNs are bonds with floating coupon payments, typically indexed to some reference interest rate (for instance, T-bills). They are particularly common on the corporate bond market. Since the coupons are not known in advance, the approach in this section is not applicable. In a way, they are more similar to a combination of a coupon bond plus an interest rate swap (discussed below).

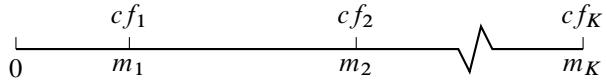


Figure 16.10: Timing convention of bond portfolio

Example 16.17 (Coupon bond prices) For the bonds with 1 and 2 years until maturity, (16.17) can be written

$$\begin{bmatrix} P(1) \\ P(2) \end{bmatrix} = \begin{bmatrix} c(1) + 1 & 0 \\ c(2) & c(2) + 1 \end{bmatrix} \begin{bmatrix} B(1) \\ B(2) \end{bmatrix},$$

where we use $P(m)$ and $c(m)$ to indicate the price and coupon rate for the m -year coupon bond. For instance, $(B(1), c(1)) = (0.95, 0)$ and $(B(2), c(2)) = (0.90, 0.06)$ we have that

$$\begin{bmatrix} P(1) \\ P(2) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.06 & 1.06 \end{bmatrix} \begin{bmatrix} 0.95 \\ 0.90 \end{bmatrix} \text{ gives } \begin{bmatrix} P(1) \\ P(2) \end{bmatrix} \approx \begin{bmatrix} 0.95 \\ 1.01 \end{bmatrix}.$$

Example 16.18 (Coupon bond price at par) Suppose $B(1) = 1/1.06$ and $B(2) = 1/1.091^2$. The price of a bond with a 9% annual coupon with two years to maturity is then

$$\frac{0.09}{1.06} + \frac{0.09}{1.091^2} + \frac{1}{1.091^2} \approx 1.$$

This bond is (approximately) sold “at par”, that is, the bond price equals the face (or par) value (which is 1 in this case).

Remark 16.19 (“Bootstrapping”) Reconsider Example 16.17, but suppose we instead have information about prices (and coupons) of the coupon bonds—and that we want to know the implied prices of the zero coupon bonds. This can be done by solving the equations for $B(1)$ and $B(2)$. That means we solve

$$\begin{bmatrix} 0.95 \\ 1.01 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.06 & 1.06 \end{bmatrix} \begin{bmatrix} B(1) \\ B(2) \end{bmatrix} \text{ to get } \begin{bmatrix} B(1) \\ B(2) \end{bmatrix} \approx \begin{bmatrix} 0.95 \\ 0.90 \end{bmatrix}.$$

(More details on bootstrapping are given in a special section of the lecture notes.)

Example 16.20 (Coupon bond prices II) Example 16.17 can be expressed in terms of interest rates (instead of zero coupon bond prices)

$$\begin{bmatrix} P(1) \\ P(2) \end{bmatrix} = \begin{bmatrix} c(1) + 1 & 0 \\ c(2) & c(2) + 1 \end{bmatrix} \begin{bmatrix} 1/[1 + Y(1)] \\ 1/[1 + Y(2)]^2 \end{bmatrix}.$$

The zero-coupon prices imply that $Y(1) \approx 5.3\%$ and $Y(2) \approx 5.4\%$ so

$$\begin{bmatrix} P(1) \\ P(2) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.06 & 1.06 \end{bmatrix} \begin{bmatrix} 1/1.053 \\ 1/1.054^2 \end{bmatrix}$$

gives the same coupon bond prices as before.

Remark 16.21 (*STRIPS, Separate Trading of Registered Interest and Principal of Securities**) A coupon bond can be split up into its embedded zero coupon bonds—and traded separately (as zero coupon bonds).

Remark 16.22 (*Bond pricing with a flat yield curve**) In the special case when all spot rates are the same (flat yield curve) and the next coupon payment is one period ($m_k = k$), then (16.18) with $Y(m_k) = Y > 0$ becomes $P = 1 + (c - Y)[1 - (1 + Y)^{-K}]/Y$, where Y is the spot rate and K is the time to maturity. The term in square brackets is positive (assuming $Y > 0$ and $K > 0$), so when the interest rate is below the coupon rate, then the bond price is above the face value (here, 1) and vice versa.

16.3.2 Yield to Maturity

The effective *yield to maturity* (also referred to as the redemption yield), θ , of a bond portfolio is the internal rate of return that satisfies the following relationship

$$P = \sum_{k=1}^K \frac{cf_k}{(1 + \theta)^{m_k}}. \quad (16.21)$$

where the portfolio has the cash flow cf_k in m_1, m_2, \dots, m_K years. This equation can be solved (numerically) for θ . Bonds are commonly quoted based on their yield to maturity, rather than their price. For a *par bond* where $P = 1$, the yield to maturity is equal to the coupon rate. For a zero coupon bond, the yield to maturity equals the spot interest rate.

Example 16.23 (*Yield to maturity*) A 4% (annual coupon) bond with 2 years to maturity. Suppose the price is 1.019. The yield to maturity is 3% since it solves

$$1.019 \approx \frac{0.04}{1 + 0.03} + \frac{1.04}{(1 + 0.03)^2}.$$

Example 16.24 (*Yield to maturity of a par bond*) A 9% (annual coupon) par bond (price of 1) with 2 years to maturity. The yield to maturity is 9% since

$$\frac{0.09}{1 + 0.09} + \frac{1.09}{(1 + 0.09)^2} = 1.$$

Example 16.25 (*Yield to maturity of a portfolio*) A 1-year discount bond with a *ytm* (effective interest rate) of 7% has the price $1/1.07$ and a 3-year discount bond with a *ytm* of 10% has the price $1/1.1^3$. A portfolio with one of each bond has a *ytm*

$$\frac{1}{1.07} + \frac{1}{1.1^3} = \frac{1}{1+\theta} + \frac{1}{(1+\theta)^3}, \text{ with } \theta \approx 0.091.$$

This is clearly not the average *ytm* of the two bonds. It would be, however, if the yield curve was flat.

Remark 16.26 (*Approximate *ytm***) $\theta \approx 2[(c + (1 - P)/K]/(1 + P)$ is sometimes used as an approximation. For the bond in Example 16.23 we would get $\theta \approx 3.02\%$. However, this approximation becomes less precise when the bond price is far from par (for instance, because of large coupon payments).

16.3.3 The Return from Holding a Coupon Bond

To calculate the *return from holding a coupon bond until maturity*, it is necessary to specify *how the coupons are reinvested*. If the coupons are reinvested through forward contracts (agreed upon at the time of purchase), the return is the same as that of a zero-coupon bond. This result is intuitive because the investor purchases the bond now and receives no payments until maturity, similar to a zero-coupon bond. This is summarised in the following proposition.

Proposition 16.27 (*Return from holding a coupon bond until maturity, another special case*) If the coupons are reinvested by forward contracts, then the (annualized) return on holding the bond until maturity is the current spot rate (on a zero coupon bond with the same maturity).

Note that this result holds regardless of the coupon rate. For this reason, it can well be said that coupons do not really matter for returns. With other assumptions about how the coupons are reinvested, the result is different (but typically not very much so).

Proof (of Proposition 16.27) Consider a 2-year coupon bond. From (16.18), the price of the bond is $P_t = B_t(1)c + B_t(2)(c + 1)$. From (16.15), we know that the forward contract for the first coupon has the gross return (until maturity) $B_t(1)/B_t(2)$. The value of the reinvested coupon and the face value at maturity is then $cB_t(1)/B_t(2) + c + 1$. Dividing by the first equation (the investment) gives $1/B_t(2)$ so the return on buying and holding (and reinvesting the coupons) this coupon bond is the same as the 2-year spot interest rate. (The extension to more years is straightforward.) \square

Example 16.28 (*Holding a coupon bond until maturity*) Suppose that the spot (zero coupon) interest rates are 4% for one year to maturity and 5% for 2 years to maturity (the zero coupon bond prices are $B(1) = 0.962$ and $B(2) = 0.907$). A 3% coupon bond with 2 years to maturity must have the current price

$$\frac{0.03}{1.04} + \frac{0.03 + 1}{1.05^2} \approx 0.963.$$

However, the value of the bond portfolio at maturity, if the coupon is reinvested by a forward contract, is

$$0.03 \times \frac{0.962}{0.907} + 0.03 + 1 \approx 1.062,$$

so the gross return over two years is approximately $1.062/0.963 \approx 1.102$. Compare that to $(1 + 0.05)^2$, which is approximately the same (some small rounding differences).

A more hypothetical (text book) case is when we (in the future) can reinvest at today's yield to maturity. The next proposition summarizes this, and the proof is in an appendix.

Proposition 16.29 (*Return from holding a coupon bond until maturity, a special case*) If all coupons are reinvested in assets that generate returns equal to the bond's yield to maturity θ , then the (annualized) rate of return is θ .

The gross return from holding a coupon bond until a period before maturity depends on both the price development on the bond and the value in $t + s$ of the (reinvested) coupon payments received. When there are changes in the interest rate level and we sell the bond before maturity, then the capital gains/losses often dominate: lower interest rates mean capital gains and vice versa (just like for zero coupon bonds). For long-maturity bonds, the effects can be considerable. See Figure 16.4 for an illustration

Empirical Example 16.30 Figure 16.11 shows monthly returns on a basket of U.S. T-bonds. These returns are probably less volatile than equity returns, but still show non-trivial movements.

In the special case where the coupons are locked in by forwards, then the bond is effectively transformed into a zero-coupon bond, so the return is same as on an m_K -year zero coupon bond bought in t and sold in $t + s$ (with $s \leq m_K$). The next proposition summarizes this. (A proof and some examples are in an appendix.)

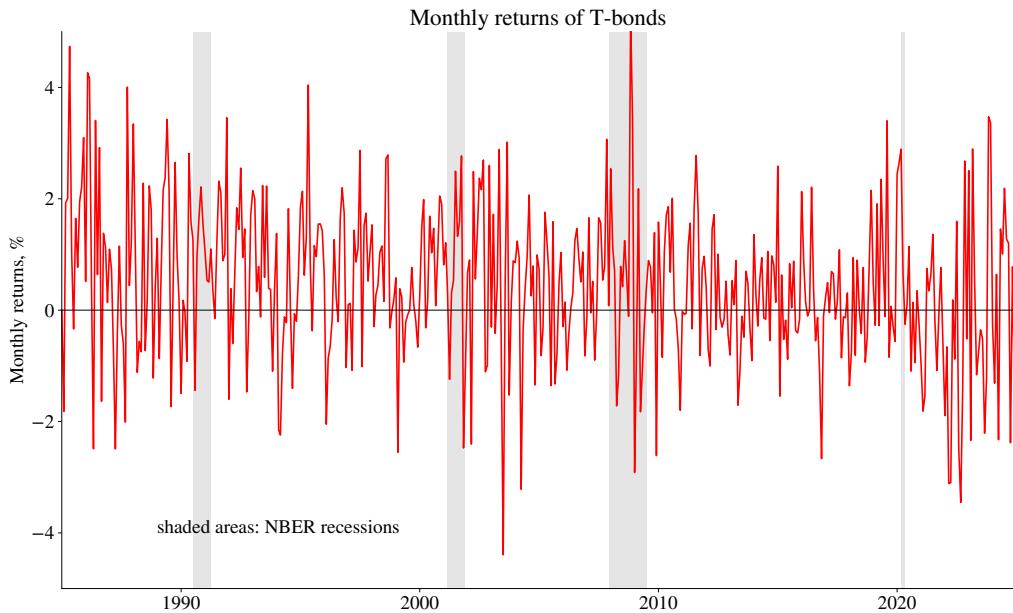


Figure 16.11: Returns on an index of U.S. Treasury bonds

Proposition 16.31 (*Bond holding return, a special case*) Suppose we reinvest the coupons with forward contracts—as if we were going to hold the bond until maturity m_K . Holding the bond until $t + s$ ($s \leq m_K$) gives the total gross return $B_{t+s}(m_K - s)/B_t(m_K)$. This implies that the portfolio has the same return as an m_K -year zero coupon bond bought in t , which becomes an $m_K - s$ zero coupon bond in $t + s$.

16.4 Other Credit Instruments

16.4.1 Overnight Indexed Swap (OIS)

Overnight indexed swaps (OIS) have supplanted the earlier LIBOR market for lending and borrowing between financial institutions, as well as for valuing derivatives. In its simplest form, such a contract (agreed upon at t) specifies a fixed payment in $t + m$ (the OIS rate) against receiving an accumulated value, which is approximately an average of the realised overnight (“floating”) interest rates between t and $t + m$. See Figure 16.12 for an example. (Also, see the appendix on bond market conventions for details on the accumulation of the floating interest rates.) These contracts typically have a notional face value that scales the payment.

Remark 16.32 With a notional value of 1000, an OIS rate of 4% and an accumulated

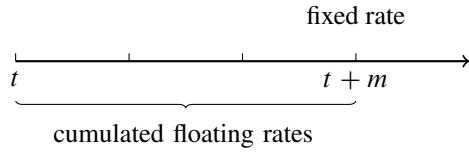


Figure 16.12: Timing convention of an OIS swap with one payment

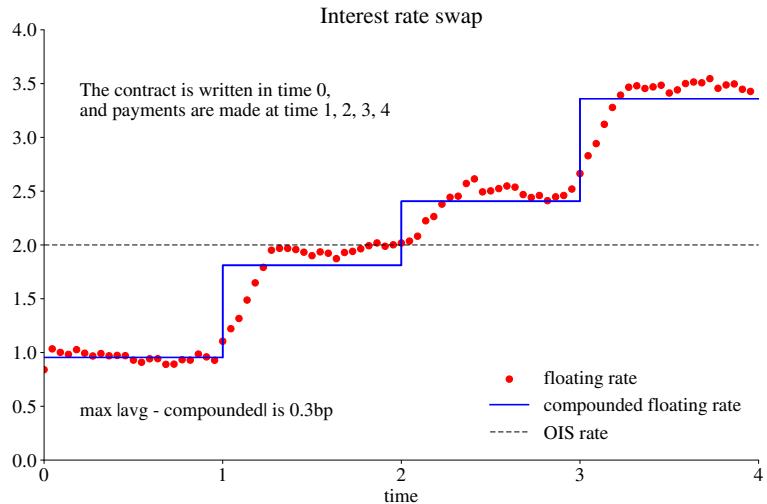


Figure 16.13: OIS with several payments

floating rate of 3%, the payment at the end of a 3-month contract is $1000 \times (0.04 - 0.03) \times 0.25$.

For longer-maturity contracts (m exceeding a year), the structure often differs, involving periodical payments (typically every three months), where the fixed OIS rate is compared with the cumulative overnight interest rates since the last payment. See Figure 16.13 for an illustration.

Empirical Example 16.33 Figure 16.14 shows the Euro OIS rates (1m to 12m) since late 2019.

16.4.2 Repo

In a repurchase agreement (*repo*), investor A sells a security to investor B, with an agreement to repurchase it at a predetermined price at some specific future time (the next

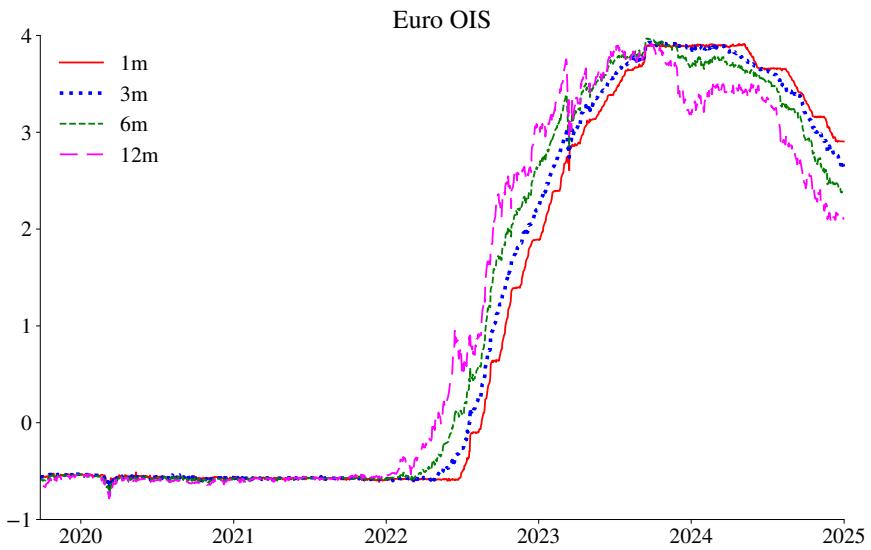


Figure 16.14: Euro OIS rates

day, after a week, etc.). The *repo rate* is calculated as the relative difference between the initial and the repurchase price.

This contract essentially implies that investor A borrows cash, while investor B borrows the asset. Investor B is said to have made a reverse repo, and can sell the asset to someone else. This is a way of shortening the security, so the repo rate is low if there is a demand for shortening the security.

A *haircut* (for instance, 3%) indicates that the collateral (security) has market value that is 3% higher than the agreed price in the repo. This provides a safety margin to the lender—since the market price of the security could decrease over the life span of the repo.

Example 16.34 (Long-short bond portfolio). First, buy bond X and use it as collateral in a repo (the repo borrowing finances the purchase of the bond). Second, enter a reverse repo where bond Y is used as collateral and sell the bond (selling provides cash for the repo lending).

16.4.3 Collateralized Debt Obligations

A collateralized debt obligations (CDO) is a repackaging of a portfolio of bonds (referred to as “collaterals”), in which the claims are divided into tranches with varying degrees of seniority. For instance, with junior, mezzanine and senior tranches, the higher tranches are

often protected against any losses (unless they are dramatic/total). In contrast, the junior tranche is similar to equity.

CDOs are created for two main reasons. First, it is a way for the issuer (typically a bank), to “package and sell off.” This is a way to shrink the balance sheet for the bank (securitisation) but still earn a fee. Second, a CDO transforms a portfolio of risky bonds to (a) some safe bonds and (b) some very risky ones. This opens up new possibilities for investors. For instance, it may allow risk averse investors (including pension funds) to invest into the safe tranches, while they would otherwise not dare (or be allowed to) invest into the original bonds.

The correlation between the defaults of the bonds within the CDO is a critical factor. The idea of tranching (in particular, to regard the senior tranche as safe) depends on the assumption that not all bonds default at the same time. Underestimating the correlation can result in significant overpricing of the senior tranches, as was frequently observed during the financial crisis 2008–9.

Another important aspect of the CDO is whether the originator (bank) holds the junior tranche or not. If it does, then it has the incentives to screen the borrowers/monitor the loans, otherwise not.

16.4.4 Credit Default Swaps

A credit default swap (CDS) is a financial instrument that provides insurance against the default on a bond. Often, the CDS is a contract where one investor pays a premium (say, every quarter) in return for an insurance in case a bond defaults. Many CDS contracts are priced under the convention that a default implies only 40% of the face value can be recovered (referred to as the recovery rate), with the remainder considered lost. In such cases, it is the probability of default which is the main driver of the pricing.

If you hold a portfolio of one risky bond and a CDS on it, then you effectively own a risk-free bond. The other way around is to buy one risk-free bond and issue a CDS, which gives effectively the same as owning the risky bond. This straightforward observation is essential for understanding how the CDS premium is calculated.

16.5 Appendix – Estimating the Yield Curve*

The (zero coupon) spot rate curve is of particular interest: it helps us price any bond or portfolio of bonds—and it has a clear economic meaning (“the price of time”).

year	Prob of survival to year t end	Prob of default in year t	Expected spread payment	Expected payment from insurance	Expected PV of net payment
1	0.98	0.02	$0.98s$	0.02×0.6	$0.98s - 0.012$
2	0.95	0.03	$0.95s$	0.03×0.6	$0.95s - 0.018$
Sum					$1.93s - 0.03$

Table 16.1: Example of the payment flows of a 2-year CDS with an assumed recovery rate of 0.4 and a risk-free interest rate of zero. The CDS spread is denoted s .

In some cases, the spot rate curve is actually observable—for instance from swaps and STRIPS. In other cases, the instruments traded on the market include some zero coupon instruments (bills) for short maturities (up to a year or so), but perhaps only coupon bonds for longer maturities. This means that the spot rate curve needs to be calculated (or estimated). This section describes different methods for doing that.

16.5.1 Direct Calculation of the Yield Curve (“Bootstrapping”)

We can sometimes calculate large portions of the yield curve directly from bond prices by a method called “bootstrapping.”

For instance, with coupon bonds maturing in the next three periods, (16.17) can be used to write

$$\begin{bmatrix} P(1) \\ P(2) \\ P(3) \end{bmatrix} = \begin{bmatrix} c(1) + 1 & 0 & 0 \\ c(2) & c(2) + 1 & 0 \\ c(3) & c(3) & c(3) + 1 \end{bmatrix} \begin{bmatrix} B(1) \\ B(2) \\ B(3) \end{bmatrix},$$

which is a recursive (triangular) system of equations. We can solve for the zero-coupon bond prices $B(1)$, $B(2)$ and $B(3)$ and then use (16.1) to transform to spot interest rates.

Example 16.35 (Bootstrapping) Suppose we know that $B(1) = 0.95$ and that the price of a bond with a 6% annual coupon with two years to maturity is 1.01. Since the coupon bond must be priced as

$$0.95 \times 0.06 + B(2) \times 0.06 + B(2) = 1.01,$$

we can solve for the price of a two-period zero coupon bond as $B(2) \approx 0.90$. The spot interest rates are then $Y(1) \approx 0.053$ and $Y(2) \approx 0.054$. In this case the system of

equations is

$$\begin{bmatrix} 0.95 \\ 1.01 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.06 & 1.06 \end{bmatrix} \begin{bmatrix} B(1) \\ B(2) \end{bmatrix}.$$

Unfortunately, the bootstrap approach is tricky to use. First, there are typically gaps between the available maturities (at least outside the US treasury market). One way around that is to interpolate. Second (and quite the opposite), there may be several bonds with the same maturity but with different coupons/prices, so it is hard to calculate a unique yield curve. This could be solved by forming an average across the different bonds or by simply excluding some data. Alternatively, we use another method than the bootstrap (see below).

16.5.2 Estimating the Yield Curve with Regression Analysis

Recall equation (16.17) which expresses the coupon bond price in terms of a series of discount bond prices. It is reproduced here

$$P = \sum_{k=1}^K B(m_k)c + B(m_K). \quad (16.22)$$

If we attach some random error to the bond prices, then this looks very similar to regression equation: the coupon bond price is the dependent variable; the coupons are the regressors, and the discount function (discount bond prices) are the coefficients to estimate—perhaps with OLS. This is a way of overcoming the second problem discussed above since multiple bonds with the same maturity, but different coupons, are just additional data points in the estimation.

The first problem mentioned above, gaps in the term structure of available bonds, is harder to deal with. If there are more coupon dates than bonds, then we cannot estimate all the necessary zero coupon bond prices from data (fewer data points than coefficients). The way around this is to decrease the number of coefficients by assuming that the discount function, $B(m)$, is a linear combination of some J predefined functions of maturity, $g_1(m), \dots, g_J(m)$,

$$B(m) = 1 + \sum_{j=1}^J a_j g_j(m), \quad (16.23)$$

where $g_j(0) = 0$ since $B(0) = 1$ (the price of a bond maturing today is one).

Once the $g_j(m)$ functions are specified, (16.23) is substituted into (16.17) and the j coefficients a_1, \dots, a_j are estimated by minimizing the squared pricing error (see, for instance, Campbell, Lo, and MacKinlay (1997) 10). One possible choice of $g_j(m)$

functions is a polynomial, $g_j(m) = m^j$. Another common choice is to make the discount bond price a spline (see McCulloch (1975)).

Example 16.36 (*Quadratic discount function*) *With a quadratic discount function*

$$B(m) = a_0 + a_1m + a_2m^2,$$

we get from (16.17)

$$\begin{aligned} P(m_K) &= \sum_{k=1}^K B(m_k)c + B(m_K) \\ &= \sum_{k=1}^K (a_0 + a_1m_k + a_2m_k^2)c + (a_0 + a_1m_K + a_2m_K^2). \end{aligned}$$

Collect all constants (that does not depend on m) into a first regressor, then all terms that are linear in m into a second regressor and finally all terms that are quadratic in m into a third regressor

$$P(m_K) = a_0 \underbrace{(Kc + 1)}_{\text{term 0}} + a_1 \underbrace{(c \sum_{k=1}^K m_k + m_K)}_{\text{term 1}} + a_2 \underbrace{(c \sum_{k=1}^K m_k^2 + m_K^2)}_{\text{term 2}}.$$

For a 1-year bonds that pays no coupons and a 2-year bond that pays a 6% coupons at $m_1 = 1$ and $m_2 = 2$, we have the following matrix of regressors (the bonds are on different rows)

Bond ↓	<u>term 0</u>	<u>term 1</u>	<u>term 2</u>
1-year, 0%	1	1	1
2-year, 6%	$2 \times 0.06 + 1$	$0.06 \times (1 + 2) + 2$	$0.06 \times (1^2 + 2^2) + 2^2$
(=)	1.12	2.18	4.30.

The a_0 , a_1 , and a_2 can be estimated by OLS if we have data on at least two bonds. This method can, however, lead to large errors in the fitted yields (if not the prices).

Empirical Example 16.37 Figure 16.15 shows the estimation of the German yield curve for one trading day, based on a cross-section of government bonds.

Example 16.38 (*Cubic discount function**) *With a cubic discount function*

$$B(m) = a_0 + a_1m + a_2m^2 + a_3m^3,$$

we get

$$P(m_K) = a_0(Kc + 1) + a_1 \left(c \sum_{k=1}^K m_k + m_K \right) + a_2 \left(c \sum_{k=1}^K m_k^2 + m_K^2 \right) + a_3 \left(c \sum_{k=1}^K m_k^3 + m_K^3 \right).$$

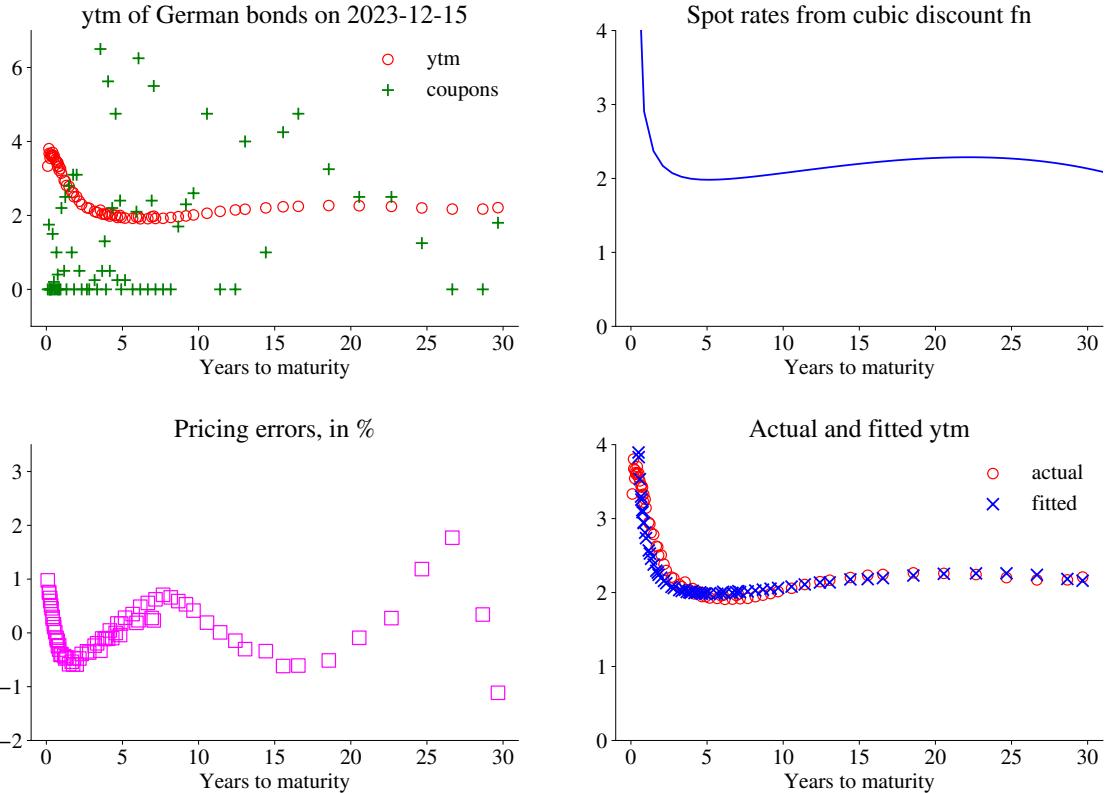


Figure 16.15: Estimated yield curves

16.5.3 Estimating a Parametric Forward Rate Curve*

Yet another approach to estimating the yield curve is to start by specifying a function for the instantaneous forward rate curve, and then calculate what this implies for the discount bond prices (discount function). (These will typically be complicated and not satisfy the simple linear structure in (16.23).)

Let $f(m)$ denote the instantaneous forward rate with time to settlement m . The *extended Nelson and Siegel forward rate function* ([Svensson \(1995\)](#)) is

$$f(m) = \beta_0 + \beta_1 \exp\left(-\frac{m}{\tau_1}\right) + \beta_2 \frac{m}{\tau_1} \exp\left(-\frac{m}{\tau_1}\right) + \beta_3 \frac{m}{\tau_2} \exp\left(-\frac{m}{\tau_2}\right), \quad (16.24)$$

where $\beta_0, \beta_1, \beta_2, \tau_1, \beta_3, \tau_2$ are parameters (β_0, τ_1 and τ_2 must be positive, and $\beta_0 + \beta_1$ must also be positive—see below). The original Nelson and Siegel function sets $\beta_3 = 0$. Note that in either case

$$\lim_{m \rightarrow 0} f(m) = \beta_0 + \beta_1, \text{ and}$$

$$\lim_{m \rightarrow \infty} f(m) = \beta_0,$$

so $\beta_0 + \beta_1$ corresponds to the current very short spot interest rate (an overnight rate, say) and β_0 to the forward rate with settlement very far in the future (the asymptote).

The spot rate implied by (16.24) is (integrate to see that)

$$y(m) = \beta_0 + \beta_1 \frac{1 - \exp(-m/\tau_1)}{m/\tau_1} + \beta_2 \left[\frac{1 - \exp(-m/\tau_1)}{m/\tau_1} - \exp\left(-\frac{m}{\tau_1}\right) \right] \\ + \beta_3 \left[\frac{1 - \exp(-m/\tau_2)}{m/\tau_2} - \exp\left(-\frac{m}{\tau_2}\right) \right]. \quad (16.25)$$

One way of estimating the parameters in (16.24) is to substitute (16.25) for the spot rate in (16.4), and then minimize the sum of the squared price errors (differences between actual and fitted prices), perhaps with 1/maturity (or 1/modified duration) as the weight for the squared error (a practice used by many central banks). Alternatively, one could minimize the sum of the squared yield errors (differences between actual and fitted yield to maturity).

Empirical Example 16.39 *Figure 16.16 shows the estimation of the German yield curve for one trading day, based on a cross-section of government bonds. Compare with Figure 16.15, especially the fitted rates at short maturities.*

16.5.4 Par Yield Curve

A par yield is the coupon rate at which a bond would trade at par (that is, have a price equal to the face value). Setting $P = 1$ in (16.17) and solving for the implied coupon rate gives

$$c = \frac{1}{\sum_{k=1}^K B(m_k)} [1 - B(m_K)], \text{ or} \quad (16.26)$$

$$= \frac{1}{\sum_{k=1}^K \frac{1}{[1+Y(m_k)]^{m_k}}} \left[1 - \frac{1}{[1+Y(m_K)]^{m_K}} \right]. \quad (16.27)$$

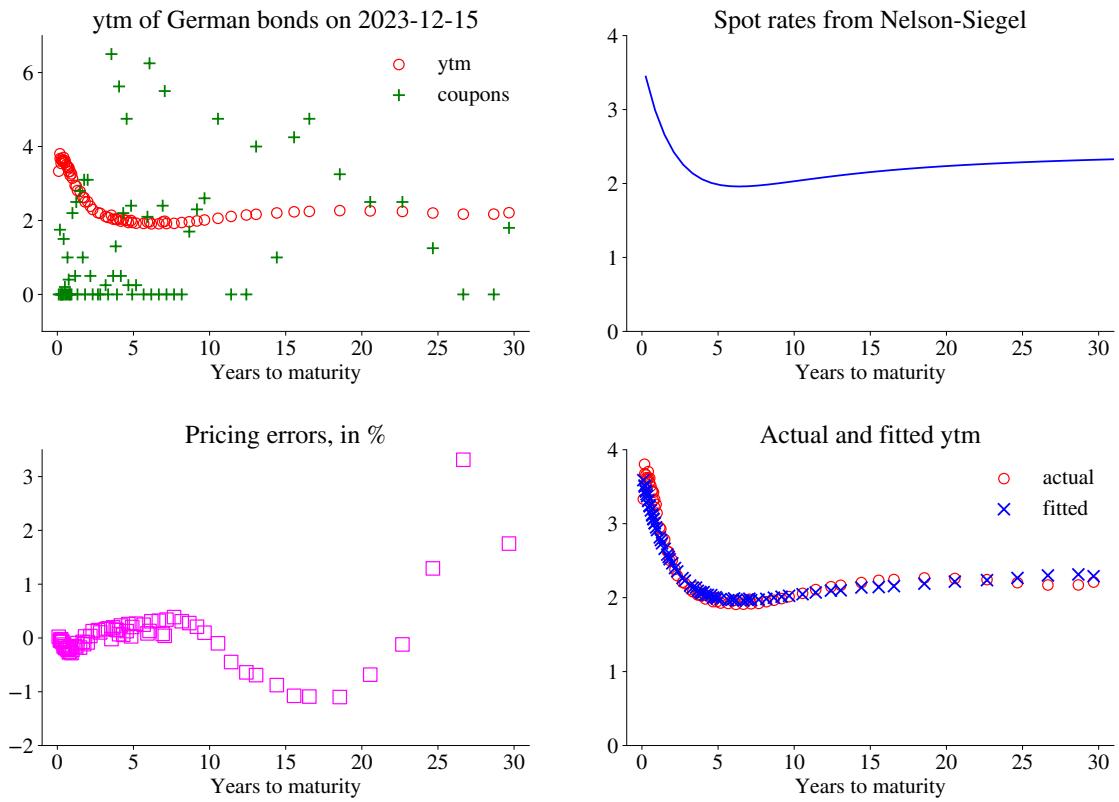


Figure 16.16: Estimated yield curves

Typically, this is very similar to the effective spot interest rates (on zero coupon bonds).

Example 16.40 Suppose $B(1) = 0.95$ and $B(2) = 0.90$. We then have

$$1 = (0.95 + 0.9)c + 0.9, \text{ so } c = \frac{1}{0.95 + 0.9}(1 - 0.9) \approx 0.054.$$

When many bonds are traded at (approximately) par, the par yield curve (16.26) can be obtained by just plotting the coupon rates. In practice, the yield to maturity is used instead (to partly compensate for the fact that the bonds are only approximately at par)—and the gaps (across maturities) are filled by interpolation. (Recall that for a par bond, the yield to maturity equals the coupon rate.) This is basically the way the Constant Maturity Treasury yield curve, published by the US Treasury, is constructed.

16.5.5 Swap Rate Curve

The swap rates for different maturities can also be used to construct a yield curve.

16.6 Appendix – Conventions on Important Markets*

16.6.1 Compounding Frequency

Suppose the interest rate r is compounded n times per year. By comparing with the definition of the effective interest rate (with annual compounding) in (16.1) we have

$$\frac{1}{B} = \left(1 + \frac{r}{n}\right)^n = 1 + Y. \quad (16.28)$$

Clearly, as $n \rightarrow \infty$, the expression in (16.28) goes to e^r , where r is the continuously compounded rate.

This shows how we can transform from semi-annual ($n = 2$) or quarterly ($n = 4$) compounding to annual compounding (and vice versa).

16.6.2 US Treasury Notes and Bonds

The convention for *US Treasury notes and bonds* (issued with maturities longer than one year) is that coupons are paid semi-annually (as half the quoted coupon rate), and that yields are semi-annual effective yields. (This applies also to most US corporate bonds and UK Treasury bonds.)

However, both are quoted on an annual basis by multiplying by two. The quoted *yield to maturity*, ϕ , solves

$$P = \sum_{k=1}^K \frac{c/2}{(1 + \phi/2)^{n_k}} + \frac{1}{(1 + \phi/2)^{n_K}}, \quad (16.29)$$

where the bond pays coupons $c/2$, in n_1, n_2, \dots, n_K half-years. By using (16.28), the yield quoted, ϕ , can be expressed in terms of an annual effective rate.

Example 16.41 A 9% US Treasury bond (the coupon rate is 9%, paid out as 4.5% semi-annually) with a yield to maturity of 7%, and one year to maturity has the price

$$\frac{0.09/2}{1 + 0.07/2} + \frac{0.09/2}{(1 + 0.07/2)^2} + \frac{1}{(1 + 0.07/2)^2} = 1.019.$$

From (16.28), we get that the yield to maturity rate expressed as an annual effective interest is $(1 + 0.035)^2 - 1 \approx 0.071$.

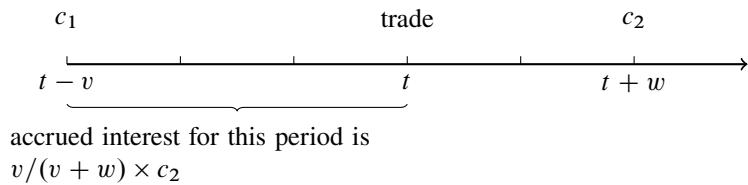


Figure 16.17: Accrued interest

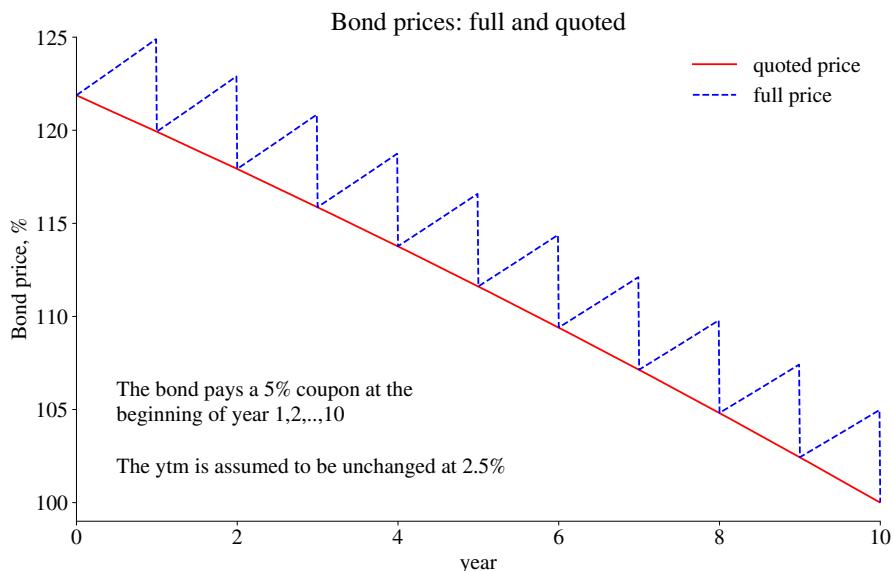


Figure 16.18: Full and quoted bond prices

16.6.3 Accrued Interest on Bonds

The quotes of bond prices (as opposed to yields) are not the full price (also called the dirty price, invoice price, or cash price) the investor pays. Instead, the full price is

full price = quoted price + accrued interest.

The buyer of the bond (buying in t) will typically get the next coupon (trading is “cum-dividend”). The accrued interest is the fraction of that next coupon that has been accrued during the period the seller owned the bond. It is calculated as

accrued interest = next coupon \times days since last coupon/days between coupons.

For instance, for US Treasury notes bonds, the next coupon is half the coupon rate and the days count uses actual days. See Figures 16.17 –16.18.

16.6.4 US Treasury Bills

Discount Yield

US Treasury bills have no coupons and are issued in 3, 6, 9, and 12 months maturities—but the time to maturity does of course change over time. They are quoted in terms of the (banker's) *discount yield*, Y_{db} , which satisfies

$$B = 1 - mY_{db}, \text{ where } m = \text{days}/360, \text{ so} \quad (16.30)$$

$$Y_{db} = (1 - B)/m. \quad (16.31)$$

Notice the convention of $m = \text{days}/360$. (If the face value is different from one, then we have $Y_{db} = [\text{face} - B]/(\text{face} \times m)$.)

From (16.1) and (16.30) it is clear that the effective interest rate and the continuously compounded interest rates can be solved as

$$Y = [1 - mY_{db}]^{-1/m} - 1 \quad (16.32)$$

$$y = -\ln(1 - mY_{db})/m. \quad (16.33)$$

Sometimes, the bills are quoted in terms of a *bond equivalent yield*, which is the simple interest rate (16.9) but using the convention of 365 days per year.

Example 16.42 A *T-bill* with 44 days to maturity and a quoted discount yield of 6.21% has the price $1 - (44/360) \times 0.0621 \approx 0.9924$. The bond equivalent (simple) interest rate is $(1/0.9924 - 1)365/44 \approx 6.35\%$.

16.6.5 European Bond Markets

The major continental European bond markets (in particular, France and Germany) typically have annual coupons and the accrued interest is calculated according to the “actual/actual” convention, that is, as

$$\text{accrued interest} = \text{next coupon} \times \text{days since last coupon}/365 (\text{or } 366).$$

(The computation is slightly more complicated for the UK and the Scandinavian countries, since they have ex-dividend periods.)

16.6.6 Short Term Reference Rates

The short term reference rates in the U.S. (SOFR) and EU (ESTR), used in overnight indexed swaps (OIS) and other contracts are based on *backward looking* compounding of overnight (one day) rates. These overnight rates are considered almost risk-free (repos in the US, unsecured in the euro area) because they apply to large financial institutions and are so short term.

The compounding is done by the formula (written using the notation in these notes)

$$\text{compounded rate(over } d_b \text{ business days)} = [\prod_{i=1}^{d_b} (1 + m_i \tilde{Y}_i) - 1] / m,$$

where \tilde{Y}_i is a simple interest rate applicable over m_i of calendar days, measured as a fraction of the year ($m_i = 1/360$ or $1/365$ if it's just one day, but $3/360$ or $3/365$ for Fridays and similarly for other business days followed by holidays), d_b is the number of business days and m the total number of calendar days as a fraction of the year ($m = 10/360$ or $10/365$ if the contract spans 10 calendar days). Notice that the $1/m$ term makes this an annualised rate.

This formula is a mix of effective compounding over business days, since $1 + m_i \tilde{Y}_i = (1 + Y_i)^{m_i}$ where Y_i is an effective rate, and simple averaging on non-business days and with the scaling by $1/m$.

Remark 16.43 (*The traditional formula*) *The formula for the compounded rate is often written*

$$[\prod_{i=1}^{d_b} (1 + \frac{n_i}{N} \tilde{Y}_i) - 1] \frac{N}{d_c},$$

where d_b is the number of business days, n_i the number of calendar days for which the rate \tilde{Y}_i applies, N number of days per year (according to the market convention) and d_c the total number of calendar days. (Both FED and ECB use this expression.)

Example 16.44 (*Compounded rate*) *For two business days, the compounding in euro area or US (where $N = 360$) could be*

$$\left[\left(1 + \frac{1}{360} 0.02 \right) \left(1 + \frac{1}{360} 0.03 \right) - 1 \right] \frac{360}{2} \approx 0.025.$$

The difference to a simple average increases as the variability of the one-day rates does and the number of days increases.

16.7 Appendix – More Proofs and Details*

16.7.1 Proof and Details of Proposition 16.29

Proof (of Proposition 16.29) Consider a 2-year coupon bond with ytm θ . From (16.21), the price of the bond is

$$P = \frac{c}{1 + \theta} + \frac{c + 1}{(1 + \theta)^2}.$$

If we can reinvest the first coupon payment to give the return θ , it is worth $c(1 + \theta)$ at maturity—and we also receive $c + 1$ at maturity. Divide the end value with the initial investment (the bond price P)

$$\frac{c(1 + \theta) + c + 1}{c/(1 + \theta) + (c + 1)/(1 + \theta)^2} = (1 + \theta)^2.$$

□

16.7.2 Proof and Details of Proposition 16.31

For instance, with a 3-year bond, the gross return on holding the bond for one year is $B_{t+1}(2)/B_t(3)$, while the gross return from holding it for two years is $B_{t+2}(1)/B_t(3)$.

Clearly, the strategy to reinvest the coupons with forward contracts essentially turns this into an m_K -year zero coupon bond (where you invest in t but do not receive any payoffs until $t + m_K$). The return of the strategy is thus the same as on holding this zero coupon bond for s years. Once again, with other assumptions about how the coupons are reinvested, the result is different.

Proof (of Proposition 16.31*) Consider a 3-year coupon bond which we hold for 1 year. Enter forward contracts like in the proof of Proposition 16.27. The value of this portfolio in $t + 1$ must be the present value of the value at maturity, that is,

$$B_{t+1}(2) \left[\frac{B_t(1)}{B_t(3)}c + \frac{B_t(2)}{B_t(3)}c + c + 1 \right],$$

where $B_{t+1}(2)$ denotes the price in $t + 1$ of a two-year zero coupon bond. Dividing by the bond price in t

$$P_t = B_t(1)c + B_t(2)c + B_t(3)(c + 1)$$

gives the gross return

$$1 + R_{t+1} = B_{t+1}(2)/B_t(3).$$

□

Example 16.45 (*Holding a coupon bond for one year*) Use the same numbers as in Example 16.28 and assume that the interest rates are unchanged. The present value in

$t + 1$ of the value at maturity is

$$0.962 \times 1.062 = 1.022.$$

Dividing by the bond price P_t , the gross return is

$$\frac{1.022}{0.963} \approx 1.06.$$

Using Proposition 16.31 directly gives $B_{t+1}(1)/B_t(2)$, which is approximately the same. Instead, if the interest rates change so $B_{t+1}(1) = 0.957$, then the return is $0.957 \times 1.062/0.963 \approx 1.055$, which is the same as $B_{t+1}(1)/B_t(2)$.

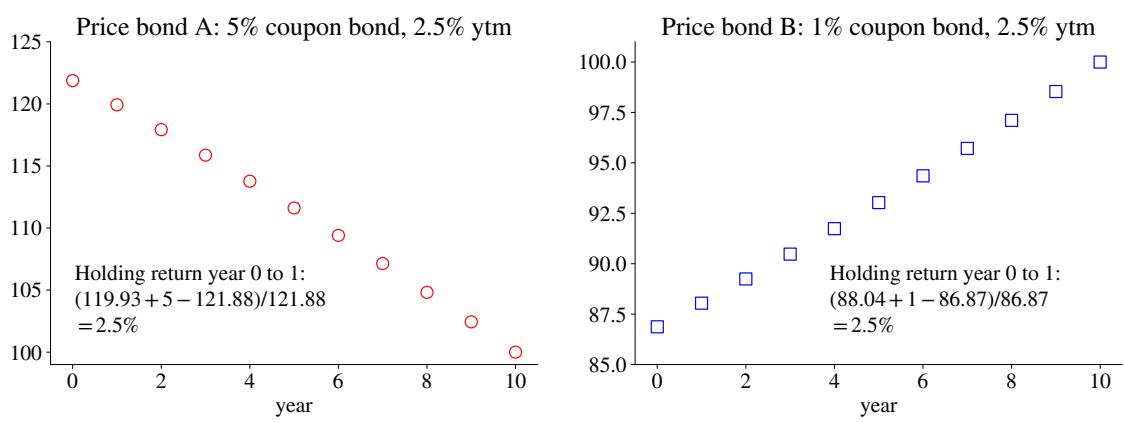
Notice that in the *special case* of holding the bond until maturity ($s = m_K$), then Proposition 16.31 shows that $1 + R_{t+s} = 1/B(m_K)$ (since $B_{t+s}(0) = 1$), which is the same result as in Proposition 16.27). In this case, the bond earns the spot interest rate $Y(m_K)$ per year.

Also, notice that in the very *special case* of a flat and unchanged yield curve (with the interest rate Y for all maturities), then Proposition 16.31 shows that the return is

$$1 + R_{t+s} = (1 + Y)^s, \quad (16.34)$$

so the return is just accumulated interest rates. See Figure 16.19 for an illustration.

Remark 16.46 (*Realized forwards**) Sometimes another set of assumptions (labelled “realized forwards”) is used to analyse the return on holding a coupon bond. In this case, the coupons are reinvested at the spot rates prevailing at the time of the coupon payment. However, it is assumed that those future spot rates will actually be equal to today’s forward rates (hence “realized”). This is clearly unrealistic, but can be used to gauge the expected return on holding the bond, at least if today’s forwards are close approximations of the expected future spot rates. The result is similar to Proposition 16.31.



Both bonds mature in year 10
 Prices are measured directly after coupon payments
 The ytm is assumed to be unchanged over time

Figure 16.19: Bond price and yield to maturity

Chapter 17

Hedging Bonds

17.1 Bond Hedging

In this chapter, we aim to hedge against price movements of a bond portfolio or a liability stream. This is called immunization. The basic idea is to form a new portfolio by combining the bond portfolio/liability stream with other bonds, making the overall portfolio “immune” to changes in the interest rates.

To simplify, the analysis in this chapter is focused on changes over a short time period, and we often make strong assumptions about how the yield curve changes (for instance, only parallel movements).

Example 17.1 (*Why a liability is not hedged by putting its present value on a bank account*) Suppose our liability is an annuity that pays 0.2 every year (starting a year from now) for 10 years. At a 5% interest rate for all maturities, the present value is

$$\sum_{k=1}^{10} \frac{0.2}{1.05^k} = 1.54.$$

Instead, with an interest rate of 3%, the present value is

$$\sum_{k=1}^{10} \frac{0.2}{1.03^k} = 1.71.$$

Putting 1.54 on a bank account will not cover the liability payments if we only get a 3% interest rate.

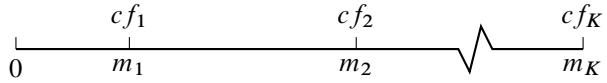


Figure 17.1: Timing convention of bond portfolio

17.2 Duration: Definitions

The “duration” of a bond portfolio is used to analyse how the price of the portfolio will change in response to changes in the yield curve. This section gives the definitions of the most commonly used duration measures.

Consider a bond portfolio with the cash flow $c f_k$ in m_k years (for $k = 1 \dots K$) as illustrated in Figure 17.1. Recall that the price P and the yield to maturity θ are related according to

$$P = \sum_{k=1}^K \frac{c f_k}{(1 + \theta)^{m_k}}. \quad (17.1)$$

The change of the price, ΔP , due to a small change in the yield, $\Delta\theta$, is approximately

$$\Delta P \approx \frac{dP(\theta)}{d\theta} \times \Delta\theta, \quad (17.2)$$

where the derivative will be calculated below. We will later also discuss how/when it makes sense to think of changes in the yield to maturity as driving bond prices.

The *dollar duration*, $D^\$$, is defined as the negative of the derivative

$$D^\$ = -\frac{dP(\theta)}{d\theta} \quad (17.3)$$

$$= \frac{1}{1 + \theta} \sum_{k=1}^K m_k \frac{c f_k}{(1 + \theta)^{m_k}}. \quad (17.4)$$

To calculate the dollar duration $D^\$$ we need all the cash flows and the times to them ($c f_k$ and m_k for $k = 1$ to K) and also the yield to maturity (θ). The latter is typically calculated by (numerically) solving (17.1) for θ .

The change of the price in (17.2) can then be written

$$\Delta P \approx -D^\$ \times \Delta\theta. \quad (17.5)$$

This expression says that an increase in the interest rate (more precisely, the yield to maturity, θ) translates into a decrease in the price—and more so if the duration ($D^\$$) is

long.

It is common to divide the dollar duration by the price, P , to get the *adjusted (or modified) duration*, D^a ,

$$D^a = D^\$ / P. \quad (17.6)$$

By dividing both sides of (17.5) by the bond price and using the definition of the adjusted duration we see that the relative change of the price (return) due to a small change in the yield is approximately

$$\frac{\Delta P}{P} \approx -D^a \times \Delta\theta \quad (17.7)$$

It is also common to multiply the dollar duration by $(1 + \theta)/P$ to get *Macaulay's duration*, D^M ,

$$D^M = D^\$ (1 + \theta) / P \quad (17.8)$$

$$= \sum_{k=1}^K w_k m_k, \text{ where } w_k = \frac{c f_k}{(1 + \theta)^{m_k} P}. \quad (17.9)$$

Macaulay's duration is a weighted average of the times to the cash flows (m_1, m_2, \dots, m_K), where the weight w_k is the fraction of the bond price accounted for by the payment in m_k ($c f_k / [(1 + \theta)^{m_k} P]$). The weights sum to unity. See Elton, Gruber, Brown, and Goetzmann (2014) 21–22, Hull (2022) 4 and McDonald (2014) 9 for more detailed discussions.

Macaulay's duration is therefore an average “time to payment” of the bond portfolio. For bond portfolios with coupons or other intermediate payments (payment of the face value of some of the bonds in the portfolio) before the last one, Macaulay's duration is less than the time to maturity, and this effect is more pronounced at large intermediate payments and at high yields to maturity. In contrast, for zero coupon bonds, Macaulay's duration equals the time to maturity. This is illustrated in Figure 17.2.

Example 17.2 (Duration) The liability in Example 17.1 has a yield to maturity (ytm) of 5% under the assumption that all interest rates are 5%. The dollar duration is

$$D^\$ = \frac{1}{1.05} \sum_{k=1}^{10} k \frac{0.2}{1.05^k} = 7.5$$

and Macaulay's duration is

$$D^M = \sum_{k=1}^{10} k \frac{0.2}{1.05^k \times 1.54} = 5.1.$$

By multiplying both sides of (17.5) by $(1 + \theta)/P$ and using the definition of Macaulay's duration we see that the relative change of the price (return) due to a small relative change

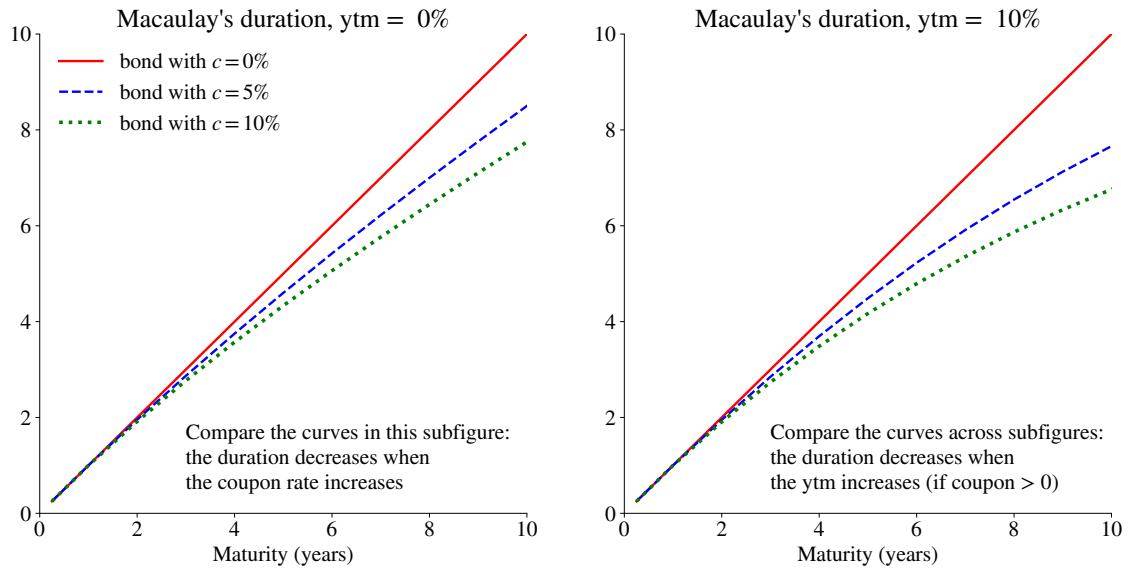


Figure 17.2: Macaulay's duration

in the yield is approximately

$$\frac{\Delta P}{P} \approx -D^M \times \frac{\Delta \theta}{1 + \theta}. \quad (17.10)$$

The term last term, $\Delta \theta / (1 + \theta)$, is the relative change in the gross yield since $\Delta \theta = \Delta(1 + \theta)$. This is the expression we will mostly work with in the rest of this chapter. See Figure 17.3 for an illustration of the fact that bonds portfolios with the same *duration* (not maturity) react similarly to interest rate changes.

Example 17.3 (Approximate price change) When the ytm changes from 5% to 3%, then (17.10) says that the liability in Examples 17.1 and 17.2 has a relative value change

$$\frac{\Delta P}{P} = -5.1 \times \frac{-0.02}{1.05} \approx 0.097.$$

From Example 17.1, we know that the exact change is $(1.71 - 1.54)/1.54 = 0.105$.

17.2.1 Duration in Special Cases*

Remark 17.4 (Duration of a zero coupon bond) For a zero-coupon bond with a face value of unity and time to maturity m , the price is $B = 1/(1 + \theta)^m$, where θ is the yield to

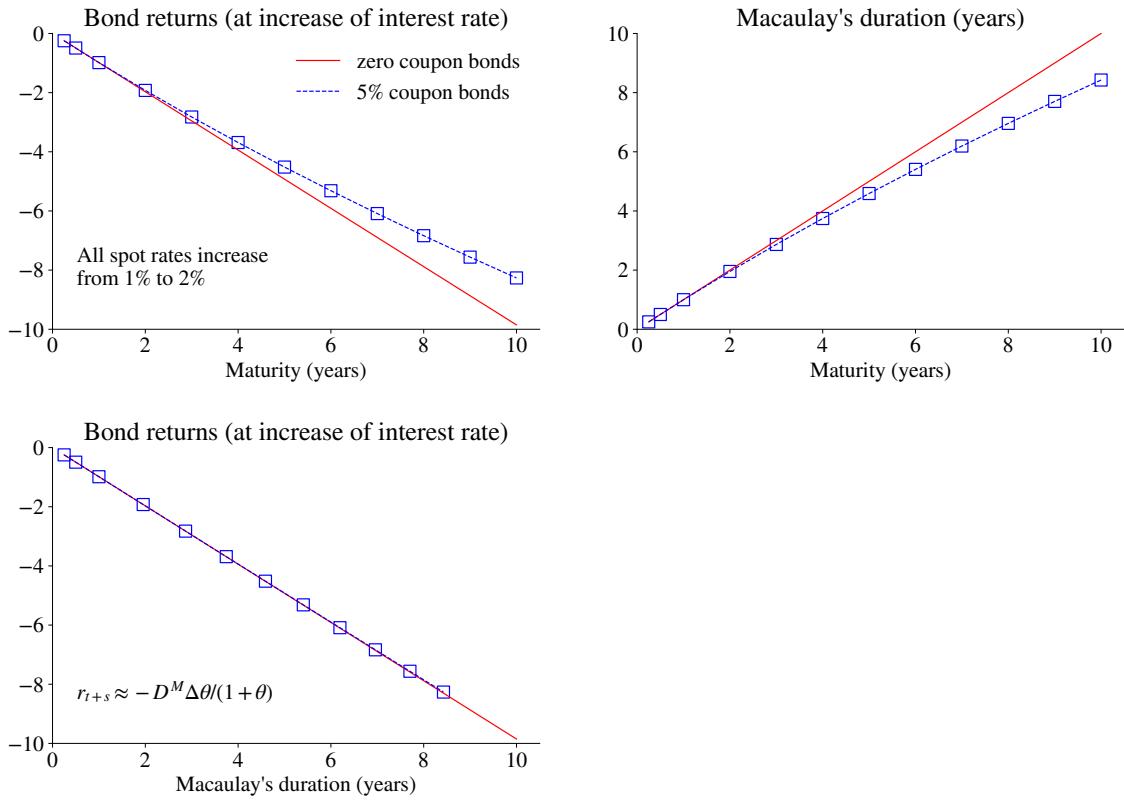


Figure 17.3: Returns after interest rate changes

maturity. The duration measures are

$$D^{\$} = \frac{m}{1 + \theta} B, D^a = \frac{m}{1 + \theta}, \text{ and } D^M = m.$$

In particular, Macaulay's duration is the same as the maturity.

The duration of a bond portfolio can be calculated by the formulas above. However, in the special case where all the bonds in the portfolio have the same yield to maturity, then there is another way. It is summarised in the next proposition.

Proposition 17.5 (*Duration of a portfolio**) If the yield to maturities of bond i and j (with prices denoted by P_i and P_j) are the same, then a portfolio of both bonds has the dollar duration $D_i^{\$} + D_j^{\$}$ and the Macaulay's duration $P_i/(P_i + P_j)D_i^M + P_j/(P_i + P_j)D_j^M$, which is the value weighted average of the different Macaulay's durations. If the ytm's are different, this does not hold.

Proof (Duration of a portfolio*) The first part of the proposition is intuitive since

the dollar duration is linear in the cash flows, see (17.4). For the second part of the proposition, multiply the dollar duration $D_i^{\$} + D_j^{\$}$ by $(1 + \theta)$ and divide by the portfolio value $(P_i + P_j)$. This is Macaulay's duration of the portfolio. Now, rewrite by using $D^{\$} = PD^M/(1 + \theta)$ to get the result in the proposition. \square

17.3 Duration to Hedge a Bond Portfolio

17.3.1 Basic Setup

This section considers how we could hedge a liability. A liability is the same as being short one unit of a bond portfolio with price P_L . We will hedge this portfolio by buying v units of a bond portfolio, denoted H , with price P_H . The value of the overall position is then

$$V = vP_H + M - P_L, \quad (17.11)$$

where M is a short-term money market account. The choice of M is typically such that the initial value of V is zero, that is, on the first day of the hedge. The subsequent amount on the money market account will change as payments are made and received and the valuation of the bonds change, as the positions are marked-to-market every day.

The purpose of setting up this portfolio is to make the value of V stable, even if interest rates change. The portfolio will typically have to be rebalanced over time in order to stay hedged.

In a first step, we choose the hedge (H) bond portfolio. Choosing a bond portfolio with a duration similar to the liability is typically a good idea. In a second step, we find v so that vP_H and P_L are equally sensitive to changes in interest rates.

One way of hedging is to hold a bond portfolio so as to *match every cash flow* of the liability, so portfolio L and H are identical ($v = 1$ and $M = 0$). However, that may be both difficult and costly because of transaction costs. The subsequent analysis will therefore focus on a case where we buy some other bond portfolio H to use as a hedge.

Example 17.6 (*Cash flow matching*) To match each cash flow of the liability in Example 17.1, we need to buy 0.2 1-year zero coupon bonds, 0.2 2-year zero coupon bond etc.

Remark 17.7 (*Overall portfolio value over several subperiods**) Start by creating a portfolio with a zero initial value

$$M_t = 0 - v_t P_{H,t} + P_{L,t},$$

where M_t is the amount held in a money market account (almost zero duration) with an interest rate Y_t . In $t + 1$ (say, one day later, $m = 1/365$), this portfolio is worth

$$V_{t+1} = v_t(P_{H,t+1} + cf_{H,t+1}) + M_t(1 + Y_t)^m - (P_{L,t+1} + cf_{L,t+1}),$$

where $cf_{H,t+s}$ and $cf_{L,t+s}$ are any cash flows (coupons) and the bond prices are measured after coupons. After rebalancing in $t + 1$, the amount on the money market account has changed to

$$M_{t+1} = P_{H,t+1}(v_t - v_{t+1}) + v_t cf_{H,t+1} + M_t(1 + Y_t)^m - cf_{L,t+1},$$

which is the same as $V_{t+1} - v_{t+1}P_{H,t+1} + P_{L,t+1}$.

Using the approximate relation of the (bond portfolio) price change (17.10), we have that the change of value, due to a sudden change in the interest rates, of the overall position is

$$\Delta V = v\Delta P_H - \Delta P_L \tag{17.12}$$

$$\approx -vD_H^M P_H \times \frac{\Delta\theta_H}{1 + \theta_H} + D_L^M P_L \times \frac{\Delta\theta_L}{1 + \theta_L}, \tag{17.13}$$

where the durations are Macaulay's duration.

The yield to maturity θ depends on the yield curve, so $\Delta\theta_H$ and $\Delta\theta_L$ may be different. For certain yield curve changes, the effect on θ is fairly straightforward. In particular, several of the hedging approaches discussed below assume that $\Delta\theta_L/(1 + \theta_L) = \Delta\theta_H/(1 + \theta_H)$, that is, a parallel shift of the yield curve. The weakness of that assumption is also discussed.

17.3.2 Yield Curve Shifts and Yield to Maturity

Bond hedging/duration analysis typically focuses on the yield to maturity (ytm, θ) as a key driver of price changes. This short section discusses how that is related to general yield curve changes.

First, the simplest case is when the yield curve is flat, meaning that spot interest rates are the same across all maturities, and shifts in parallel. Then all ytm's will change equally much, as they are equal to the interest rate. See Figure 17.4, upper left subfigure, for an illustration. Second, when the yield curve is not flat but the shift is parallel, then ytm's will change approximately the same (see upper right subfigure of Figure 17.4). Thirdly,

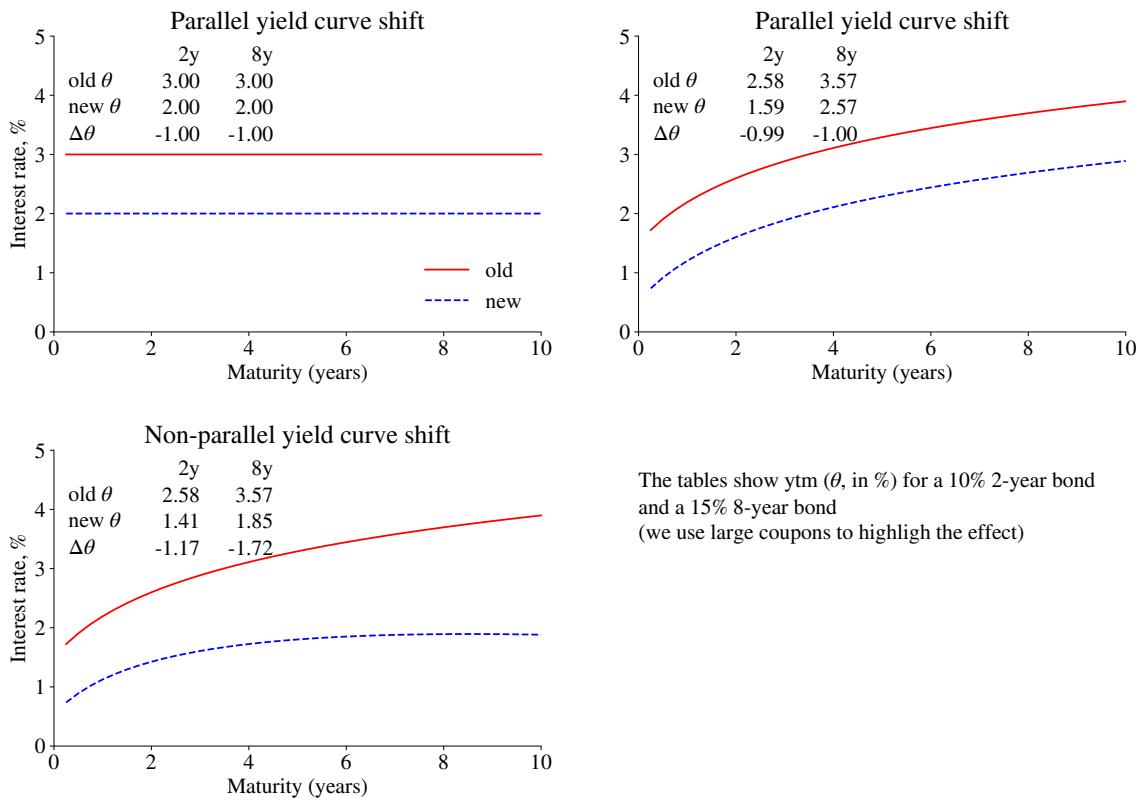


Figure 17.4: Yield curve shifts

then the yield curve shift is non-parallel, then ytms will *not* change the same (see lower left subplot of Figure 17.4)).

Example 17.8 (ytm changes) Suppose bond A pays 1 in one year and another 1 in two years, while bond B is a 2-year zero coupon bond. In the parallel yield curve shift (from the base case to scenario 1), both ytms change by approximately 1 percentage point. In the nonparallel shift (from the base case to scenario 2), only the ytm for bond A changes.

	$y(1)$	$y(2)$	θ_A	θ_B
Base case:	3%	2%	2.34	2%
Scenario 1:	2%	1%	1.33%	1%
Scenario 2:	2%	2%	2%	2%

17.3.3 Duration Matching

In this case, we choose a hedge bond (portfolio) with the same duration at the liability ($D_H^M = D_L^M$), and invest the same amount in the hedge bond as the value of the liability ($vP_H = P_L$). This means that the initial position on the money market account is zero. While the two bonds have the same durations, their cash flow streams might differ.

If the yield curve shifts up in a parallel fashion, so $\Delta\theta_L/(1 + \theta_L) = \Delta\theta_H/(1 + \theta_H)$, then (17.13) gives

$$\frac{\Delta V}{P_L} \approx 0, \quad (17.14)$$

so the duration hedge makes the overall portfolio approximately immune to interest rate changes.

As interest rates change, the duration does too. This means that a hedge bond that had the same duration as the liability in t may not be a duration match in a later period. This requires either switching hedge bond or to move over to a duration hedging (discussed below).

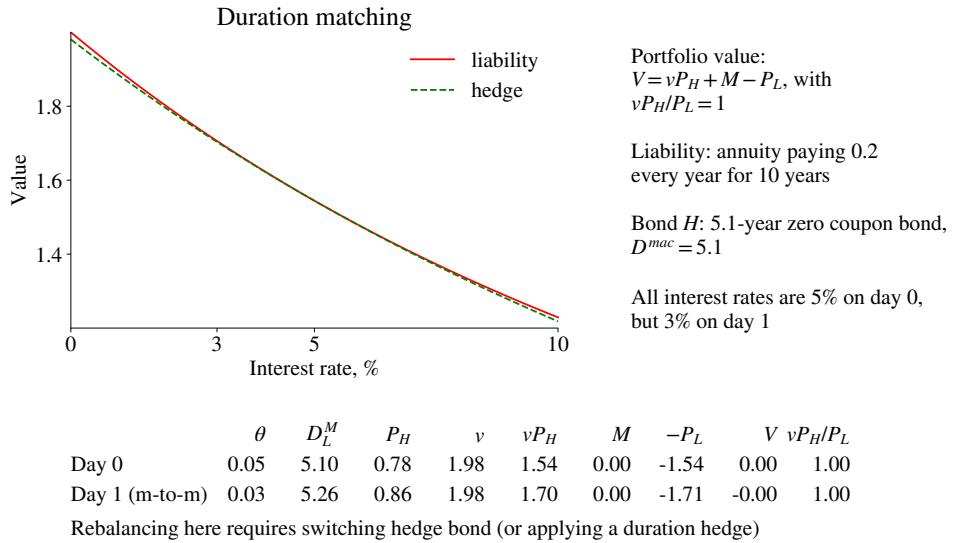


Figure 17.5: Example of duration matching. “m-to-m” stands for the marking-to-market stage

Example 17.9 (Duration matching) Figure 17.5 illustrates a case where a liability stream is hedged by a (here, zero-coupon) bond with the same duration. This appears to give a very precise hedge (the value of V stays close to zero). Notice, however, that the duration

of the liability changes as the interest rates do, so we must rebalance to be immune to further interest rate changes.

17.3.4 Naive Hedging

Suppose we again invest the same amount in the hedge bond as the value of the liability ($vP_H = P_L$), but this time we do not pay any attention to the durations.

This will typically make the overall portfolio vulnerable to interest rate changes. To illustrate that, assume, for simplicity, that the yield curve shifts up in a parallel fashion. Then (17.13) gives

$$\frac{\Delta V}{P_L} \approx (D_L^M - D_H^M) \times \frac{\Delta \theta}{1 + \theta}, \quad (17.15)$$

which depends on the duration mismatch. For instance, suppose interest rates decrease ($\Delta\theta < 0$) and the duration of the liability is longer than that of the hedge bond ($D_L^M > D_H^M$). Then, the portfolio will lose money. See Figure 17.6 for an example. The reason is that the value of the liability increases more than the value of the hedge bond, as longer-duration bonds are more sensitive to interest rate changes than shorter-duration bonds.

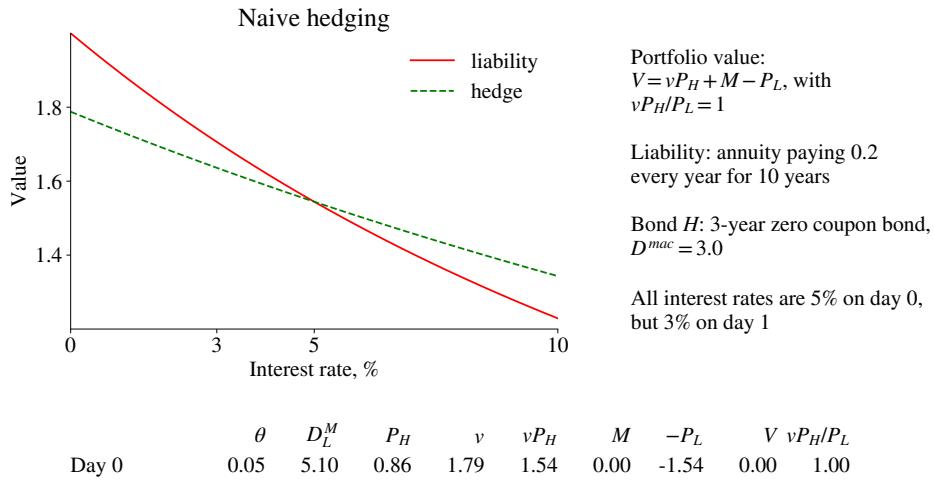


Figure 17.6: Example of naive hedging

Example 17.10 (Naive hedging) Figure 17.6 shows a case of naive hedging when we have a duration mismatch. This makes the overall portfolio (V) sensitive to interest rate changes. In this case, interest rates decrease (from 5% to 3%) so the liability increases more in value

than the hedge bond (which has too low duration—and is thus not sufficiently sensitive to interest rate changes). We face losses. In terms of (17.15), we have $D_L^M - D_H^M > 0$ and $\Delta\theta < 0$.

Remark 17.11 (*Effect of yield curve shift on a bank*) A bank typically has liabilities with short duration (deposits, inter-bank lending) and assets (plays the same role as the “hedge” above) with long duration (loans to companies and households), so $D_L^M - D_H^M < 0$. Equation (17.15) shows that an increase in the interest rate level will hurt the bank ($D_L^M - D_H^M < 0$ and $\Delta\theta > 0$) since the assets decrease more than the liabilities. This can also be phrased as follows: the bank has fixed incomes from the loans it has made, but it now needs to refinance itself (deposits and inter-bank loans) at a higher cost.

17.3.5 Duration Hedging

Instead of the naive hedge, suppose we instead choose offset the duration differences by the size of the position

$$v = \frac{D_L^M}{D_H^M} \times \frac{P_L}{P_H}, \text{ so} \quad (17.16)$$

$$\frac{vP_H}{P_L} = \frac{D_L^M}{D_H^M}. \quad (17.17)$$

Again, consider the case of $D_L^M > D_H^M$. The duration hedge in (17.17) then suggests that the *amount* invested into the hedge bond (vP_H) should exceed the value of the liability (P_L). In this way, by having a larger position, we offset the hedge bond’s lower interest rate sensitivity. The initial position on the money market account is typically nonzero. As in the other cases, the portfolio needs to be rebalanced over time.

Combine (17.13) and the hedge ratio (17.16) to get

$$\frac{\Delta V}{P_L} \approx D_L^M \times \left(\frac{\Delta\theta_L}{1 + \theta_L} - \frac{\Delta\theta_H}{1 + \theta_H} \right). \quad (17.18)$$

Suppose again that the yield curve shifts up in a parallel fashion. Then, (17.18) shows that the overall portfolio value will not change ($\Delta V/P_L \approx 0$). See Figure 17.7 for an example how the duration hedging works.

Example 17.12 (*Duration hedging*) Figure 17.7 illustrates a case where we have a duration mismatch (similar to the case of naive hedging), but where this compensated for by a hedge ratio that takes the mismatch into account. The hedge bond has a too short duration,

we therefore take a larger position in it (the amount invested into the hedge bond, vP_H , is much larger than the value of the liability)—so as to increase the interest rate sensitivity of the position.

Empirical Example 17.13 Figures 17.8–17.9 show an example based on the German yield. The value of the (artificial) liability is calculated by using estimated yield curves for each trading day. In contrast, the hedge bond is one of the bonds in the data set. Notice that the duration of the liability jumps up just after a cash flow has been made. (The average time to future cash flows is then longer.) The poor initial performance of the naive hedge is explained by the steepening of the yield curve (see Figure 17.9), which means that longer maturity bonds (the hedge bond) loose more value than shorter maturity bonds (effectively, the liability).

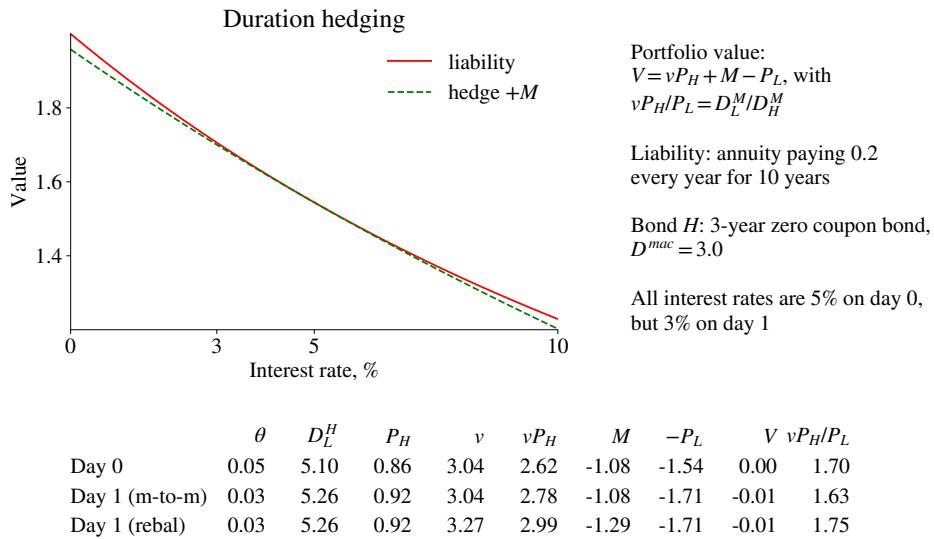


Figure 17.7: Example of duration hedging

Remark 17.14 (Using the dollar duration instead*) Recall that $D^M = D^\$(1 + \theta)/P$, so (17.13) can be rewritten as

$$\Delta V \approx -vD_H^\$ \times \Delta\theta_H + D_L^\$ \times \Delta\theta_L.$$

Set $\Delta V = 0$ to get the hedge ratio $v = (D_L^\$/D_H^\$) \times (\Delta\theta_L/\Delta\theta_H)$. If we assume that both yields change equally much, then $v = D_L^\$/D_H^\$$.

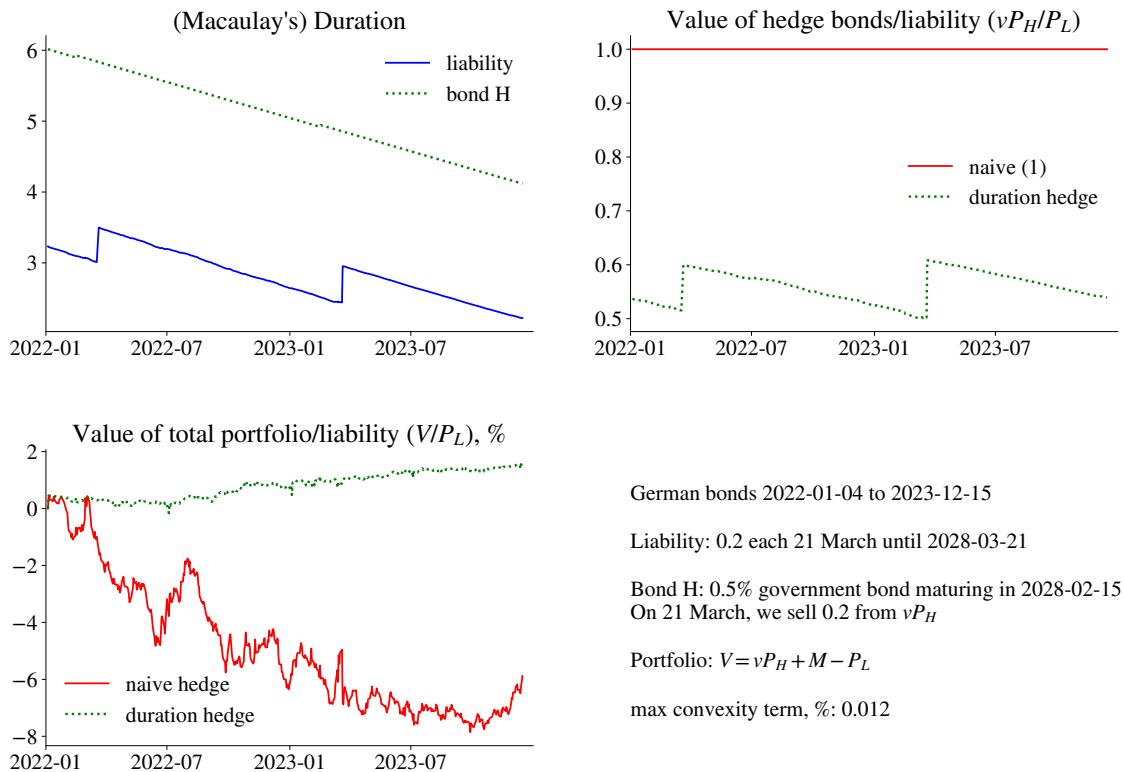


Figure 17.8: Duration hedging

17.4 Addressing Issues in Duration Hedging

This section discusses potential problems with the duration hedging.

17.4.1 Problem 1: Approximation Error

The formula for the price change (17.10) is a first-order Taylor approximation of the form

$$\Delta P \approx \frac{dP}{d\theta} \times \Delta\theta. \quad (17.19)$$

Obviously, a second-order Taylor approximation is more precise. It would be

$$\Delta P \approx \frac{dP}{d\theta} \times \Delta\theta + \frac{1}{2} \frac{d^2 P}{d\theta^2} \times (\Delta\theta)^2. \quad (17.20)$$

where the last term includes the second derivative of the bond price with respect to the yield to maturity. See Figure 17.7 for an illustration of the non-linear effect.

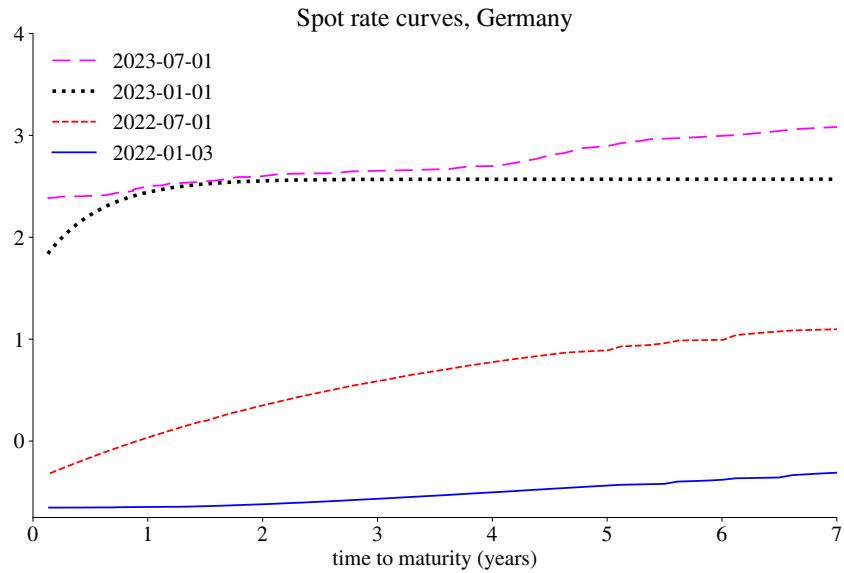


Figure 17.9: Duration hedging

Dividing (17.20) by the bond price and using (17.7) gives

$$\frac{\Delta P}{P} \approx -D^M \times \frac{\Delta \theta}{1 + \theta} + \frac{1}{2} C \times (\Delta \theta)^2, \quad (17.21)$$

where C (often called “convexity”) is the second derivative in (17.20) divided by the bond price.

The convexity is easily calculated as

$$C = \frac{1}{P} \sum_{k=1}^K m_k (m_k + 1) \frac{c f_k}{(1 + \theta)^{m_k + 2}}. \quad (17.22)$$

It is clear that the convexity is positive (since $c f_k \geq 0$), but tends to be lower if much of the cash flow comes early, similarly to the duration. Often, the convexity effect is modest compared to the duration effect, at least for bonds with short duration, see Figure 17.10. Still, choosing the hedging bond (portfolio) so that it has a similar convexity to the bond portfolio to be hedged may reduce the approximation error.

Example 17.15 (Convexity) The convexity of the 10-year bond in Example 17.1 is (when interest rates are 2%)

$$C = \frac{1}{1.54} \sum_{k=1}^{10} k(k+1) \frac{0.2}{1.05^{k+2}} \approx 35.6.$$

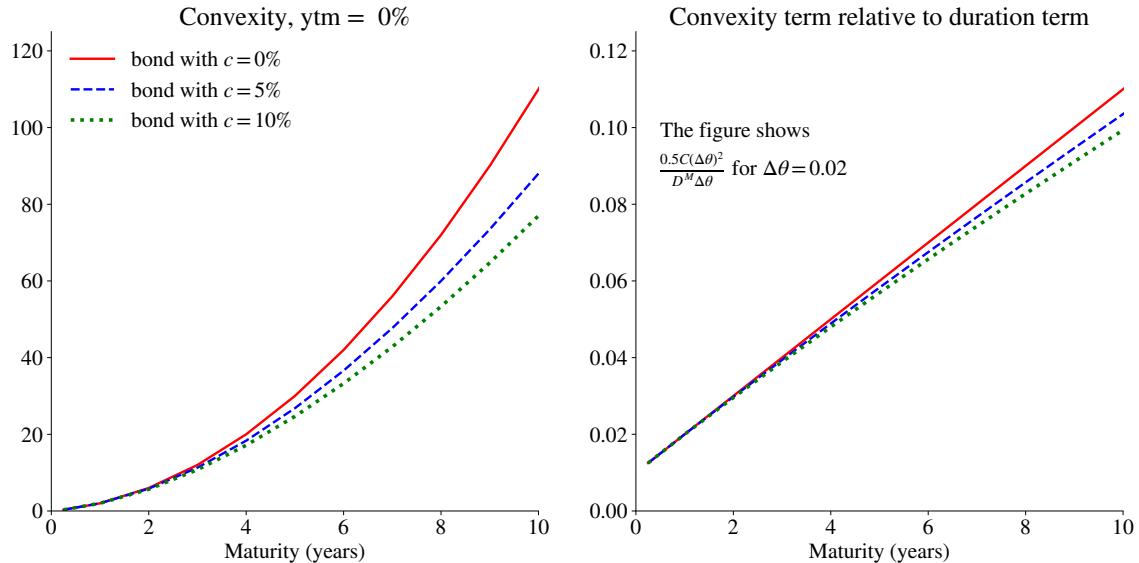


Figure 17.10: Convexity

If interest rates decrease from 5% to 3%, then the second-order term in (17.21) is

$$\frac{35.6}{2} \times 0.02^2 = 0.007,$$

which is fairly small compared to the duration effect (see Example 17.3).

17.4.2 Problem 2: Changing Cash Flows

The duration measures assume that the cash flow is unaffected by the yield change. That is true for many instruments, like most government bonds, but not for callable bonds and effectively not for bonds with (time varying) default risk. In such cases, another approach is needed.

17.4.3 Problem 3: Yield Curve Changes vs. Changes in Yields to Maturity

An important problem with using duration for hedging is that the hedge ratio in (17.23) depends on how the yields change.

The ideal case for duration hedging is when the yields to maturity move in parallel. In reality, level shifts of the entire yield curve make up a sizeable fraction of the overall variability of the curve. However, there are also other important aspects, for instance, changes in the slope of the curve.

Equation (17.18) shows how the value of the overall portfolio depends on the yields of the liability and the hedge bond. For instance, suppose the yield curve changes from being flat to being downward sloping and the hedging bond has shorter duration than the liability. In this case, the overall portfolio loses value. The reason is that the value of the hedging portfolio increases less, as the yield decreases less, in price than the liability. See Figure 17.4 for an illustration.

To overcome this problem, the hedge ratio should be (set $\Delta V = 0$ in (17.13))

$$v = \frac{D_L^M}{D_H^M} \times \frac{P_L}{P_H} \times \frac{\Delta\theta_L/(1 + \theta_L)}{\Delta\theta_H/(1 + \theta_H)}. \quad (17.23)$$

This is consistent with the duration hedging equation (17.16) when all changes of the yield curve are parallel shifts, rendering the last term in (17.23) equal to unity. Otherwise, we need to model how the yield curve changes (level, slope, curvature) in response to the overall economic situation.

Chapter 18

Interest Rate Models

18.1 Empirical Properties of Yield Curves

Yield curves in the US and most other developed countries tend to exhibit the following features: first, the yield curve is usually upward sloping; second, it changes over time, primarily due to general level shifts but occasionally due to changes in its slope.

Empirical Example 18.1 *Figures 18.1–18.2 show U.S. yield curves.*

Yield curve movements are commonly described in terms of three factors: level, slope, and curvature. One way of measuring these factors is by defining

$$\begin{aligned}\text{Level} &= y(10\text{-year}) \\ \text{Slope} &= y(10\text{-year}) - y(3\text{-month}) \\ \text{Curvature} &= [y(2\text{-year}) - y(3\text{-month})] - [y(10\text{-year}) - y(2\text{-year})].\end{aligned}\tag{18.1}$$

This means that we measure the level by a long rate, the slope by the difference between a long (maturity) and a short (maturity) rate—and the curvature (or rather, concavity) by how much the medium/short spread exceeds the long/medium spread.

Empirical Example 18.2 *Figure 18.3 shows the U.S. yield curve factors over time.*

Most evidence from US data suggests that changes in the level factor dominate, accounting for 80–90% of the total variation in yields. The slope ranks second, contributing 10%, while the curvature accounts for only a few percent.

Interest rates are strongly related to business cycle conditions, so it often makes sense to include macro economic data in the modelling.

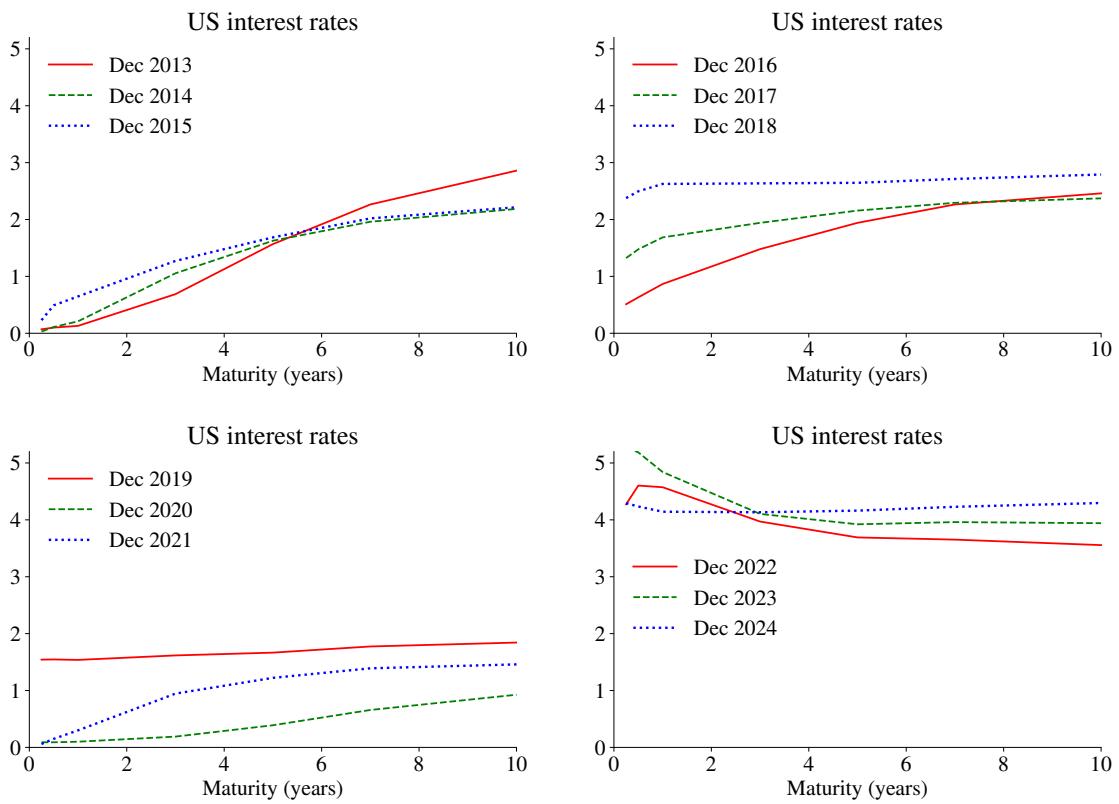


Figure 18.1: Estimated yield curves

The slope is an example of a *term spread*, that is, a difference between the interest rates for two maturities. Analysing and plotting such spreads is sometimes more meaningful than examining interest rate levels.

Empirical Example 18.3 *Figure 18.4 shows how the U.S. slope factor (long rate minus a short rate) is related to recessions. The slope factor is often very small or even negative at the beginning of recessions and then increase towards the end.*

18.2 Yield Curve Models

Yield curve models aim to describe the dynamics of the yield curve. The previous empirical evidence suggests that accounting for a level (parallel) shift is important, but that we should also try to model changes in the slope. The curvature and further factors might be less important. Such models can, among other things, improve the hedging of bond portfolios.

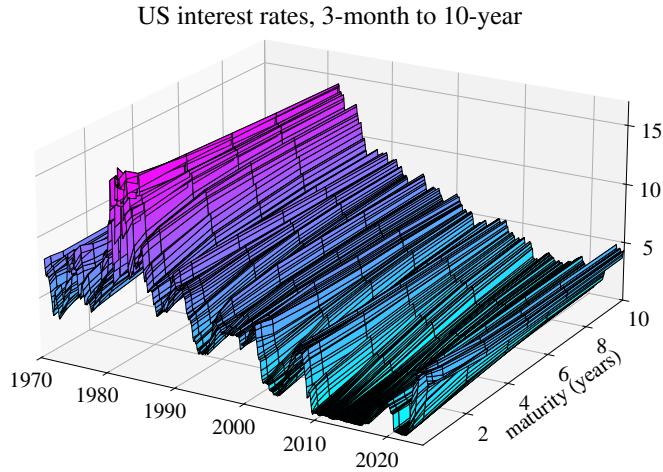


Figure 18.2: US yield curves

18.2.1 The Expectations Hypothesis of Interest Rates

The expectations hypothesis (EH) of interest rates posits that long bonds either have no risk premia or only a constant risk premium. The empirical evidence is mixed, so the expectations hypothesis is best thought of as an approximation.

EH implies that the n -period interest rate $y_t(n)$ equals the average of the 1-period (the shortest maturity) rates over t to $t + n$

$$y_t(n) = \lambda(n) + \frac{1}{n} \sum_{s=0}^{n-1} E_t r_{t+s}, \quad (18.2)$$

where r_t is short hand notation for the 1-period rate. See Figure 18.5 for an illustration.

In (18.2), the period length (from t to $t + 1$) corresponds to the maturity of the short interest rate. For instance, if r_t is a 1-month rate today, then r_{t+1} is the 1-month rate a month later and $y_t(120)$ is today's 120-month (10 year) interest rate. As usual, all interest rates are annualized rates of returns of keeping the bond until maturity. These features require some care when using $y_t(n)$ in bond pricing formulas.

Example 18.4 (*The expectations hypothesis*) Suppose the $(r_t, E_t r_{t+1}, E_t r_{t+2}, E_t r_{t+3}) = (3\%, 2\%, 2\%, 1\%)$ are the expected 1-month interest rates, then the 4-month rate is $\lambda(4) + 2\%$.

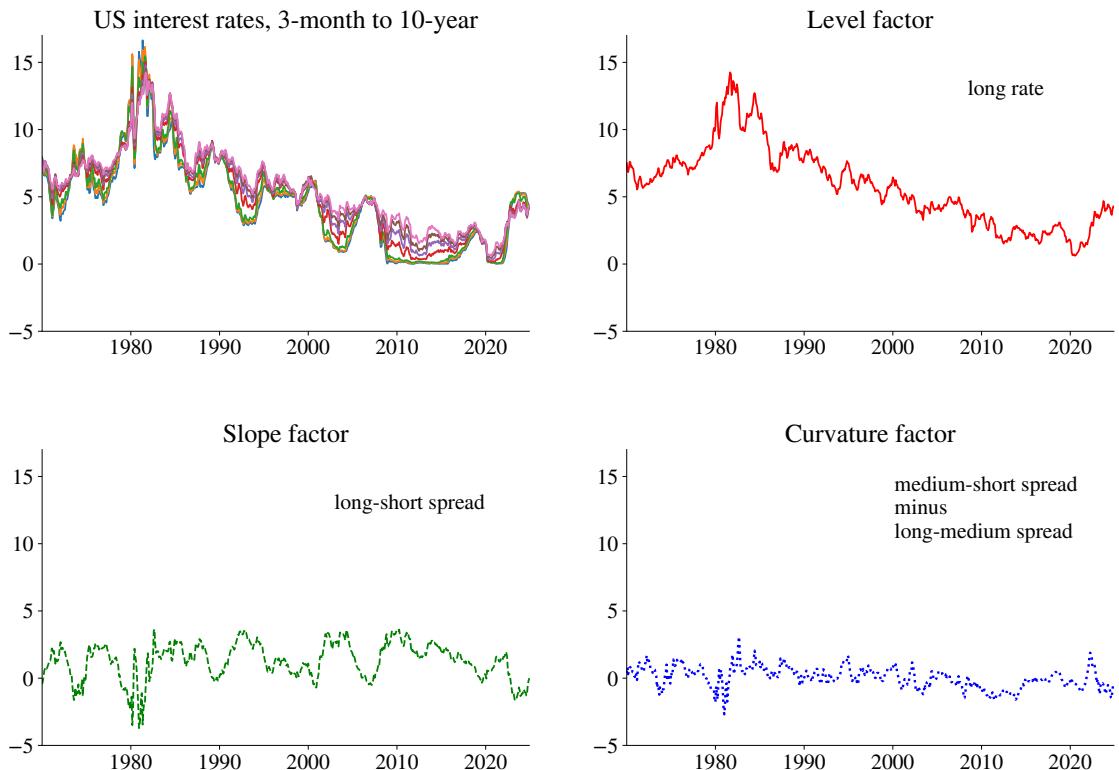


Figure 18.3: US yield curves: level, slope and curvature

The expectations hypothesis allows for constant risk premia ($\lambda(n) \neq 0$), which may differ across maturities (n). If $\lambda(n) = 0$, then the *pure* expectations hypothesis is said to hold.

18.2.2 Risk Premia

There are several reasons for why bonds should have risk premia. First, long bonds are risky for investors who do not intend to hold them until maturity and therefore carry term premia. Second, some bonds are infrequently traded (for instance, off-the-run bonds and many index-linked bonds) and are likely to have liquidity premia. Third, the real return of a long bond is very sensitive to inflation changes, likely more so than equities. Bonds are therefore likely to have inflation risk premia. In general, the typical upward sloping yield curve observed in data is consistent with the view that long-maturity bonds have risk premia.

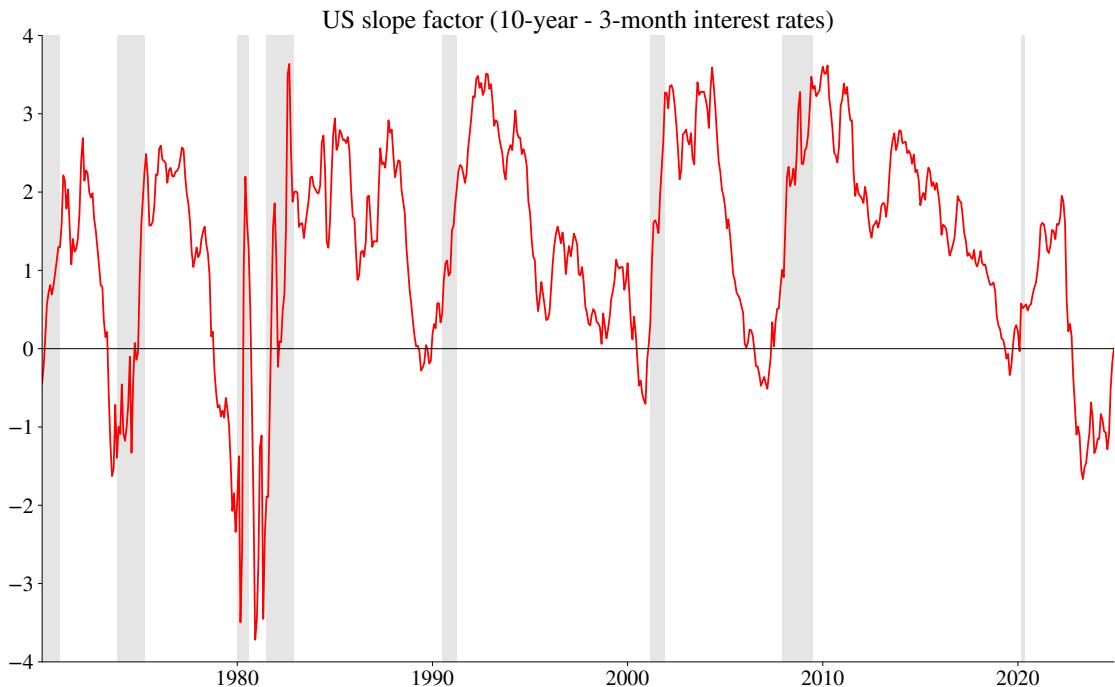


Figure 18.4: US slope factor and recessions

18.2.3 A Simple One-Factor Model: The Vasicek Model

The [Vasicek \(1977\)](#) model uses a single factor to model the entire yield curve: the short interest rate, which is assumed to follow an autoregression of the first order, an AR(1).

To present a simplified version of the model, the current section applies some unspecified constant risk premia. The more general formulation (discussed in an appendix) derives the risk in terms of the mean reversion and volatility of the short rate.

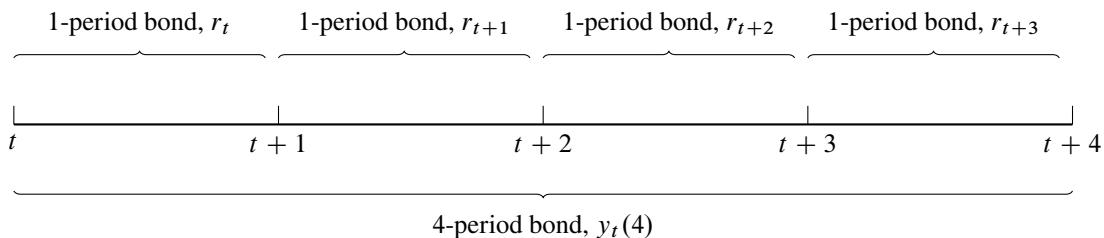


Figure 18.5: Timing for expectations hypothesis

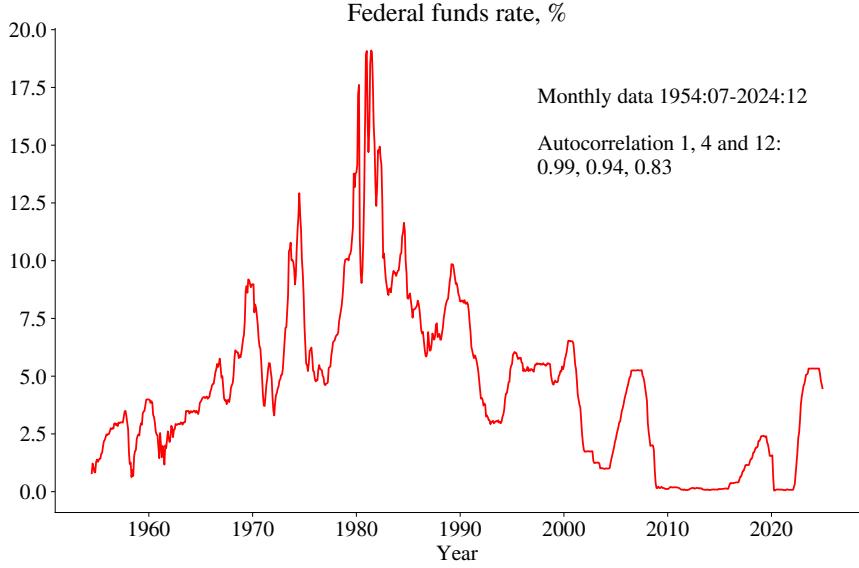


Figure 18.6: Federal funds rate, monthly data

To simplify the notation, let the short interest rate, r_t , follow an AR(1)

$$r_{t+1} - \mu = \rho(r_t - \mu) + \varepsilon_{t+1}, \quad (18.3)$$

where μ is the average short interest rate, and ρ describes the dynamics. Typically, we consider the mean-reverting (stationary) case when $0 < \rho < 1$, but we will also discuss the borderline case of $\rho = 1$.

Empirical Example 18.5 *Figure 18.6 shows how the U.S. Federal Funds rate has developed over time. It shows significant persistence.*

Remark 18.6 *(Alternative formulation of (18.3)*) The process is sometimes specified in terms of changes as $r_{t+1} - r_t = a(\mu - r_t) + \varepsilon_{t+1}$. Clearly, this can be written $r_{t+1} - \mu = (1 - a)(r_t - \mu) + \varepsilon_{t+1}$, where $1 - a$ corresponds to ρ in (18.3). With $0 < a < 1$ (that is, with $0 < \rho < 1$) the process is mean reverting.*

The forecast made in t of r_{t+s} is

$$\mathbb{E}_t r_{t+s} = (1 - \rho^s)\mu + \rho^s r_t, \quad (18.4)$$

where \mathbb{E}_t denotes expectations formed in t . Notice that when r_t is a 1-month rate, then (18.4) is today's expectation of the 1-month rate in s months. See Figure 18.7 for an

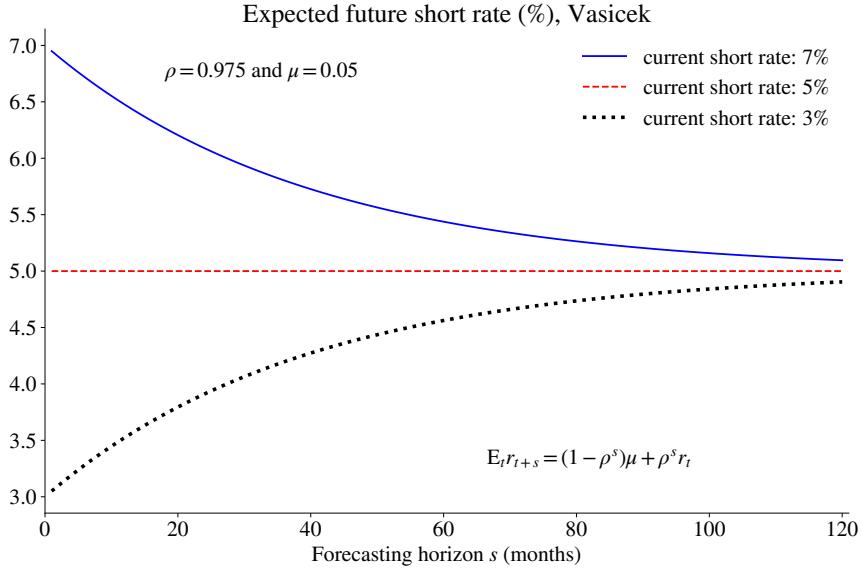


Figure 18.7: Expected future short rate in Vasicek model, for different initial short rates

illustration of how these expectations depend on the starting value r_t and the horizon (for some specific (μ, ρ) parameters).

Example 18.7 (*Predictions from an AR(1)*) With $\mu = 0.05$, $\rho = 0.975$ and $r_t = 0.07$, then $E_t r_{t+50} = 0.72 \times 0.05 + 0.28 \times 0.07 = 0.056$.

Remark 18.8 (*Calibrating the AR(1) to data**) Notice that (18.3) implies that $\text{Corr}(r_t, r_{t-s}) = \rho^s$, so we could thus estimate ρ by $\text{Corr}(r_t, r_{t-s})^{1/s}$. If the AR(1) is a very good fit to data, then it should not matter (much) if you use $s = 1$ or $s = 12$ (say). In practice, the results may well differ. For instance, suppose monthly data gives $\text{Corr}(r_t, r_{t-1}) = 0.99$ but $\text{Corr}(r_t, r_{t-12}) = 0.80$, which imply $\rho = 0.99$ and $\rho = 0.982$ respectively. This matters for the pricing of long-maturity bonds: with 120 months (10 years) we get $0.99^{120} = 0.3$ while $0.98^{120} = 0.09$. Which value we choose to use depends on whether we are most interested in the short maturities (use $\rho = 0.99$) or the long maturities (use $\rho = 0.982$).

We now assume that the expectations hypothesis holds for continuously compounded rates. Using this in (18.2) gives the long interest rate. For instance, the two-period (annualized, continuously compounded) rate is

$$\begin{aligned} y_t(2) &= \lambda(2) + \frac{1}{2} [r_t + (1 - \rho) \mu + \rho r_t] \\ &= \lambda(2) + \mu (1 - \rho) / 2 + r_t (1 + \rho) / 2, \end{aligned} \tag{18.5}$$

where we have collected the terms that are constant first and those that involve r_t last. The general expression for a maturity of n periods is

$$\begin{aligned} y_t(n) &= a(n) + b(n)r_t, \text{ where} \\ a(n) &= \lambda(n) + \mu [1 - b(n)] \text{ and} \\ b(n) &= (1 + \rho + \dots + \rho^{n-1})/n = (1 - \rho^n)/[(1 - \rho)n]. \end{aligned} \tag{18.6}$$

Again, notice that the period length is defined by the maturity of the short rate. For instance, when r_t is a 1-month rate, $y_t(120)$ is a 120-month (10 year) rate.

Remark 18.9 (*A recursive expression for $b(n)$) Equation (18.6) implies $b(n) = [1 + \rho(n-1)b(n-1)]/n$, where the recursion starts at $b(1) = 1$.

In this model, all movements of the yield curve are driven by the short rate, so it is a *one-factor model*. The shifts of the yield curve are parallel if $\rho = 1$ (the random walk model) since then $b(n) = 1$ in (18.6), so we get

$$y_t(n) = \lambda(n) + r_t, \text{ if } \rho = 1. \tag{18.7}$$

For lower values of ρ , the short rate process r_t is mean-reverting, so the expected future short rates (and therefore the current long rates) are always closer to the mean than the current short rate. See Figures 18.8–18.9 for an illustration. Also, see Hull (2022) 31 for a more detailed discussion.

Example 18.10 (Vasicek model) For $\rho = 0.975$ and $\mu = 0.05$, (18.6) gives (assuming no risk premia)

$$\begin{bmatrix} y_t(1) \\ y_t(2) \\ y_t(3) \\ y_t(4) \end{bmatrix} \approx \begin{bmatrix} 0 \\ 0.00062 \\ 0.00124 \\ 0.00184 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.988 \\ 0.975 \\ 0.963 \end{bmatrix} r_t.$$

18.3 The Vasicek Model: Hedging a Bond

The Vasicek model allows us to calculate a potentially better way of *hedging a bond portfolio* than the duration hedging. The model can account for both level and slope changes of the yield curve, while the duration hedging was based on the assumption of only level (parallel) shifts.

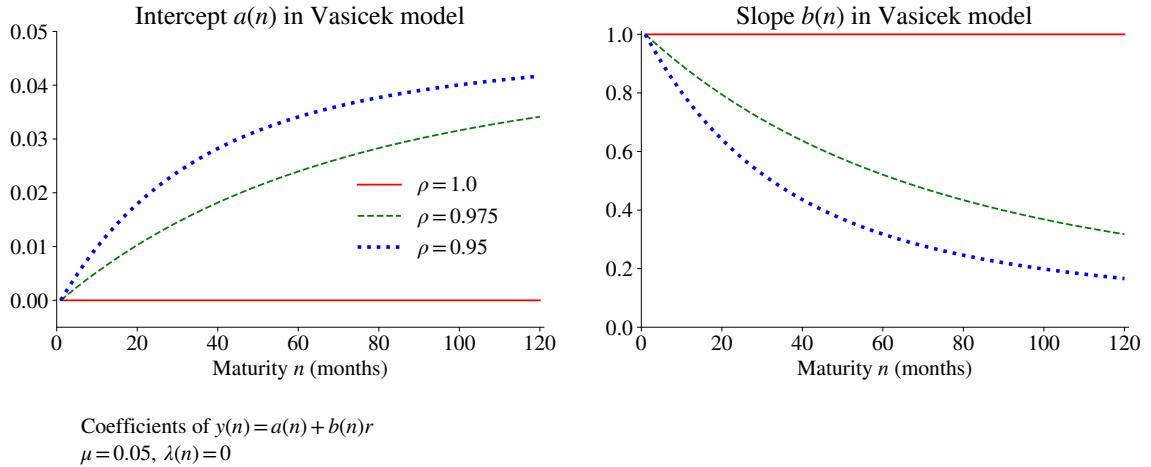


Figure 18.8: Intercept and slope in the Vasicek model

Recall we have a liability worth P_L , and we buy v units of a bond portfolio (denoted H) with price P_H . The value of the overall position is

$$V = vP_H + M - P_L, \quad (18.8)$$

where M is a short-term money market account.

The change of the hedge portfolio (over a short time interval) is

$$\Delta V = v\Delta P_H - \Delta P_L, \quad (18.9)$$

and a bond price can be calculated as

$$\begin{aligned} P &= \sum_{k=1}^K B(m_k) c f_k \\ &= \sum_{k=1}^K \frac{c f_k}{\exp[m_k y(m_k)]}, \end{aligned} \quad (18.10)$$

where $c f_k$ is the cash flow at $t + m_k$, and $y(m_k)$ is the continuously compounded interest between t and $t + m_k$. Notice that time (m_k) is here measured in *years* since the interest rates $y(m_k)$ are annualized rates.

Once we know the parameters of the Vasicek model, it is straightforward to numerically calculate what ΔP_H and ΔP_L are, as functions of the change in the current short rate interest rate (Δr_t). In practice, this can be done by the following steps.

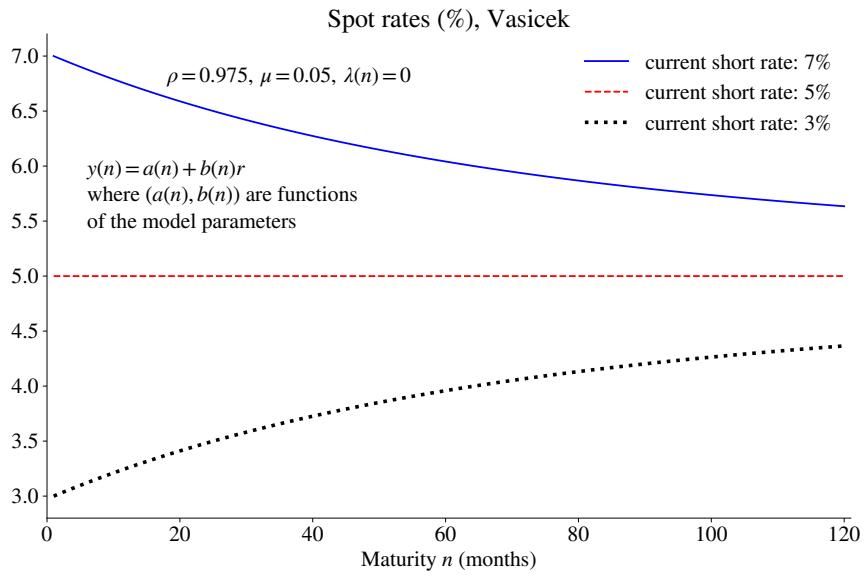


Figure 18.9: Vasicek model, spot rates for different initial short rates

1. For an initial value of the short interest rate r , use (18.6) to calculate all spot rates $y(m_k)$ needed in (18.10). Notice that the periods in the Vasicek model (n) might be shorter than years. For instance, if $m_k = (0.5, 1, 10)$ years but the Vasicek model is for monthly data, then calculate $y(n)$ for $n = (6, 12, 120)$ months according to (18.6) and use them for $m_k = (0.5, 1, 10)$ in (18.10). See Figure 18.10.
2. Use the interest rates $y(m)$ to calculate the prices of the hedge bond and the liability according to (18.10).
3. Redo points 1 and 2, but starting from another short rate, say, the earlier r_t plus 1%.
4. Calculate the difference of the prices at the two different short rates ($\Delta P_H, \Delta P_L$). We then set v so that $\Delta V = 0$, that is, $v = \Delta P_L / \Delta P_H$.

This approach identifies the sensitivity of P_L and P_H to the primary driver of the yield curve: the short interest rate (r). Effectively, $v = \Delta P_L / \Delta P_H$ will capture how the yield to maturity is driven by the short interest rate (r), but also the duration.

Remark 18.11 (*Duration hedging with the Vasicek model**) The Vasicek model can also be used to calculate the yield changes in a duration hedge. Recall that the following value (dollars) invested into a hedge bond is (H) relative to the value of the liability (L) should

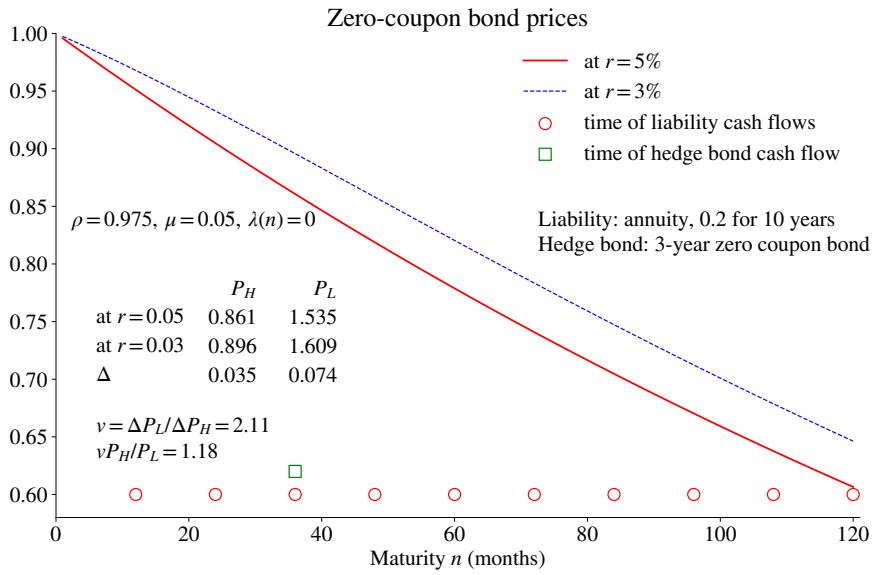


Figure 18.10: Bond prices in the Vasicek model

provide a good hedge:

$$vP_H / P_L = \frac{D_L^M}{D_H^M} \times \frac{\Delta \theta_L / (1 + \theta_L)}{\Delta \theta_H / (1 + \theta_H)}$$

where D_i^M is Macaulay's duration, θ_i the yield to maturity and P_i the price of bond i . In the typical duration hedge we assume that all yield curve moments are parallel, so the last term in this expression equals one. Follow the same steps as above, but also calculate the durations (only at the initial short interest rate) and the yield to maturities. Then

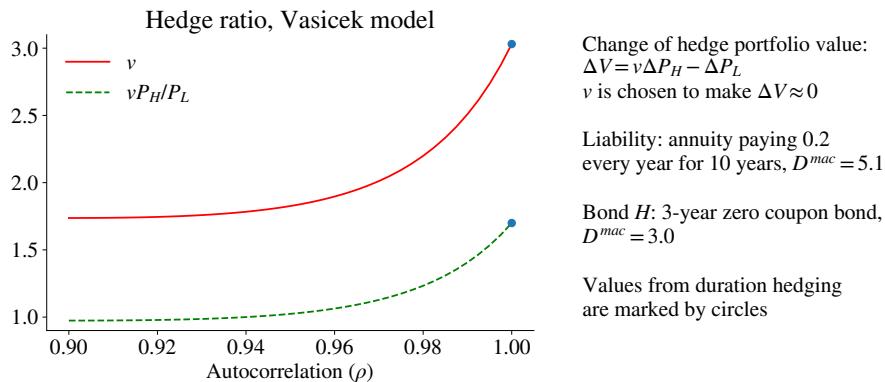


Figure 18.11: Hedge ratios in the Vasicek model

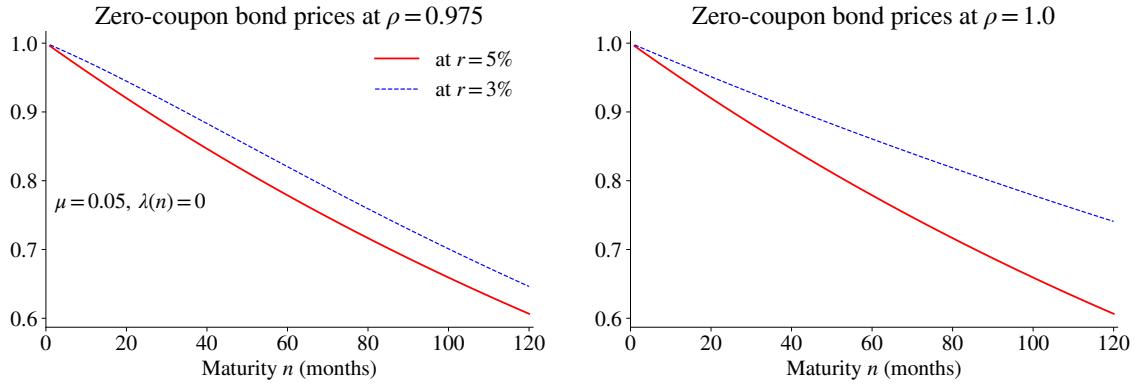


Figure 18.12: Bond price changes in the Vasicek model

calculate vP_H/P_L according to the equation above. The results are very similar to the easier approach discussed above.

Figure 18.11 gives an illustration. The hedge ratio v converges to the duration hedge ratio as the autocorrelation (ρ) in the short rate process (18.3) increases towards unity: in that limiting case all yield curve movements are indeed parallel. For lower values of the autocorrelation, the hedge ratio is lower. The main reason is that mean-reversion, that is, low autocorrelation makes interest rates on long-maturity bonds (here, the liability) move less than interest rates on short-maturity bonds (here, the hedge bonds). As a result, we need not invest so much into the (shorter maturity) hedge bond. See Figure 18.12 for how this result is affected by the autocorrelation ρ .

Notice, however, that all one-factor models (including the Vasicek model) imply that all yields are perfectly correlated (there is a common single driving force) and only fairly limited yield curve movements are possible. For instance, if the current short rate is low, then the yield curve must be upward-sloping. *Multi-factor models* overcome most of those limitations, for instance, the two-factor Nelson and Siegel (1987) model.

18.4 Interest Rates and Macroeconomics*

This section outlines several (not mutually exclusive) macroeconomic approaches to modelling the yield curve.

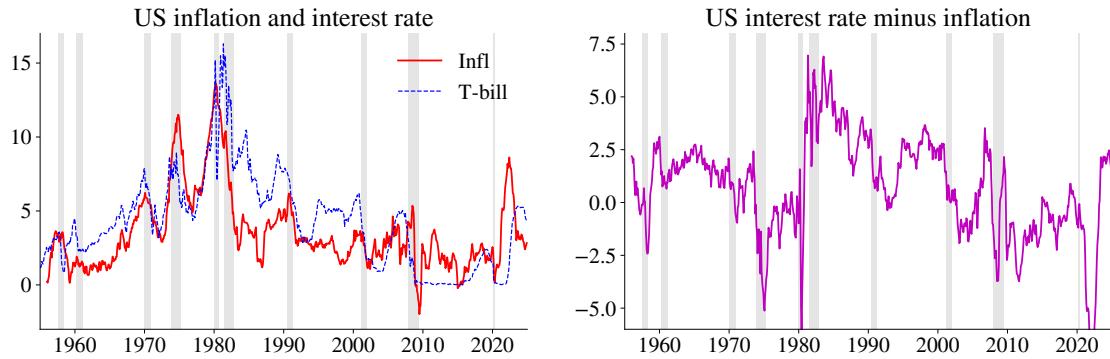


Figure 18.13: US inflation and 3-month interest rate

18.4.1 The Fisher Equation and Index-Linked Bonds

Let π_{t+n} be the annualised inflation rate over t to $t + n$, and $y_t^r(n)$ the *real interest rate* for the same period. The real interest rate is a return in terms of real purchasing power. Note the difference between a real interest rate and a traditional interest rate, where the latter (also called a nominal interest rate) is in terms of monetary units (dollars, say).

The *Fisher equation* says that the nominal interest rate includes compensation both for inflation expectations, $E_t \pi_{t+n}$, the real interest rate, $y_t^r(n)$, and possibly a constant risk premium, $\psi(n)$,

$$y_t(n) = E_t \pi_{t+n} + y_t^r(n) + \psi(n). \quad (18.11)$$

Example 18.12 (Fisher equation) Suppose the nominal interest rate is $y(n) = 0.07$, the real interest rate is $y^r(n) = 0.03$, and the nominal bond has no risk premium ($\psi = 0$), then the expected inflation is $E_t \pi_{t+n} = 0.04$.

The Fisher equation suggests a framework for analysing nominal interest rates in terms of real interest rates and inflation expectations. Information about real interest rates could possibly be elicited from *index-linked bonds*, that is, bonds which give automatic compensation for actual inflation.

Empirical results typically indicate non-trivial fluctuations in the real interest rate and risk premia (possibly driven by liquidity concerns), especially for short horizons. This holds also when inflation expectations as measured by surveys, are used as the dependent variable. It is therefore not straightforward to extract inflation expectations from nominal interest rates.

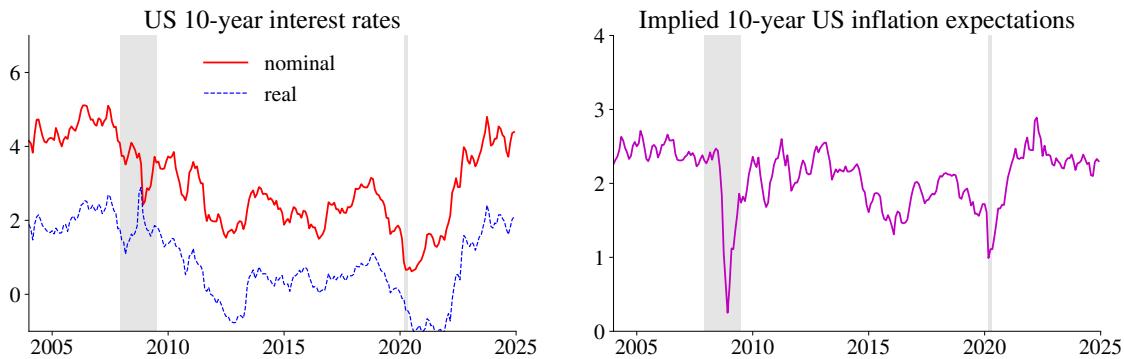


Figure 18.14: US nominal and real interest rates

Empirical Example 18.13 Figures 18.13–18.15 illustrate the relation between U.S. nominal and real interest rates, as well as inflation. A potential conclusion is that there are considerable movements in real interest rates (and/or liquidity premia on index linked bonds).

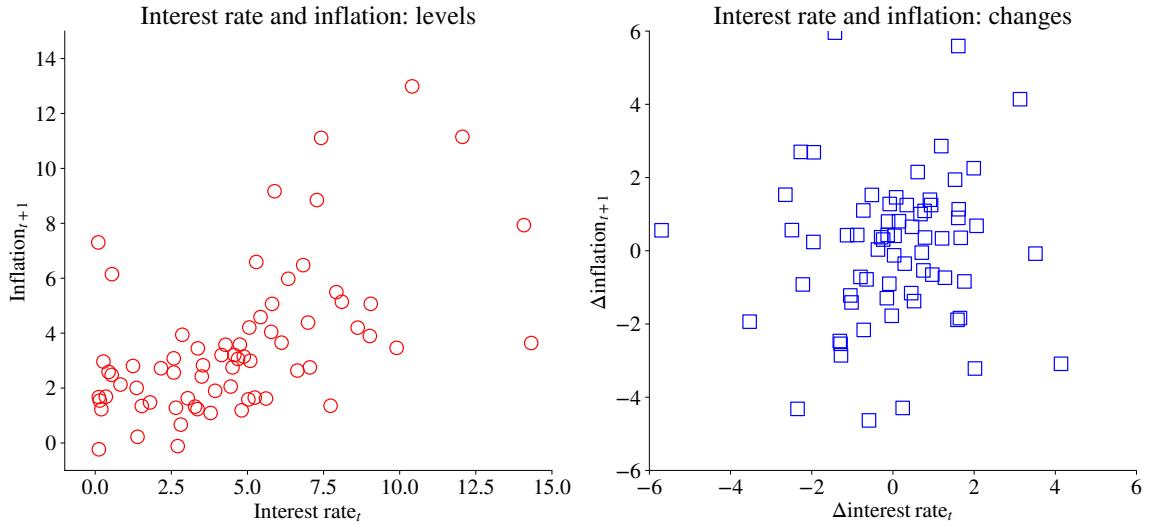
The Fisher equation is sometimes embedded in a macro model to construct a sophisticated model of the yield curve. This involves using macro theory/empirics to model how real interest rates and inflation expectations depend on the state of the economy.

18.4.2 The Expectations Hypothesis of Interest Rates

The expectations hypothesis of interest rates says that long interest rates equal an average of expected future short rates, possibly with a constant (across time, not maturities) risk premium as in (18.2). This can help interpreting yield curve changes around, for instance, interest rate hikes by a central bank. Suppose the central bank increases its policy rate, a short-maturity rate. The impact on longer rates depends of several factors.

First, one possibility is that only the very short interest rates change, and that all longer interest rates stay unchanged. This would happen if the policy move was well anticipated.

Second, another possibility is that long interest rates increase. Under the expectations hypothesis of interest rates, the interpretation is that the market now expects high short interest rates also in the future. That is, that the central bank will not reverse its policy action in the foreseeable future. If we are willing to assume that the real interest rate was not affected by the policy move, then one possible interpretation is that the central bank has received information about a long-lasting inflation pressure.



Sample: US 1-year interest rates and next-year inflation 1955–2024

Figure 18.15: US nominal interest rates and subsequent inflation

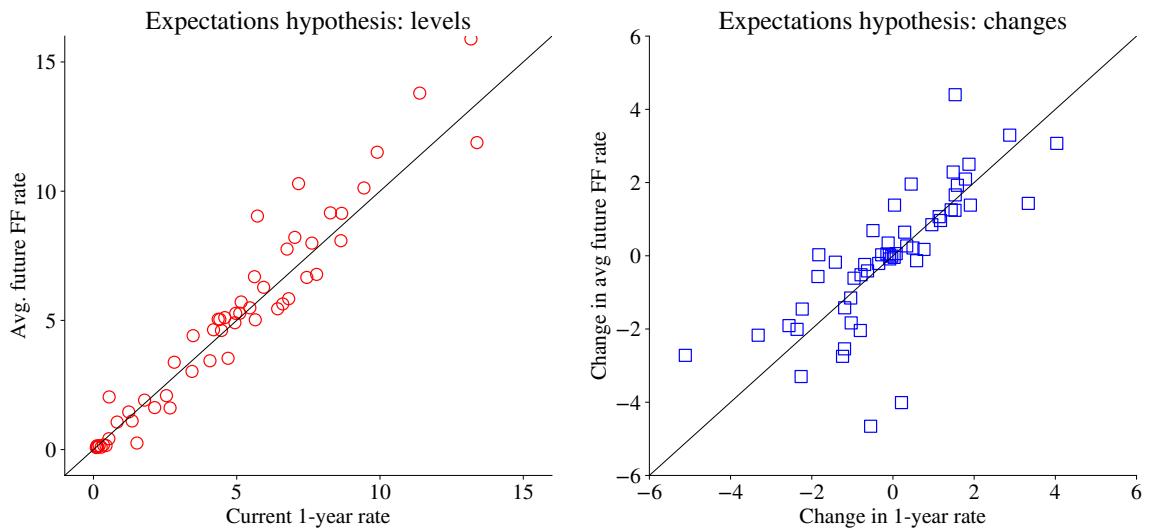
Third, and finally, short rates may increase, but long interest rates decrease. A common interpretation of this scenario is that the central bank has become more inflation averse. It therefore raises the policy rate to bring down inflation. If the market believes that it will succeed, then it follows that it will eventually be possible to lower interest rates (when inflation and inflation expectations are lower).

The expectations hypothesis has been tested many times, typically by an ex post linear regression (realized interest rates regressed on lagged forward rates). The results often give mild support to the hypothesis.

Empirical Example 18.14 *Figure 18.16 shows scatter plots of long interest rates and average future short rates—in levels and in changes. The evidence suggests some mild support of the expectations hypothesis.*

18.4.3 A New-Keynesian Model of Monetary Policy

Monetary policy is a crucial part of the macroeconomic setting, so it is important to understand how the policy is formed. It has not always been this way: there are long periods when many countries adopted a very simple (or so it seemed) monetary policy by pegging the currency to another currency. Macroeconomic policy was then synonymous with fiscal policy.



US 1-year interest rates and next-year average federal funds rate: 1970:01–2024:12

Figure 18.16: US 12-month interest and average federal funds rate (next 12 months)

Modern macro models are often smaller than the older macroeconomic models and they pay more attention to theory, the supply side of the economy and the role of expectations. These models try to capture the key elements in the way central banks (and most other observers) reason about the interaction between inflation, output, and monetary policy.

In these models, inflation depends on expected future inflation (some prices are set today for a long period and will therefore be affected by expectations about future costs and competitors' prices), lagged inflation, and a "Phillips effect" where an *output gap* (output less trend output) affects price setting via demand pressure. For instance, inflation (π_t) is often modelled as

$$\pi_t = \alpha E_t \pi_{t+1} + \beta \pi_{t-1} + \phi x_t + \varepsilon_{\pi t}, \quad (18.12)$$

where x_t is the output gap and $\varepsilon_{\pi t}$ can be interpreted as "cost push" shocks (wage demands, commodity price shocks). This equation can be said to represent the supply side of the economy and it is typically derived from a model where firms with some market power want to equate marginal revenues and marginal costs, but choose to change prices only gradually.

The demand side of the economy is modelled from consumers' savings decision, where

the trade-off between consumption today and tomorrow depends on the real interest rates. Simplifying by setting consumption equal to output we get something like the following equation for the output gap

$$x_t = x_{t-1} - \gamma(i_t - E_t \pi_{t+1}) + u_t, \quad (18.13)$$

where i_t is the nominal interest rate (set by the central bank) and u_t is a shock to demand. Note that the expected *real* interest rate affects demand (negatively).

In some cases, the real exchange rate is added to both (18.12) and (18.13), capturing price increases on imported goods and foreign demand for exports, respectively. The exchange rate is then linked to the rest of the model via an assumption of uncovered interest rate parity.

Some of the important features of this simple model are: (i) inflation expectations matter for today's inflation (think about wage inflation), (ii) the instrument for monetary policy, the short interest rate i_t , can ultimately affect inflation only via the output gap; (iii) it is the real, not the nominal, interest rate that matters for demand.

To make the model operational, two more things must be added: the monetary policy rule and a formalization of how expectations in (18.12)–(18.13) are formed.

It is common to assume that the central bank has some instrument rule like the “Taylor rule”

$$i_t = \theta_0 + 0.5x_t + 1.5\pi_t + v_t. \quad (18.14)$$

The residual v_t is a “monetary policy shock,” which picks up factors left out of the model, for instance, the central bank's concern for the banking sector or simply changes in the central bank's preferences.

Another approach to find a policy rule is to assume that the central bank has some loss function that it minimizes by choosing a policy rule. This loss function is often a weighted average of the variance of inflation and the variance of the output gap.

The expectations in (18.12)–(18.13) can be handled in many ways. The perhaps most straightforward way is to assume that the expectations about the future equal the current value of the same variable (a “random walk”). A more satisfactory way is to use survey data on inflation expectations. Finally, many model builders assume that expectations are “rational” (or “model consistent”) in the sense that the expectation equals the best guess we could do under the assumption that the model is correct. This latter approach typically requires a sophisticated way of solving the model, as the model both generates the best guesses and depends on them.

18.5 Forecasting Interest Rates*

The expectations hypothesis of interest suggests that current long rates can help predict future short rates. Empirically, this has some support. However, there are also a number of other forecasting approaches.

There is a two-way causality: inflation and the real economy affect monetary policy, and monetary policy can surely affect inflation and the real economy. This makes it difficult to analyse and forecast interest rates. However, for short term forecasting, the emphasis is typically on forecasting the next monetary policy move. Long run forecasting relies more on understanding the determinants of real interest rates and inflation, which depends on the general business cycle prospects, but also on the long run stance of monetary policy (“tough on inflation or not?”).

18.6 Risk Premia on Fixed Income Markets

There are many different types of risk premia on fixed income markets.

Nominal bonds are risky in real terms, and are therefore likely to carry *inflation risk premia*. Long bonds are risky because their market values fluctuate over time, so they probably have *term premia*. Corporate bonds and some government bonds (in particular, from developing countries) have *default risk premia*, depending on the risk for default. Interbank rates may be higher than T-bill of the same maturity for the same reason (see the TED spread, the spread between 3-month Libor and T-bill rates) and illiquid bonds may carry *liquidity premia* (see the spread between off-the run and on-the-run bonds).

Empirical Example 18.15 Figures 18.17–18.19 illustrate some U.S. data. In particular, there seems to be considerable (and business cycle related) default risk premia in the corporate sector, and also within the banking sector. in addition, the evidence on the on/off-the run interest rates suggests important liquidity risk premia, even across comparable bonds with the same issuer (the U.S. Treasury).

18.7 Appendix – Formal Derivation of the Vasicek Model*

Remark 18.16 This section uses a slightly different notation, namely a subscript_n to indicate the maturity and P to indicate a zero coupon bond price. For instance, y_{nt} for

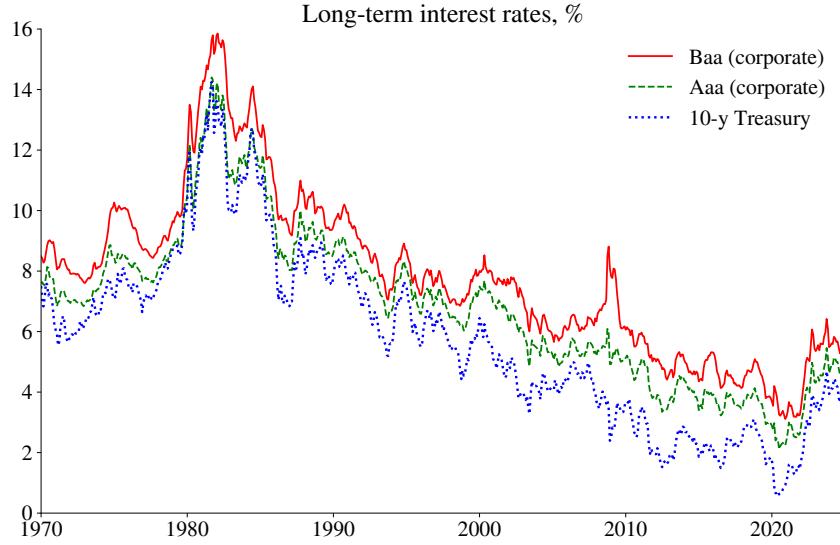


Figure 18.17: US interest rates

the n -period interest rate in t (same as $y_t(n)$ in the rest of this chapter) and P_{nt} and $P_{n-1,t+1}$ to indicate a bond price.

Write (18.6) as

$$y_{nt} = a_n + b_n r_t, \text{ where } a_n = A_n/n \text{ and } b_n = B_n/n. \quad (18.15)$$

The expressions for A_n and B_n will be derived below.

The price of an n -period zero coupon bond equals the cross-moment between the stochastic discount factor (SDF) and the value of the same bond next period (when it's an $n - 1$ -period bond)

$$P_{nt} = E_t e^{m_{t+1}} P_{n-1,t+1}, \quad (18.16)$$

where m_{t+1} is the *logarithm* of the stochastic discount factor $e^{m_{t+1}}$. Notice that this notation differs from some other chapters.

The *Vasicek model* assumes that the log SDF (m_{t+1}) is a linear function of r_t and an iid shock

$$-m_{t+1} = r_t + \gamma \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} \text{ is iid } N(0, \sigma^2) \text{ and} \quad (18.17)$$

$$r_{t+1} = (1 - \rho) \mu + \rho r_t + \varepsilon_{t+1}. \quad (18.18)$$

The short rate process is the same as in (18.3).

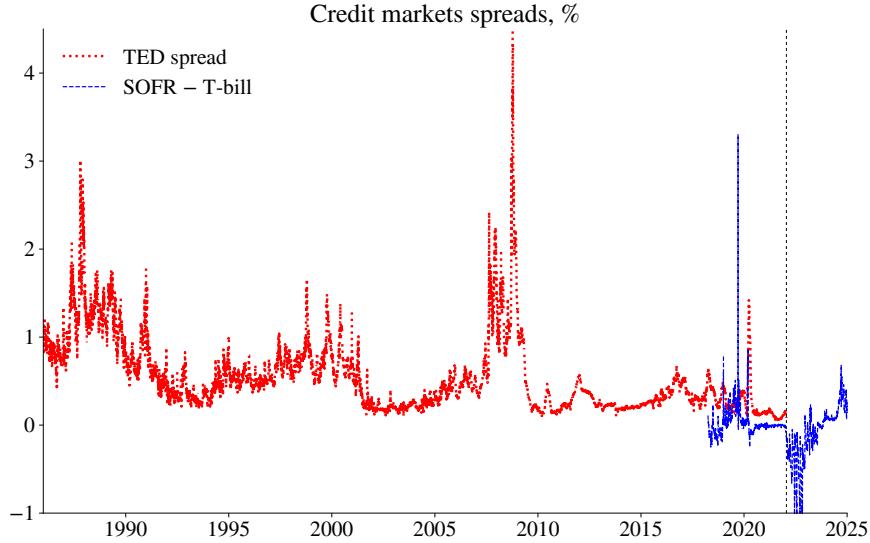


Figure 18.18: TED spread

Remark 18.17 If $x \sim N(\mu, \sigma^2)$, then $E e^x = e^{\mu + \sigma^2/2}$. Take logs to get $\ln E e^x = \mu + \sigma^2/2$.

The model values of (A_n, B_n) are found by using (a) $P_n = e^{-ny_n}$; (b) the proposed model (18.15); (c) the dynamics in (18.17)–(18.18) to calculate the logarithm of (18.16) as

$$p_{nt} = E_t(m_{t+1} + p_{n-1,t+1}) + \text{Var}_t(m_{t+1} + p_{n-1,t+1})/2, \quad (18.19)$$

where p_{nt} is the log bond price ($\ln P_{nt}$). This is an application of Remark 18.17 with $m + p$ playing the role of x . The result (see below for a proof) is that

$$B_n = 1 + \rho B_{n-1} \text{ and} \quad (18.20)$$

$$A_n = A_{n-1} + B_{n-1} (1 - \rho) \mu - (\gamma + B_{n-1})^2 \sigma^2 / 2, \quad (18.21)$$

where the recursion starts at $B_0 = 0$ and $A_0 = 0$. Notice that the expression for B_n is the same as in Remark 18.9, but that we have another expression for the A_n which involves both the mean μ and the volatility (risk) σ^2 .

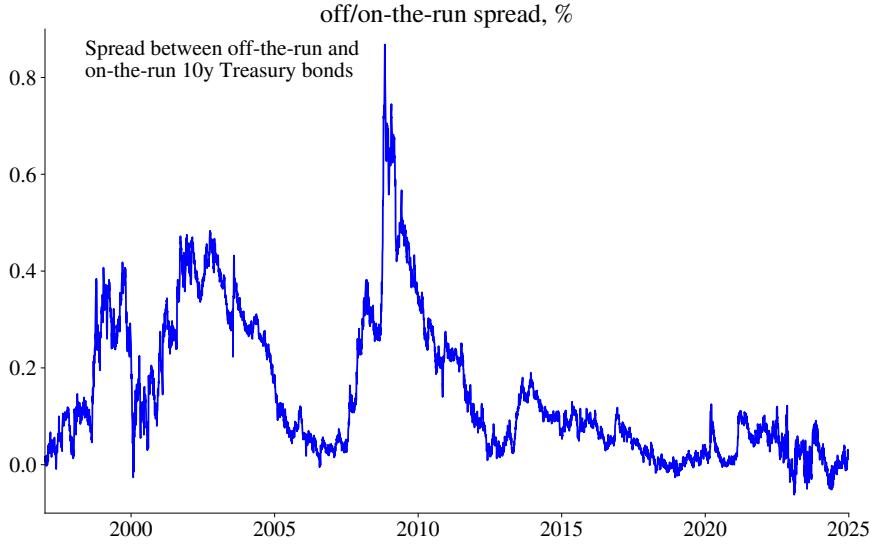


Figure 18.19: Off-the-run liquidity premium

Example 18.18 (A_n and B_n in the Vasicek model) (18.20 –(18.21)) give

$$B_0 = 0 \text{ and } A_0 = 0$$

$$B_1 = 1 \text{ and } A_1 = -\gamma^2 \sigma^2 / 2$$

$$B_2 = 1 + \rho \text{ and } A_2 = (1 - \rho) \mu - [\gamma^2 + (1 + \gamma)^2] \sigma^2 / 2.$$

Proof (of (18.20)–(18.21)) First, rewrite

$$\begin{aligned} m_{t+1} + p_{n-1,t+1} &= \underbrace{-r_t - \gamma \varepsilon_{t+1}}_{m_{t+1}} - \underbrace{A_{n-1} - B_{n-1} r_{t+1}}_{p_{n-1,t+1}} \\ &= -(1 + B_{n-1} \rho) r_t - (\gamma + B_{n-1}) \varepsilon_{t+1} - A_{n-1} - B_{n-1} (1 - \rho) \mu, \end{aligned}$$

where we use (18.18) to substitute for r_{t+1} . The conditional moments in (18.19) can then be calculated as

$$\begin{aligned} E_t(m_{t+1} + p_{n-1,t+1}) &= -(1 + B_{n-1} \rho) r_t - A_{n-1} - B_{n-1} (1 - \rho) \mu \\ \text{Var}_t(m_{t+1} + p_{n-1,t+1}) &= (\gamma + B_{n-1})^2 \sigma^2. \end{aligned}$$

Second, substitute $p_{nt} = -A_n - B_n r_t$ on the LHS of (18.19) and plug in the conditional moments from above on the RHS

$$-A_n - B_n r_t = -(1 + B_{n-1} \rho) r_t - A_{n-1} - B_{n-1} (1 - \rho) \mu + (\gamma + B_{n-1})^2 \sigma^2 / 2.$$

This equation must always hold (for any value of r_t : match coefficients of r_t and the “constant” to get (18.20)–(18.21). \square

Chapter 19

Basic Properties of Options

19.1 Derivatives

Derivatives are assets whose payoffs depend on an underlying asset (for instance, shares of a company). The most common derivatives are futures contracts (or similarly, forward contracts) and options. Options are sometimes written on (depend on) the price of a futures contract, not the underlying directly. See Figure 19.1.

Derivatives have zero net supply, so a contract must be issued (a short position) by someone for an investor to be able to buy it (a long position). For that reason, gains and losses on derivatives markets sum to zero.

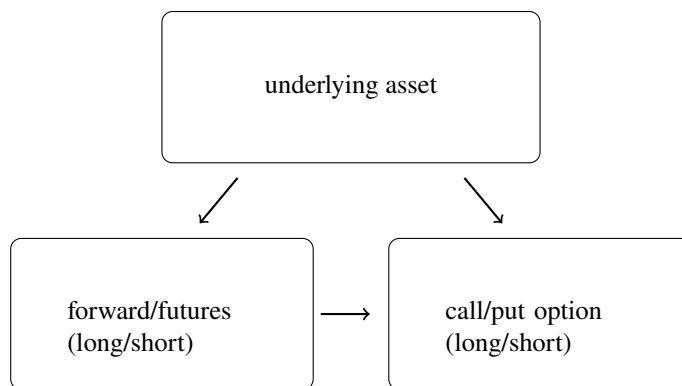


Figure 19.1: Derivatives on an underlying asset

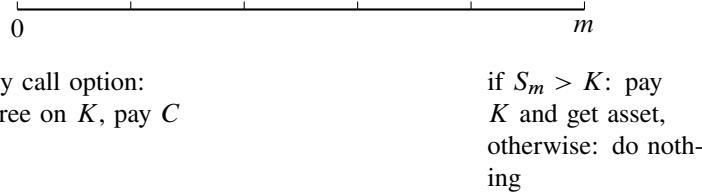


Figure 19.2: Timing convention of a European call option contract

19.2 Introduction to Options

Remark 19.1 (*On the notation*) *The notation here is typically kept short. The current period is assumed to be 0 and the derivative expires m years later. The current price of the underlying is denoted S (rather than S_0), the forward price according to a contract agreed on now and expiring in m is F (rather than $F_0(m)$) and the continuously compounded interest between 0 and m is y (rather than $y_0(m)$). However, to avoid confusion, the price of the underlying asset at expiration is denoted S_m . The more precise notation is used only when strictly needed.*

19.2.1 Definition of European Calls and Puts

A European *call* option contract traded in period 0 stipulates that the owner of the contract has the *right* (but not the obligation) to *buy* one unit of the underlying asset (“exercise the option”) from the issuer of the option on the expiration date m at the strike price K . Compare with a forward contract where the owner *must* exercise. See Figure 19.2 for the timing convention.

The analysis here normalizes all contracts to one unit of the underlying. A simple rescaling is needed for an application to typical contracts, which may be for many more units.

To the owner of a call option, the payoff at expiration is either zero (if the owner does not exercise) or the value the underlying asset S_m minus the strike price K (if the owner exercises). For a rational investor (who only exercises if $S_m \geq K$), the payoff is thus

$$\text{call payoff}_m = \max(0, S_m - K). \quad (19.1)$$

Clearly, an owner of a call option benefits from a high price of the underlying asset.

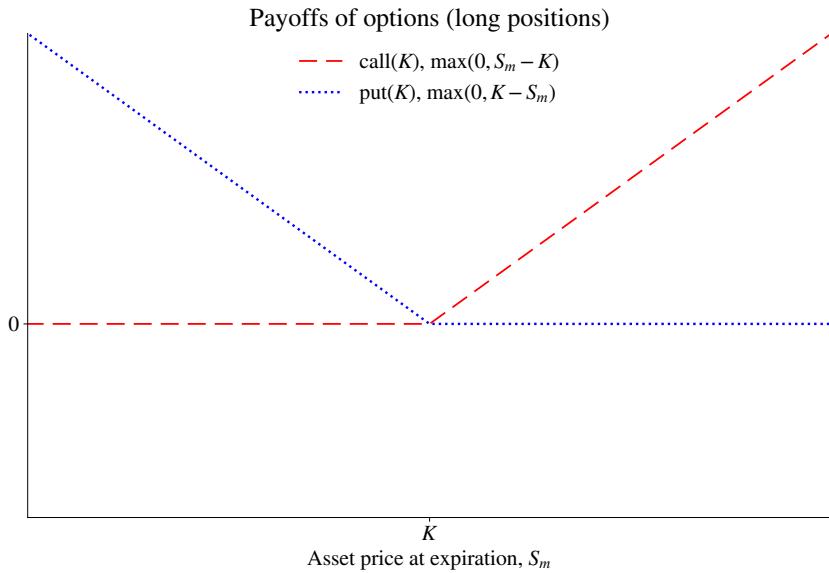


Figure 19.3: Payoffs of options, long positions

Example 19.2 (Call option payoffs) With $K = 5$ we have

S_m	Exercise	Payoff
4.5	no	0
5.5	yes	$5.5 - 5 = 0.5$

The profit at expiration is thus

$$\text{call profit}_m = \text{call payoff}_m - e^{my} C, \quad (19.2)$$

where C is the call price, typically paid period 0. (To simplify the notation, the time subscript on C is suppressed, but we could write C_0 when required.) The e^{my} factor captures the capital cost of paying the option price already on the trade date (think: borrow C in period 0 and repay with interest, $e^{my}C$, on the expiration date). Time to expiration m is measured in years, since interest rates are annualized rates.

See Figure 19.4 for an illustration. Notice that the price of the option (C) is always paid, irrespective of whether the option is exercised or not.

Remark 19.3 (In-the-money*) An option that would be profitable to exercise is called in-the-money; an option that would be unprofitable to exercise is called out-of-the-money—and an option that would just break even is called at-the-money.

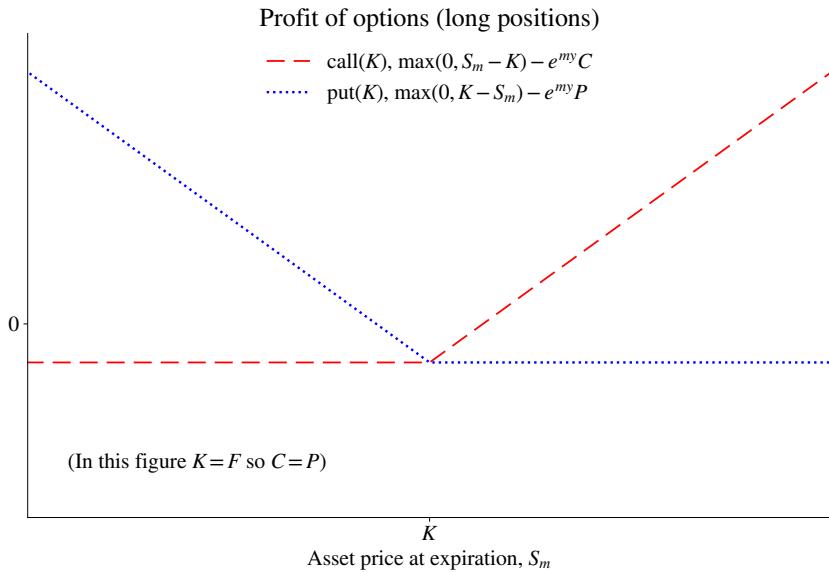


Figure 19.4: Profit of options, long positions

The payoff of the issuer is the mirror image of the owner's payoff: the owner's gain is the issuer's loss: a *zero sum game*. See Figures 19.3 for an illustration. This zero sum game property is true both for the payoff at exercise as well as the for the profit.

Remark 19.4 (*Margin requirements**) *A buyer of an option does not have to post any margin, but an issuer typically does. The reason is that a default of the issuer could create a loss for the option owner (if the option is worth exercising). In contrast, a default of the owner cannot create a loss for the issuer.*

A *put* option instead gives the owner of the contract the right to *sell* one unit of the underlying asset at the strike price K . The put price is here denoted by P . An owner of a put option benefits from a low price of the underlying asset (buy the asset cheaply and exercise the right to sell for K). The payoff is

$$\text{put payoff}_m = \max (0, K - S_m) . \quad (19.3)$$

Example 19.5 (*Put option payoffs*) With $K = 5$ we have

S_m	Exercise	Payoff
4.5	yes	$5 - 4.5 = 0.5$
5.5	no	0

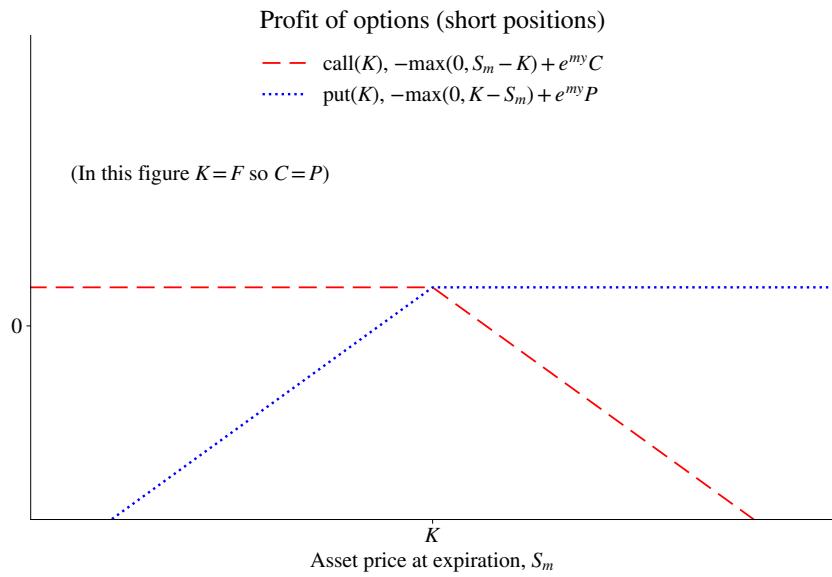


Figure 19.5: Profit of options, short positions

Remark 19.6 (*Which options are traded?*) *Most of the trade is in out-of-the-money options (high strike prices for the calls and low strike prices for the puts). Also, most of the trade happens close to the expiration date, and there is a seasonality pattern related to rolling over the investment from other (expired) options. Figure 19.6 shows how the trading volume at CBOE has developed over time. The volume seems to correlate with the general business cycle movements. The ratio of traded put contracts to traded call contracts in Figure 19.6 is sometimes used to gauge market nervousness. The idea is that investors will demand put contracts if they want to insure against a stock market decline.*

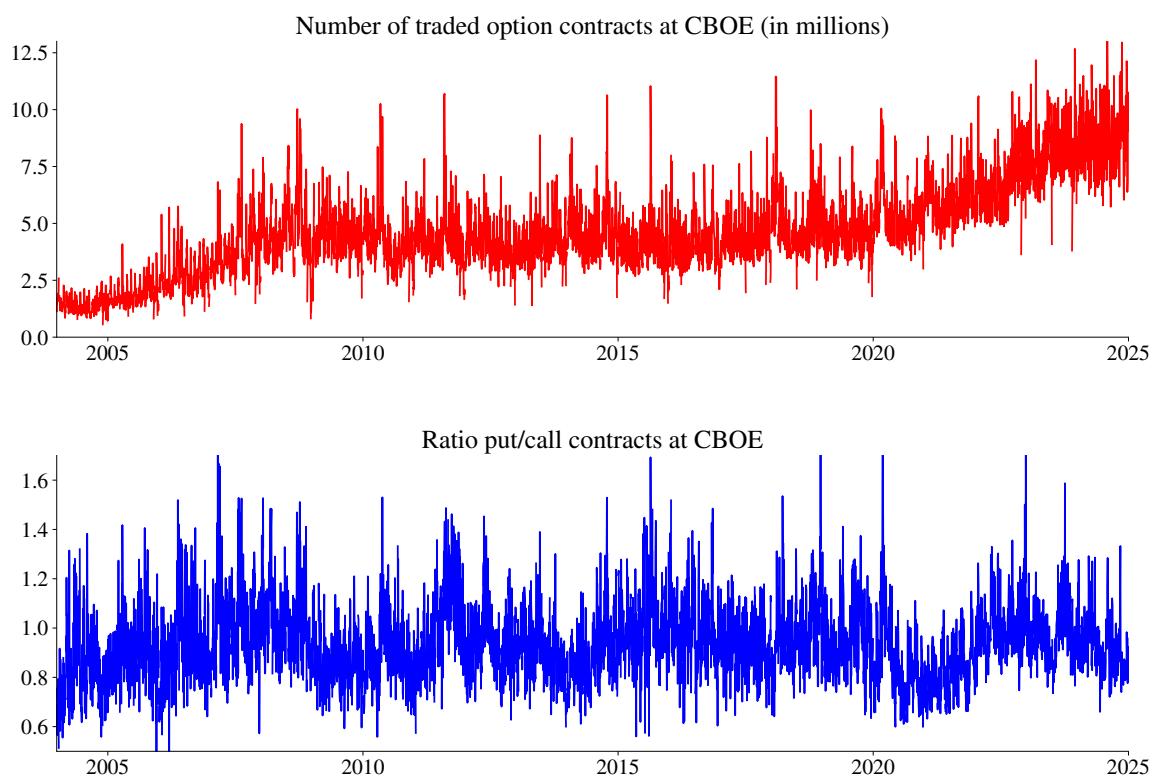


Figure 19.6: Option trade volume

19.2.2 Options Are Risky Assets

The net return on a long position in a European call option is

$$\text{return on call}_m = \frac{\max(0, S_m - K)}{C} - 1, \quad (19.4)$$

where C is the call option price. Whenever the option isn't exercised ($S_m < K$) or exercised but with a zero payoff ($S_m = K$), the whole investment is lost (and the return is -100%). In contrast, when the option is exercised ($S_m > K$), then the return can potentially become very large.

It is clear that the option return (19.4) cannot be normally (or even lognormally) distributed: the density function has a spike at -100% (whose probability mass is the same as the probability of $S_m \leq K$). This means that we cannot motivate “mean-variance” pricing of options by referring to a normal distribution of the return. (This does not rule out mean-variance pricing, which could be motivated by, for instance, mean-variance preferences.)

19.3 Financial Engineering

This section discusses the properties of some specific portfolios of options, forwards and the underlying asset.

19.3.1 Replicating a Forward

Options markets are often very liquid—and are therefore useful for constructing replicating portfolios. Let “call(K) - put(K)” be short hand notation for portfolio which is long one call option with strike price K and short one put option with the same strike price. When $K = F$, then this portfolio replicates a forward contract, so it is a synthetic forward. Clearly, we can then replicate a short position in a forward contract by selling such a portfolio. See Figure 19.7.

Example 19.7 (*Payoff of a synthetic forward*) With $K = 5$, we have the differences of the payoffs in Examples 19.2 and 19.5, that is,

S_m	Exercise call	Call payoff	Exercise put	Put payoff, short	Total Payoff
4.5	no	0	yes	$-(5 - 4.5)$	-0.5
5.5	yes	$5.5 - 5$	no	0	0.5

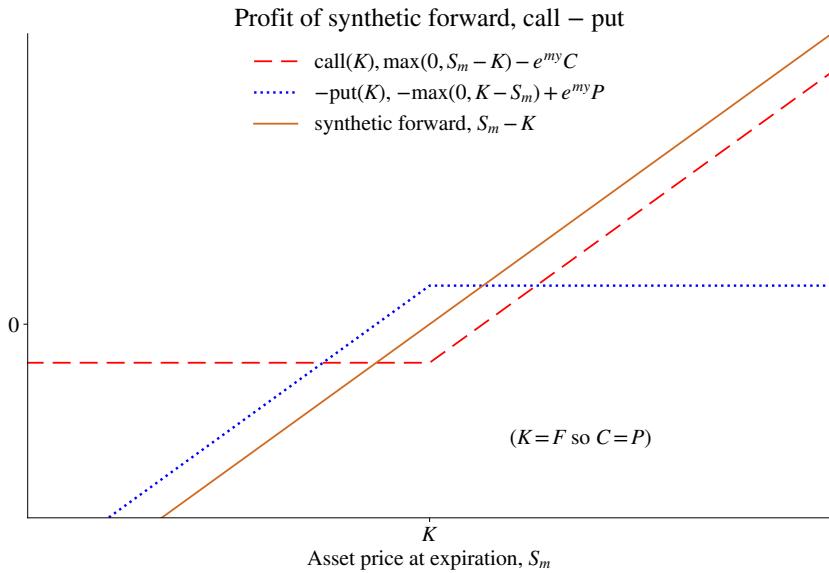


Figure 19.7: Profit of an option portfolio that replicates a forward contract

To get the profit, subtract the difference of the call and put prices from the total payoff.

19.3.2 Portfolio Insurance

A *protective put* is a combination of a put and a position in the underlying asset. This allows the owner to capture the upside of the price movement (of the underlying), at the same time as insuring against the downside. This is indeed very similar to just buying a call option. See Figure 19.8.

19.3.3 Betting on Large Changes

An option is a bet on a change in a specific direction. Option portfolios can be constructed to instead make a bet on a large change in either direction (that is, high volatility): a *straddle* is $\text{call}(K) + \text{put}(K)$, and a *strangle* is $\text{call}(K_2) + \text{put}(K_1)$ where $K_1 < K_2$ where K_1 and K_2 are two different strike prices. See Figure 19.9.

Example 19.8 (Payoff of a straddle) With $K = 5$, we have the sum of the payoffs in

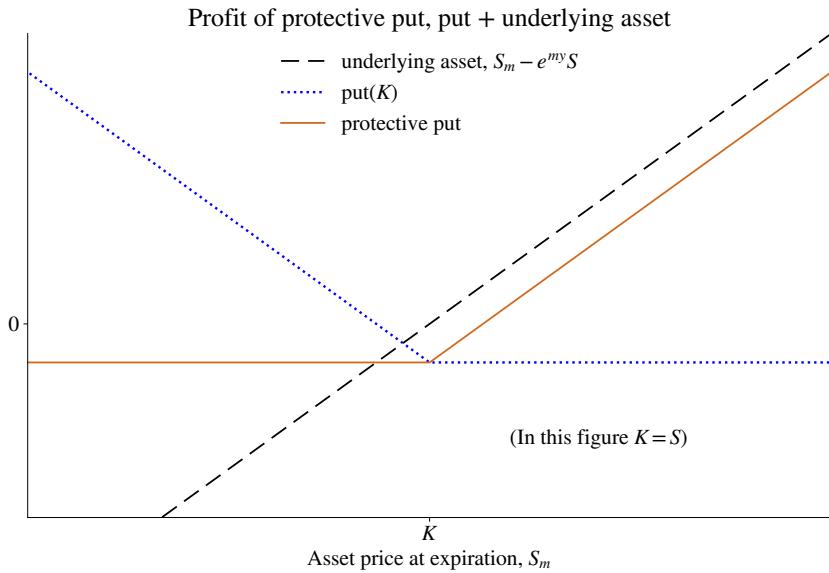


Figure 19.8: Profit of an option portfolio that insures the underlying asset

Examples 19.2 and 19.5, that is,

S_m	Exercise call	Payoff from call	Exercise put	Payoff from put	Total Payoff
4.5	no	0	yes	$5 - 4.5$	0.5
5.5	yes	$5.5 - 5$	no	0	0.5

To get the profit, subtract the sum of the call and put prices.

19.3.4 Putting a Collar on Losses and Gains

A *collared stock* is a combination of the underlying asset, a put with a low strike price (K_1) and a short call with a high strike price (K_2). This portfolio has a profit that increases one-for-one with the underlying asset as long as it is between K_1 and K_2 . The losses for values of the underlying below K_1 are limited (by the put), and the gains for values above K_2 are also capped (by the short call). See Figure 19.10.

19.3.5 Betting on a Large Price Decrease

A variation on the synthetic short forward is the *collar*: $-\text{call}(K_2) + \text{put}(K_1)$ where $K_1 < K_2$. It also looks like a short position in a forward contract, except that the payoff

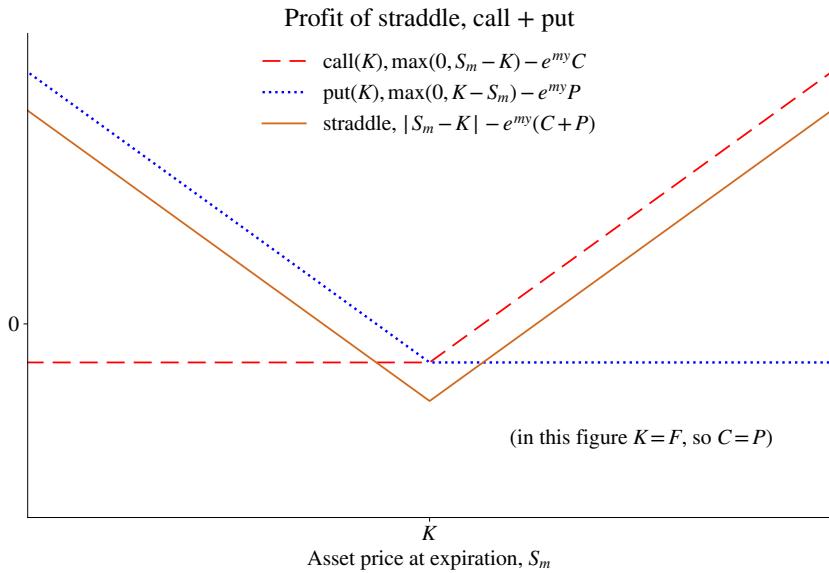


Figure 19.9: Profit of an option portfolio than bets on volatility

is flat between the strike prices. Clearly, this is betting on a large price decrease. Selling a collar (or *reversal*) is instead a bet on a large price increase.

A collar (reversal) can be used to hedge a long (short) position in the underlying asset, except that there is no hedge between the strike prices. It provides insurance outside the strike prices. See Figure 19.11.

19.3.6 Betting On a Small Price Increase

To bet on a small increase in the price of the underlying asset we can use a *bull spread*: $\text{call}(K_1) - \text{call}(K_2)$ where $K_1 < K_2$. This portfolio has flat payoffs outside the strike prices, but a payoff that increases with the underlying asset between them. Selling a bull spread creates a *bear spread*, which is a bet on a small decrease of the underlying price. (These spreads can also be constructed by combining puts.) See Figure 19.11.

19.4 Prices of Options

Much of the subsequent analysis will focus on understanding how options are priced (before the expiration date), that is, how the C and P are determined.

As an example, Figure 19.12 shows results from a particular model for option pricing (the Black-Scholes model). Before creating such models, we will first discuss (*a*) how put

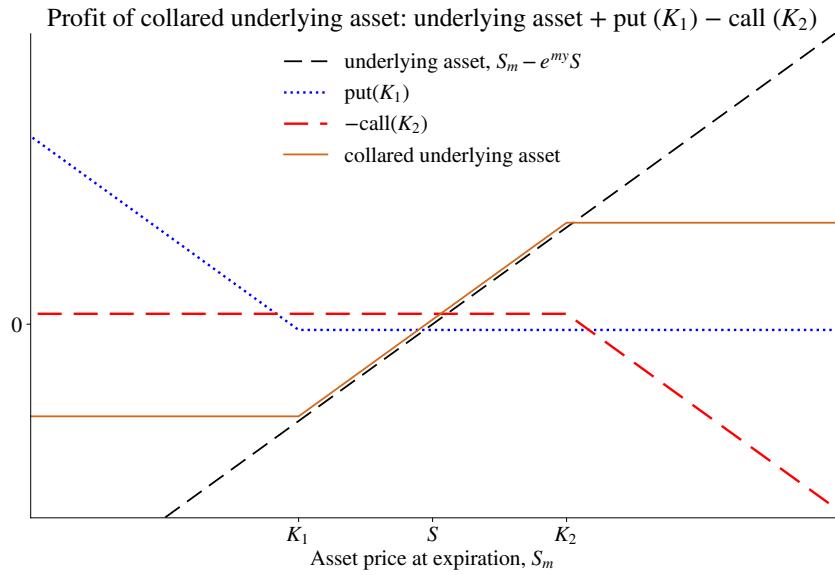


Figure 19.10: Profit of a collared underlying asset

and call prices are related; (b) the general effects of the strike price K (which decrease the call price) and volatility σ (which tend to increase call and put prices); (c) and also derive (no-arbitrage) bounds that option prices have to obey.

Option prices are often decomposed into the *intrinsic value* (what you get if you could get rid of the option today by exercising or burning it) and the *time value* (the rest). Clearly, the time value converges to zero as time approaches expiration.

19.5 Put-Call Parity for European Options

There is a tight link between European call and put prices. If you know one of them (and the forward price), then you can easily calculate what the other must be. The following proposition is more precise.

Proposition 19.9 (*Put-call parity for European options*) *The put-call parity for European options is*

$$C - P = e^{-my}(F - K), \quad (19.5)$$

where $e^{-my}(F - K)$ is the present value of the forward price minus the strike price.

Time subscripts and indicators of time to expiration have been omitted simplify the notation. The parity holds irrespective of whether the underlying asset has dividends or

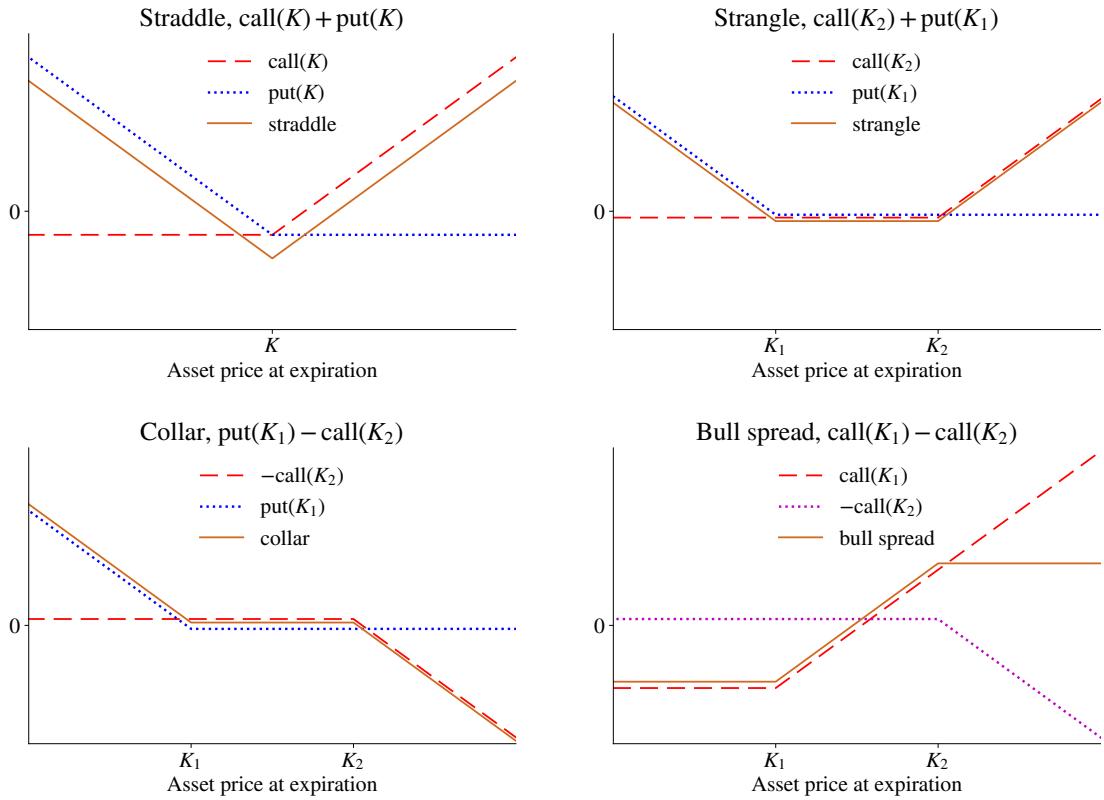


Figure 19.11: Profits of option portfolios

not (since the expression uses the forward price).

The practical importance of the proposition is that it shows how to use two assets to replicate a third asset. For instance, we can combine a call option (with strike price $K = F$) and a forward contract to replicate a put option, or buy a call and sell a put (with strike price $K = F$) to replicate a forward contract. Transaction costs can cause (relatively small) deviations from the parity condition. See Figure 19.13 for an illustration. Also, see Hull (2022) 11 and McDonald (2014) 11–12 for more detailed treatments.

Example 19.10 (Put-call parity) Let $S = 42, m = 1/2, y = 5\%, K = 38$. If the underlying asset has no dividends, then $F = e^{0.5 \times 0.05} 42 = 43.06$. With $C = 5.5$, (19.5) gives

$$5.5 - P = e^{-0.5 \times 0.05} (43.06 - 38) \text{ or } P \approx 0.56.$$

Proof (of Proposition 19.9) Portfolio A: buy one call option and sell one put option, both with the strike price K , at the cost $C - P$. This will with certainty give $S_m - K$ at maturity (since the call *or* the put will be exercised). Portfolio B: enter a forward contract

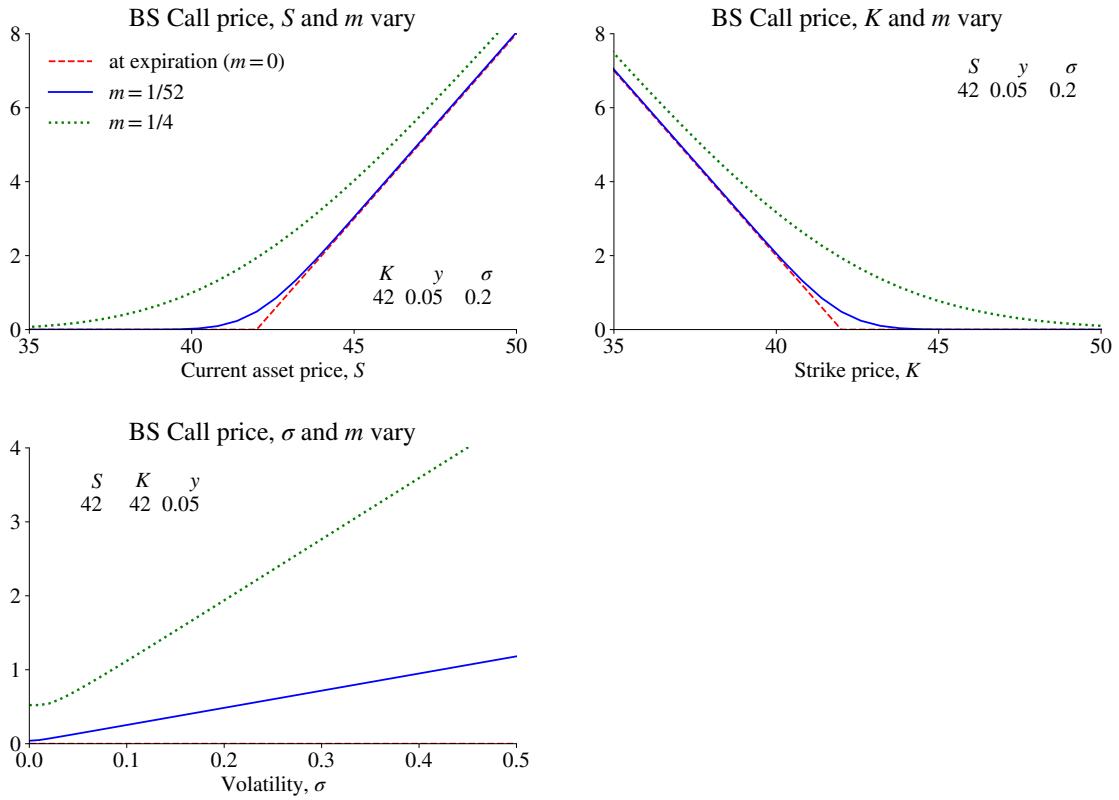


Figure 19.12: Call option prices (from the Black-Scholes model)

and put $e^{-my}(F - K)$ in the bank (your cost). At expiration, get $S_m - F$ from the forward contract plus the $F - K$ that you have in the bank: $S_m - K$. Since the two portfolios give the same at expiration, they must have the same costs today. \square

Example 19.11 (*Trading on deviations from the put-call parity*) Assume the same numbers as in Example 19.10, except that $P = 1$. Buying a call, selling a put and issuing a forward then costs $C - P = 4.5$ in $t = 0$. To finance this, we borrow and pay back $e^{0.5 \times 0.05} 4.5 = 4.61$ at expiration. The options and forwards together give $F - K = 43.06 - 38 = 5.06$ for sure at expiration. There is thus a risk-free profit. (With $P = 0.56$ there is not.)

This formula is very general, but a few special cases are of particular interest. First, when the underlying asset pays no dividends, then (19.5) together with the forward-spot parity give

$$C - P = S - e^{-my} K \text{ if no dividends.} \quad (19.6)$$

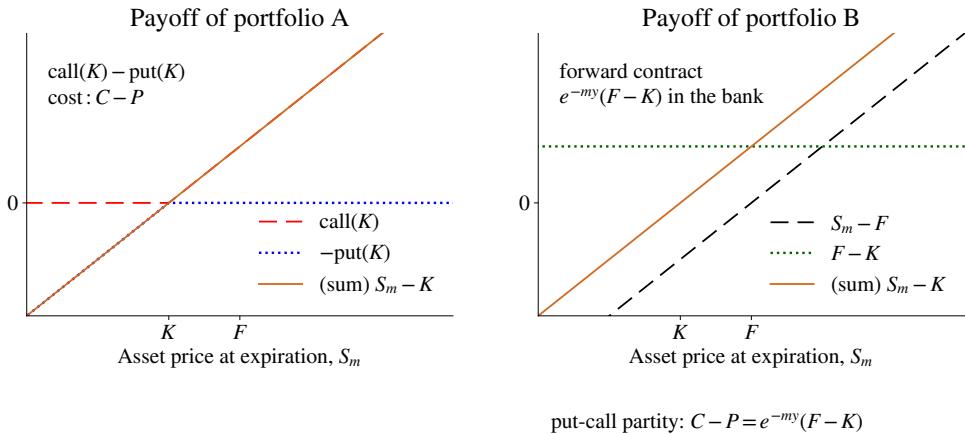


Figure 19.13: Put-call parity

Second, with dividends we get

$$C - P = S - \sum_{i=1}^n e^{-m_i y(m_i)} D_i - e^{-my} K \text{ if dividends,} \quad (19.7)$$

$$C - P = S e^{-m\delta} - e^{-my} K \text{ if continuous dividend rate } \delta. \quad (19.8)$$

19.5.1 Put-Call Parity and Synthetic Replications*

The following remarks provides details on how two assets can be used to replicate a third—since they are all tied together by the put-call parity.

Remark 19.12 (*Synthetic forward*) *Buy one call and sell one put at a strike price that equals the forward price. By (19.5), the cost of this portfolio is zero. At expiration, it will give one unit of the underlying, at the cost K . Just like a forward contract. See Figure 19.14.*

Remark 19.13 (*Synthetic call option*) *Buy one forward and one put with strike price $K = F$. By (19.5) this has the price C . If $S_m < K$ (at expiration), then the forward pays off $S_m - F$ and the put option $K - S_m$. Since $K = F$, the sum is zero. Instead, if $S_m > K$, then the forward pays off $S_m - F$ and the put nothing. In either case, this is just like a call option with strike price K . See Figure 19.14.*

Remark 19.14 (*Synthetic put option*) *Buy one call with strike price $K = F$ and sell one forward. By (19.5), this has the price P . If $S_m < K$ (at expiration) then the call pays off nothing and the short forward $-(S_m - F)$. Since $K = F$, the sum is $K - S_m$. Instead, if*

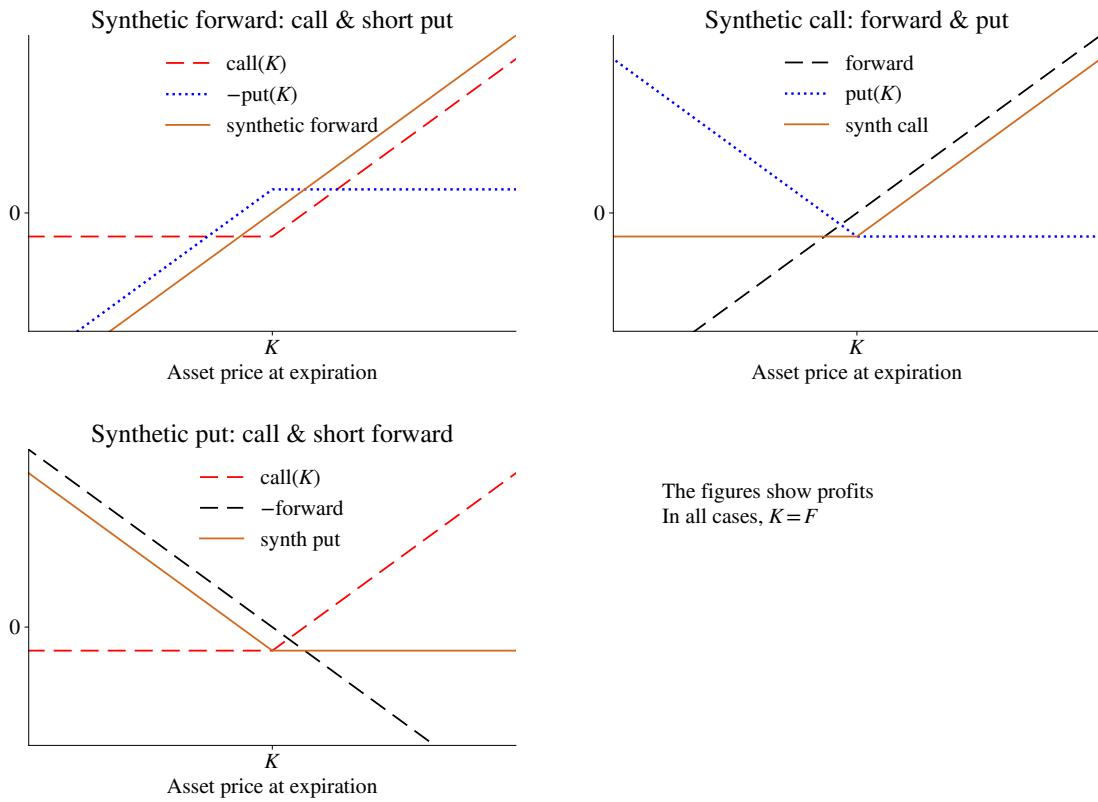


Figure 19.14: Synthetic replication

$S_m > K$, then the call pays off $S_m - K$ and the short forward $-(S_m - F)$, which sums to zero. In either case, this is just like a put option with strike price K . See Figure 19.14.

19.6 Definition of American Calls and Puts

An American option is similar to a European option, except that it *can be exercised on any day* before or on the expiration date. This means that an American option has more rights than a European option and is therefore worth at least as much

$$C_A \geq C_E \text{ and } P_A \geq P_E, \quad (19.9)$$

where we use subscripts to distinguish between American (A) and European (E) options.

You would only consider exercising an American call option if its profitable ($S > K$) so the immediate payoff is $S - K$, where S should be understood as the current price of the underlying. Instead, if you keep the option, then you know that it always worth 0 or more.

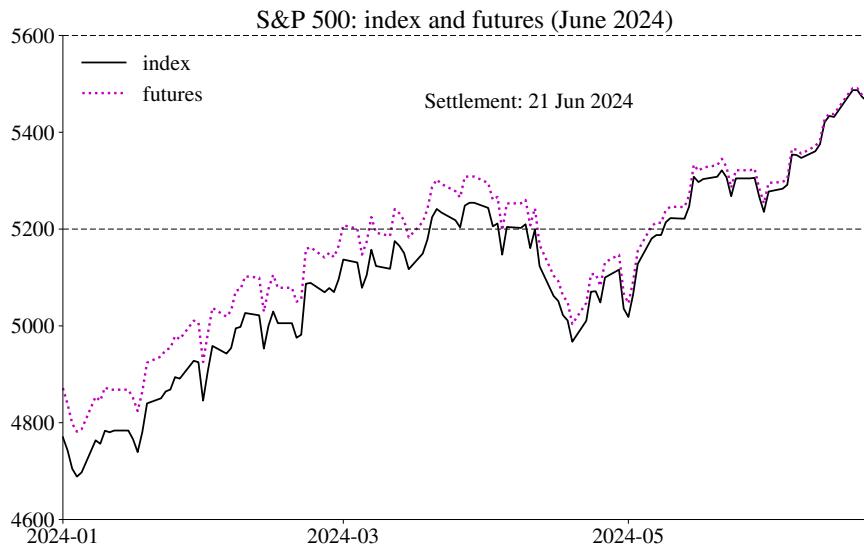


Figure 19.15: S&P 500 index level and futures

A similar logic applies to an American put option. This means that the option prices must (at any point in time) obey

$$\begin{aligned} C_A &\geq \max(0, S - K) \\ P_A &\geq \max(0, K - S). \end{aligned} \tag{19.10}$$

The right hand sides are called the “intrinsic values,” which can be thought of as what you get if you decide to get rid of the option today (exercise or burn it).

Empirical Example 19.15 Figures 19.15 and 19.16 provide an example of how the futures price (on S&P 500), the intrinsic value of the option and the option price developed over six months. Notice how the futures price converges to the index level at expiration of the futures. Before it can deviate because of delayed payment (+) and no part in dividend payments (-). Also notice that even options with zero intrinsic value can have a fairly high option price (time value)—at least if the time to expiration is long, but it converges to zero as the expiration date gets closer.

There is no put-call parity for American options. However, pricing bounds (based on the values of European options) can be derived.

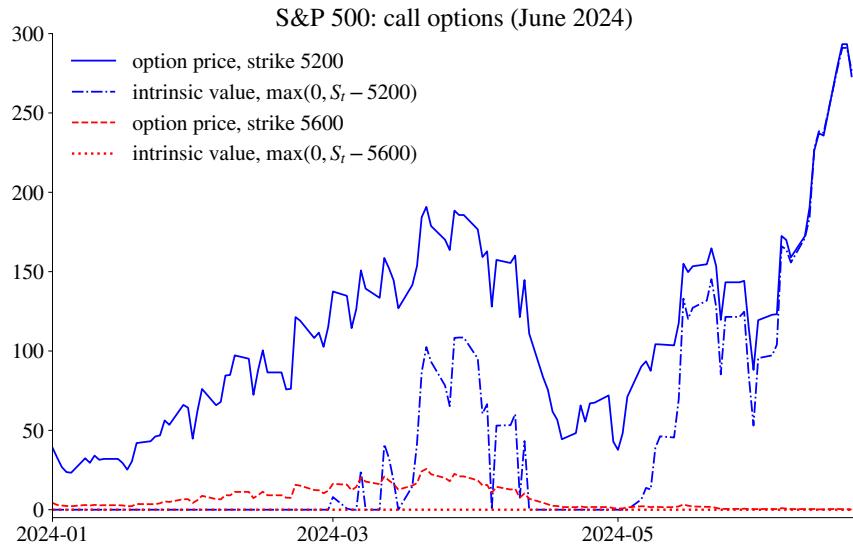


Figure 19.16: S&P 500 options

Remark 19.16 (*Put-call, American option, no dividend*) For an American option on an asset without dividends, the put price must be inside the interval

$$\underbrace{C_A - S + e^{-my} K}_{P_E} \leq P_A \leq \underbrace{C_A}_{C_E} - S + K. \quad (19.11)$$

See Hull (2022) 11 and McDonald (2014) 11 Appendix A.

19.7 Basic Properties of Option Prices

Options prices depend on many things, but there are some fairly general results, which we discuss here.

First, *call option prices are decreasing in the strike price*, while put options prices are increasing in the strike price, see Figure 19.17. The intuition is illustrated in Figure 19.18 which illustrates the perceived (by the market) distribution of the asset price at expiration. Notice that a higher strike price means that an owner of a call option will have to pay more in case of exercise—and there is also a lower chance of exercise.

Actually, it can be shown the call option price is decreasing in the strike price, but slower than the strike price itself, but that the curve flattens out at high strike prices. That

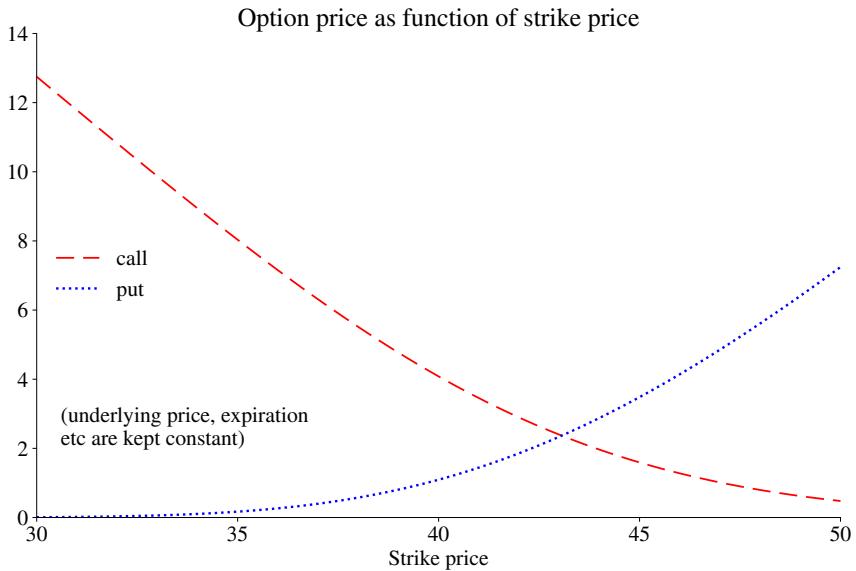


Figure 19.17: Option price as a function of the strike price

is, (if the derivatives exist) we have

$$-1 \leq dC(K)/dK \leq 0 \text{ and } dC^2(K)/dK^2 \geq 0. \quad (19.12)$$

For a put option, we instead have that $0 \leq dC(K)/dK \leq 1$. See McDonald (2014) 11 for proofs.

Second, both *call and put option prices are typically increasing in the (perceived) uncertainty* of the future price of the underlying asset, see Figure 19.19. The intuition is illustrated in Figure 19.20, which shows that a wider dispersion of the distribution increases the probability of a really high price of the underlying asset (although the figure is constructed to have the same probability of exercise in the two cases). Of course, it also increases the probability of a really low asset price, but that is of no concern since the call option payoff is bounded from below (at zero).

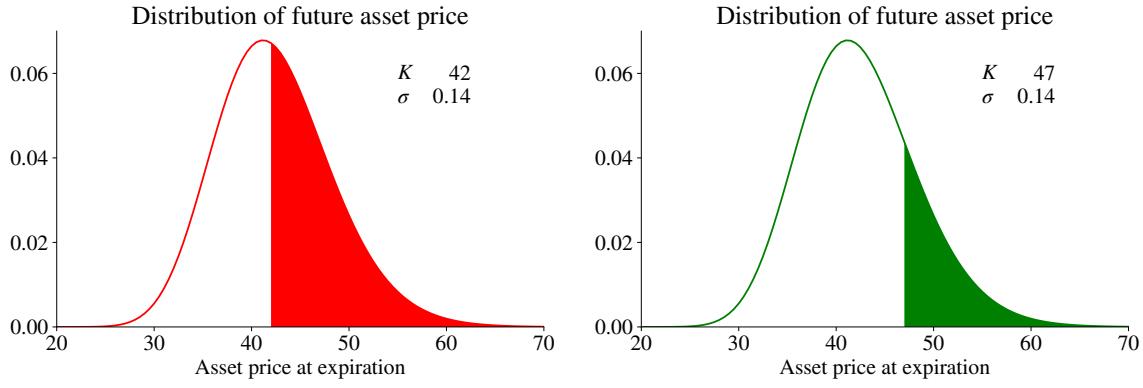


Figure 19.18: Distribution of future asset price

19.8 Pricing Bounds and Convexity

19.8.1 Pricing Bounds for (European and American) Call Options

The prices of call options must satisfy the following restrictions

$$C \leq e^{-my} F \leq S \quad (19.13)$$

$$0 \leq C \quad (19.14)$$

$$e^{-my}(F - K) \leq C. \quad (19.15)$$

These bounds hold for both American or European call options (we here use C to denote both of them.)

The motivations are basically as follows (the intuition is based on European options, but the results extend to American options as well). First, a call option with a zero strike price ($K = 0$) would be the same as owning a prepaid forward contract (which is worth as much or less than the underlying asset). Whenever the strike price is higher, the call price is lower. Second, the call option gives rights, not obligations: its price value cannot be negative. Third, the lowest possible value of a put option is zero, so the put-call parity (19.5) immediately gives that the call price must exceed the present value of $F - K$. (See below for an alternative proof.) Transaction costs can cause (relatively small) failures of the bounds.

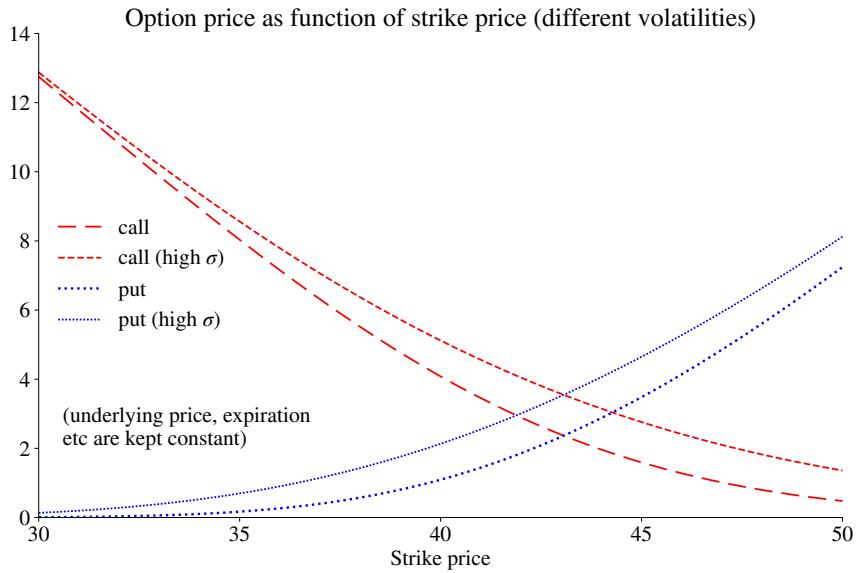


Figure 19.19: Option price as a function of the strike price

Combining the bounds, we get

$$C \leq e^{-my} F \leq S \quad (19.16)$$

$$C \geq \max[0, e^{-my}(F - K)]. \quad (19.17)$$

In particular, for a financial asset without dividends (until expiration of the option), we have $\max(0, S - e^{-my} K) \leq C \leq S$. See Figure 19.21 for an illustrations.

The pricing bounds are typically very wide, so they are of little importance in determining a fair option price. However, they may be helpful in checking data and also as a sanity check of a pricing model.

Example 19.17 (*Pricing bounds for call option*) Using the same parameters as in Example 19.10, we get $C \leq 42$ and

$$C \geq \max[0, e^{-0.5 \times 0.05}(43.06 - 38)] = 4.94.$$

Empirical Example 19.18 (*The option price bounds in Figure 19.22*) At very low strike prices, it is almost certain that the option will be exercised at expiration. Therefore, the present value of the cost, $C + e^{-my} K$, must be almost equal to the present value of a forward contract, $e^{-my} F$. Combining gives $C = e^{-my}(F - K)$. In contrast, at very high strike prices, the probability of exercise is almost zero—so the option price is too.

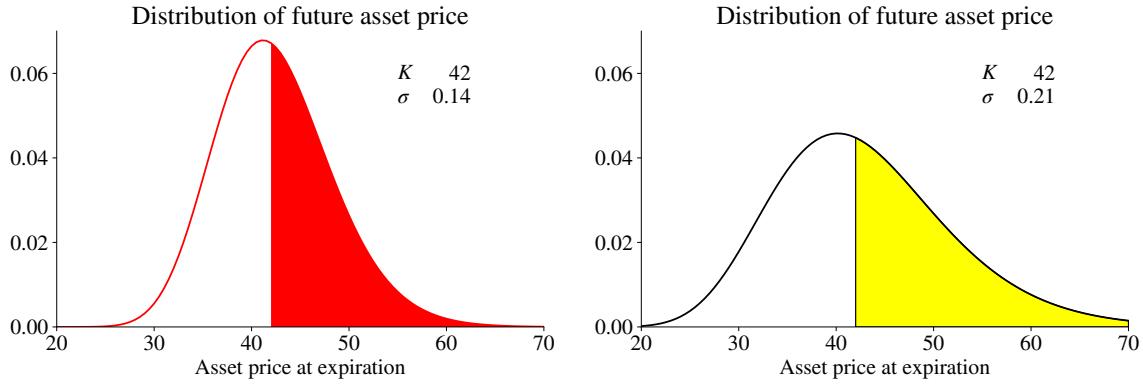


Figure 19.20: Distribution of future asset price

Proof (*of (19.15)) Portfolio A: one European call option and $e^{-my}K$ on a bank account. At expiration, this portfolio is worth S_m if the option is exercised, and K otherwise: $\max(S_m, K)$. Portfolio B: one prepaid forward contract, which is worth S_m at expiration. (Since you pay $e^{-my}F$ now, there is no payment at expiration.) Clearly, portfolio A is always worth more at expiration, so it must also be worth more right now: $C_E + e^{-my}K \geq e^{-my}F$. Rearrange to get (19.15). Since $C_A \geq C_E$, the bound holds also for an American call option. \square

19.8.2 Pricing Bounds for (European and American) Put Options

The prices of American and European put options must satisfy the following restrictions

$$P_E \leq e^{-my}K \text{ and } P_A \leq K \quad (19.18)$$

$$0 \leq P_E \text{ and } 0 \leq P_A \quad (19.19)$$

$$e^{-my}(K - F) \leq P_E \text{ and } K - S \leq P_A. \quad (19.20)$$

See Figure 19.23.

The motivations are as follows. First, the payoff from a put option is $\max(K - S, 0)$, so the maximum value is the strike price (when $S = 0$). For a European put, this payoff is received only at expiration, so the maximum value today is the present value of the strike price. Second, the put option gives rights, not obligations: its price value cannot be negative. Third, the lowest possible value of a call option is zero, so the put-call parity (19.5) immediately gives that the European put price must exceed the present value of $K - F$. (See below for an alternative proof.) In contrast, the American put can be

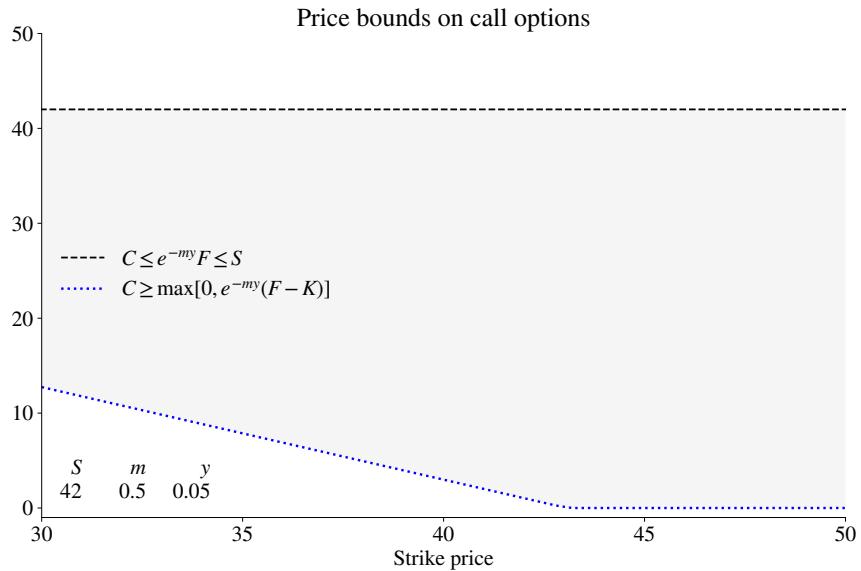


Figure 19.21: Call option price bounds as a function of the strike price

exercised now so its value must be at least as high as the intrinsic value.

Proof (*of (19.20)) Portfolio A: one European put option and a prepaid forward contract. At expiration, this portfolio is worth K if the option is exercised, and S_m otherwise: $\max(K, S_m)$. Portfolio B: $e^{-my}K$ on a bank account, which is worth K at expiration. Clearly, portfolio A is always worth more at expiration, so it must also be worth more right now: $P_E + e^{-my}F \geq e^{-my}K$. Rearrange to get (19.20). Since $P_A \geq P_E$, the bound holds also for an American put option. \square

19.9 Early Exercise of American Options

This section discusses early exercise of American options. There are some cases where we can exclude early exercise, so the American option is priced as a European option. In other cases, we cannot exclude early exercise—but we may still be able to say something about when early exercise is likely. More precise answers will require building a model for the pricing. Clearly, the answer is then model dependent.

The key results are as follows (assuming interest rates are positive):

	no dividends	with dividends
Call	no early exercise	early exercise (at high S)
Put	early exercise (at low S)	early exercise

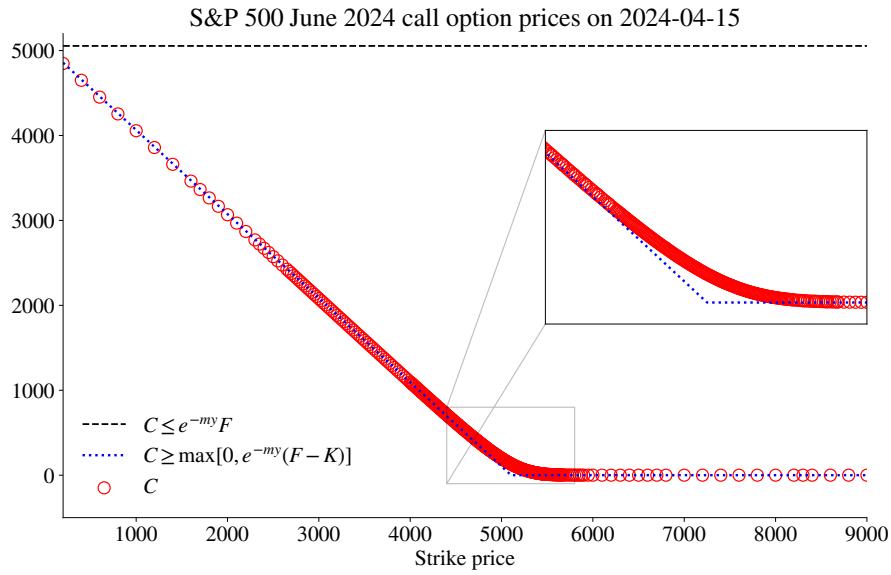


Figure 19.22: Prices and bounds for S&P 500 options

(Negative interest rates means that you could plausibly have early exercise for all four types.)

Proposition 19.19 (*No early exercise, American call, no dividends*) *An American call option on an asset without dividends should never be exercised early (if the interest rate is positive). It therefore has the same price as a European call option.*

See Figure 19.26 for an illustration of the fact that early exercise is not profitable for a call on an underlying asset without dividends since $C_A \geq C_E > \max(0, S - K)$, so the market price of the American call option will always be higher (or equal) to what you get by exercising. Rather, sell the option. Actually, if $C_A < S - K$, then there is an arbitrage opportunity: buy the option and exercise immediately to earn an instantaneous and risk-free profit of $S - K - C_A$.

Proof (of Proposition 19.19) From the put-call parity for European options (19.5), $C_E = P_E + S - e^{-my}K$, we have $C_E \geq S - K$ as long as the interest rate is positive (since $P_E \geq 0$). Since $C_A \geq C_E$, selling the option gives more than exercising it. \square

Example 19.20 (*Bankruptcy, American put, no dividends*) *Suppose the underlying asset goes bankrupt, then $S = 0$ and it is known that it will stay at $S = 0$. Exercising the American put option now gives K , whereas waiting until expiration has a present value of $e^{-my}K$ (which is lower): early exercise is optimal.*

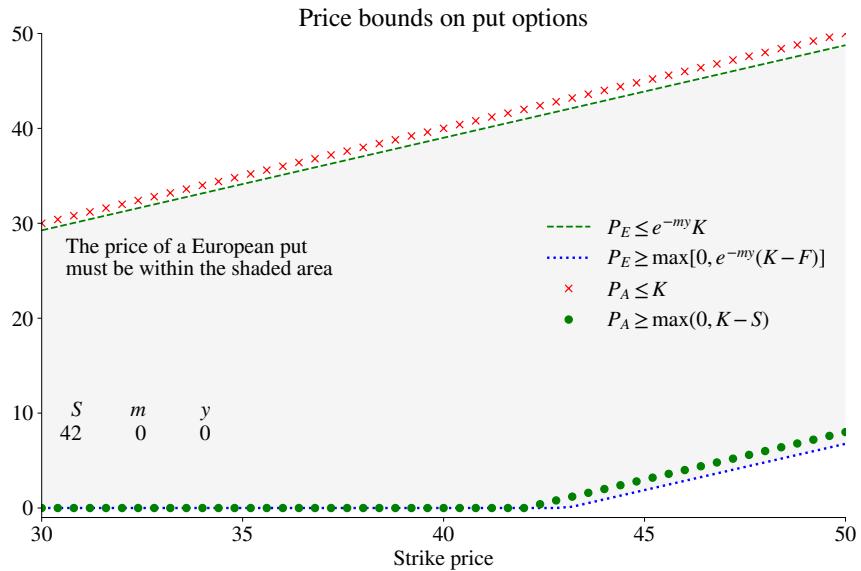


Figure 19.23: Put option price bounds as a function of the strike price

See Figures 19.24 –19.25 for an illustration, based on a numerical solution (of a specific model, so the precise results are not general, but discussed later) for the price on an American put option. In particular, Figure 19.24 shows in which nodes early exercise is optimal for an American put option: at low asset prices. In contrast, Figure 19.25 illustrates that numerical calculations verify that an American call option is not exercised early.

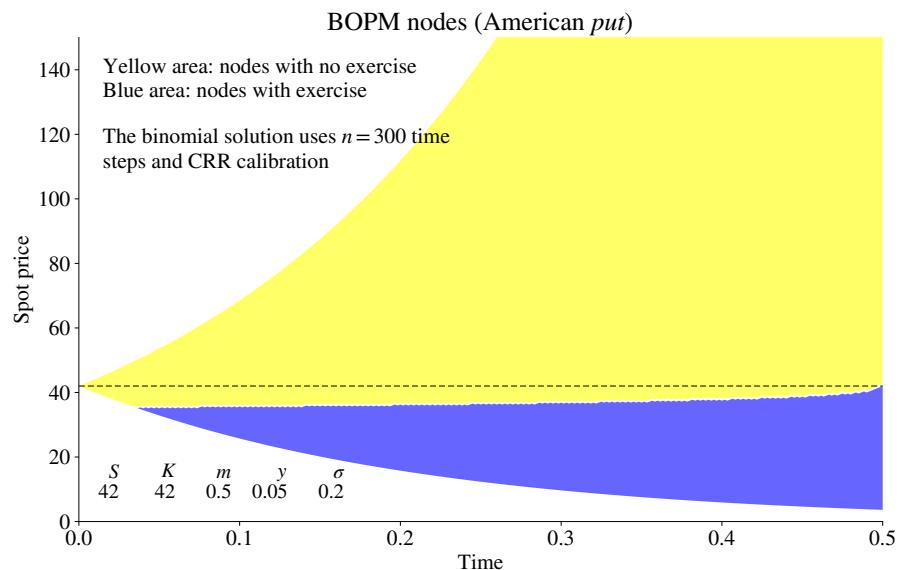


Figure 19.24: Numerical solution of an American put price (no dividends)

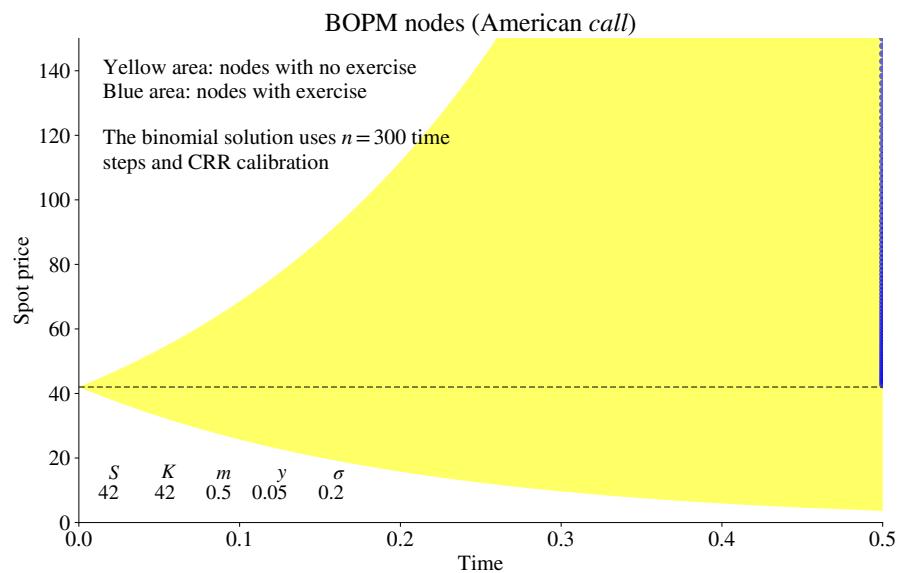


Figure 19.25: Numerical solution of an American call price (no dividends)

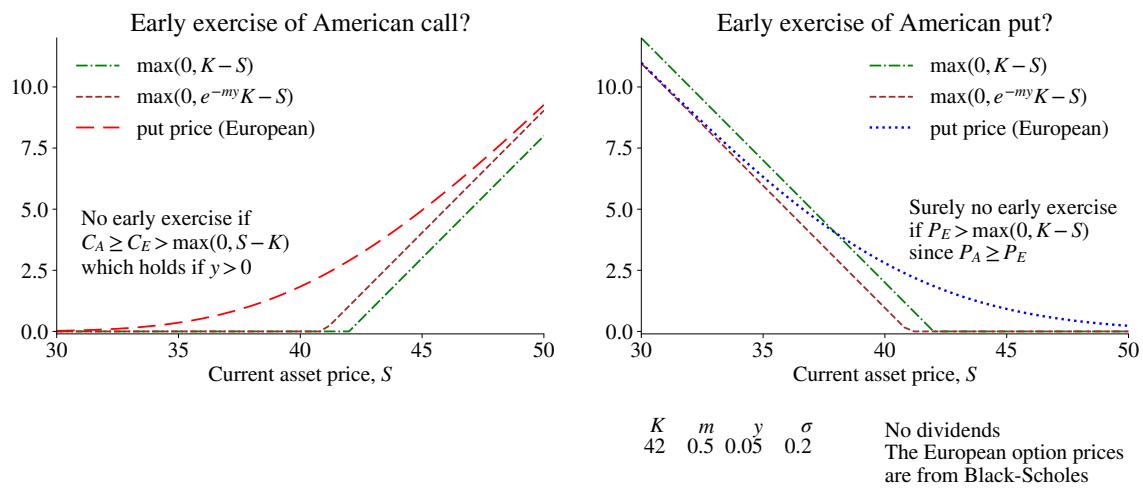


Figure 19.26: Early exercise of American call and put options (no dividends)

Chapter 20

The Binomial Option Pricing Model

20.1 Overview of Option Pricing

There are basically two ways to model option prices: a factor model (such as CAPM) or a no-arbitrage argument. These notes focus on the latter, based on the contributions by Cox, Ross, and Rubinstein (1979) and Rendleman and Bartter (1979).

20.2 The Basic Binomial Model

In the binomial model option pricing model (BOPM), the price of the underlying asset can change in only two ways. This is very stylized, but useful for establishing some key ideas of option pricing and provides a foundation for a more realistic model by cumulating many short subperiods. When applied to a European-style option, the binomial model converges to the well-known Black-Scholes model, as the subperiods become very many and short. However, in contrast to the latter, the binomial model can also be easily applied to an American-style option.

20.2.1 A Binomial Process for the Price of the Underlying Asset

The binomial tree for the underlying asset starts at the current price S and has probability q of moving to Su ($S \cdot u$, where typically $u > 1$) in the next period and a probability of $1 - q$ of moving to Sd (where $d < u$). This is illustrated in Figure 20.1. These probabilities are the true (“natural”) probabilities. Clearly, the expected value of the future asset price is $qSu + (1 - q)Sd$.

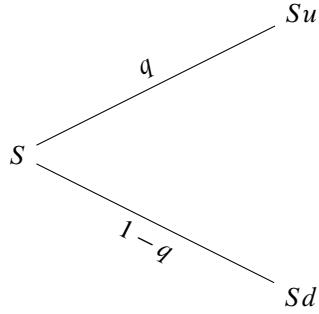


Figure 20.1: Binomial process for S

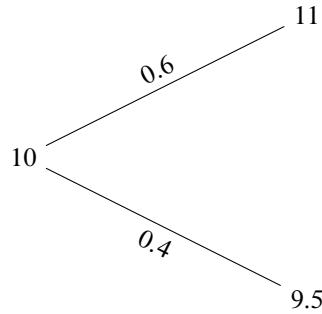


Figure 20.2: Numerical example of a binomial process for S

Example 20.1 (Binomial process) Suppose $S = 10$, $u = 1.1$, $d = 0.95$, and $q = 0.6$. Then, the process has a 60% probability of increasing from 10 to 11 and a 40% probability of decreasing to 9.5. See Figure 20.2.

We take it for granted that

$$u > e^{yh} > d. \quad (20.1)$$

If this condition is not satisfied, then trivial arbitrage opportunities arise. For instance, if $e^{yh} > u$, then we could shorten the underlying asset and buy bonds: this would guarantee a positive payoff for a zero investment (an arbitrage possibility).

20.2.2 No-Arbitrage Pricing of a Derivative

Basic Setup

Consider a derivative asset that will be worth f_u in case the underlying asset ends up at Su and f_d if it ends up at Sd . Notice that f_u is just the notation for the value (price) of the derivative in the up state (it should *not* be read as f times u).

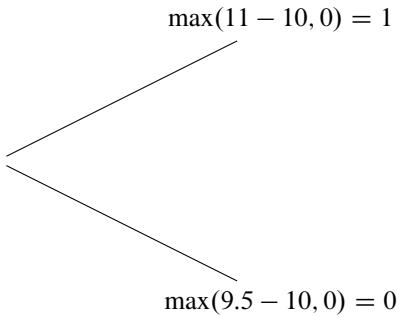


Figure 20.3: Numerical example of call option payoff, $K = 10$

As an example, suppose the derivative is a call option with strike price K and that the next period is the expiration date. Then,

$$f_u = \max(Su - K, 0) \text{ and } f_d = \max(Sd - K, 0). \quad (20.2)$$

Example 20.2 (European call option) With the parameters in Example 20.1, equation (20.2) shows that a European call option with strike price of 10 has

$$f_u = \max(11 - 10, 0) = 1 \text{ and } f_d = \max(9.5 - 10, 0) = 0,$$

while a strike price of 9 gives

$$f_u = \max(11 - 9, 0) = 2 \text{ and } f_d = \max(9.5 - 9, 0) = 0.5.$$

See Figure 20.3.

Step 1: Construct a risk-free Portfolio

We now use a no-arbitrage argument to determine the present price of the derivative, denoted f . Consider the following portfolio

$$\begin{aligned} &\Delta \text{ of the underlying asset, and} \\ &-1 \text{ of the derivative,} \end{aligned} \quad (20.3)$$

where Δ is yet to be decided. (Note that Δ here denotes a quantity, *not* a difference.)

For a given value of Δ , the payoff of the portfolio in the next period is $\Delta Su - f_u$ in the “up” state and $\Delta Sd - f_d$ in the “down” state. To make the portfolio *risk-free*, Δ must

be such that the payoff is the same in both states

$$\Delta S_u - f_u = \Delta S_d - f_d, \text{ so} \\ \Delta = \frac{f_u - f_d}{S(u - d)}. \quad (20.4)$$

With this choice of Δ (also called the “delta hedge”) the portfolio is risk-free. For future reference, we can also notice that Δ in (20.4) looks like a derivative, $\partial f / \partial S$.

Example 20.3 (European call option) Continuing from Example 20.2 we get

$$\Delta = \frac{1 - 0}{10(1.1 - 0.95)} = \frac{2}{3} \text{ for } K = 10.$$

The payoff of this portfolio is indeed safe. For instance, for the $K = 10$ option the value in the up state is $\frac{2}{3} \cdot 11 - 1 = 19/3$ and in the down state $\frac{2}{3} \cdot 9.5 - 0 = 19/3$. For a $K = 9$ call option, $\Delta = \frac{2 - 0.5}{10(1.1 - 0.95)} = 1$.

Step 2: Make the Return of the Portfolio Equal to the Risk-free Rate

Since the choice of Δ in (20.4) makes the portfolio safe, it must have *same return as the risk-free asset* (otherwise, arbitrage opportunities would arise). This is the same as requiring that the present value of the portfolio payoff (left hand side in the equation below) equals the cost of the portfolio today (right hand side)

$$e^{-yh}(\Delta S_u - f_u) = \Delta S - f, \quad (20.5)$$

where we still keep the Δ notation (to save space), but assume that Δ is determined as in (20.4). We could equally well have used the payoff in the down state, $\Delta S_d - f_d$, since it is the same. This equation defines the (current) arbitrage-free price f of the derivative.

Solve (20.5) for f and then use the value of Δ from (20.4) that ensures that the portfolio is risk-free

$$f = \Delta S(1 - e^{-yh}u) + e^{-yh}f_u \quad (20.6)$$

$$= \frac{f_u - f_d}{u - d}(1 - e^{-yh}u) + e^{-yh}f_u \quad (20.7)$$

$$= e^{-yh}[pf_u + (1 - p)f_d] \text{ with } p = \frac{e^{yh} - d}{u - d} \quad (20.8)$$

$$= e^{-yh} E^*(\text{future payoff of derivative}). \quad (20.9)$$

$$\begin{array}{c}
f_u = \max(Su - K, 0) \\
\\
f = e^{-yh}[pf_u + (1-p)f_d] \\
\\
(p = \frac{e^{yh}-d}{u-d}) \qquad f_d = \max(Sd - K, 0)
\end{array}$$

Figure 20.4: Solving for a call option price

$$\begin{array}{c}
\max(11 - 10, 0) = 1 \\
\\
\sqrt{3} \\
\\
\frac{1}{3} \times 1 + \frac{2}{3} \times 0 = \frac{1}{3} \\
\\
\frac{2}{3} \\
\\
(y = 0) \qquad \qquad \qquad \max(9.5 - 10, 0) = 0
\end{array}$$

Figure 20.5: Numerical example of call option price, zero interest rate

These are alternative ways to express the price of the derivative, f .

Equation (20.7) shows what the price of the derivative must be, and is written in terms of the possible outcomes and the interest rate. Notice that neither probabilities, nor risk preferences enter this expression, since we have used a no-arbitrage argument to price this derivative. This works because there are as many relevant assets, (risk-free and underlying asset) as there are possible outcomes (up or down), meaning that it is possible to construct a risk-free portfolio).

Equation (20.8) shows that the current price of the derivative is the present value of what *looks like* an expectation of the payoff of the derivative ($pf_u + (1-p)f_d$). This expression is quite useful since we can think of p as a “risk neutral probability” although it is not a probability in the usual sense: it is just a convenient construction. Note, though, that under the restrictions in (20.1), $0 < p < 1$, as any “probability” should be. This interpretation is highlighted in (20.9), where E^* stands for the expectations according to the *risk neutral distribution* (more about that later). The computation in (20.8) is illustrated in Figure 20.4.

The risk neutral probability p does not depend on the specific derivative considered, as long as it has the same underlying asset; rather, p depends only on the underlying asset and the interest rate. See Hull (2022) 13 and McDonald (2014) 13–14 for further details.

Example 20.4 (*European call option*) Continuing from Example 20.2 and assuming that $y = 0$, equation (20.8) provides the price of a call option with strike price 10 as

$$\begin{aligned} f &= e^{-0} [p1 + (1 - p)0] \text{ with } p = \frac{1 - 0.95}{1.1 - 0.95} = 1/3 \\ &= 1/3. \end{aligned}$$

See Figure 20.5. For the call option with a strike price of 9, we get

$$f = e^{-0} [(1/3) \times 2 + (2/3) \times (1/2)] = 1.$$

20.2.3 Applying the No-Arbitrage Pricing on Different Derivatives

This section discusses how the pricing formula (20.8) can be applied to specific derivatives.

A *forward contract* has a zero current price (nothing is paid until expiry), and the payoff at expiry is $f_u = Su - F$ in the up state (the value of the underlying asset minus the forward price) and $f_d = Sd - F$ in the down state. Using this in (20.8) gives

$$0 = e^{-yh} [p(Su - F) + (1 - p)(Sd - F)], \text{ so} \quad (20.10)$$

$$F = pSu + (1 - p)Sd. \quad (20.11)$$

This shows that the mean of the risk neutral distribution equals the forward price.

Example 20.5 (*A forward contract*) Continuing from Example 20.4, we get

$$F = (1/3) \times 11 + (2/3) \times 9.5 = 10,$$

which is the same as $S = 10$ (since the interest rate is zero).

An “Arrow-Debreu asset” (a sort of theoretical derivative often used in asset pricing models) pays off one unit in the up state and zero otherwise ($f_u = 1$ and $f_d = 0$). This is also a so-called “cash-or-nothing” call option provided the up state means that the option is in the money ($Su > K$). From (20.8) we have

$$f = e^{-yh} p. \quad (20.12)$$

20.2.4 Replicating (and Hedging) a Derivative

The no-arbitrage argument in (20.4) was based on the fact that a portfolio with Δ of the underlying asset and -1 of the derivative replicates a safe asset.

This argument can be turned around to replicate the derivative by holding the following portfolio (these are values of the positions)

$$\begin{aligned} & \Delta S \text{ in the underlying asset, and} \\ & f - \Delta S \text{ in a safe asset.} \end{aligned} \tag{20.13}$$

This means that we hold Δ underlying assets (each of which costs S) and hold $f - \Delta S$ on the money market (the latter is negative so it means borrowing). This replicates the derivative's payoff. We can therefore hedge a short position in the derivative by portfolio (20.13).

Proof (of that (20.13) replicates the derivative) The payoff of this portfolio is $\Delta Su + e^{yh}(f - \Delta S)$ in the up state and $\Delta Sd + e^{yh}(f - \Delta S)$ in the down state. Recall from (20.5) that $\Delta Su - f_u$ equals $e^{yh}(\Delta S - f)$. Use (the negative of) this in payoff in the up state to get f_u . Also, Δ is such that $\Delta Sd - f_d$ also equals $e^{yh}(\Delta S - f)$. Use in the payoff for the down state to get f_d . \square

Example 20.6 (*Replicating a call option*) For the call option with a strike price of 10 and with a zero interest rate, we have (see Examples 20.3 and 20.4) $\Delta = 2/3$, $f = 1/3$ and

$$(f - \Delta S) = \frac{1}{3} - \frac{2}{3} \cdot 10 = -6\frac{1}{3},$$

so we borrow. The value of this portfolio in the up node is $\frac{2}{3} \times 11 - 6\frac{1}{3} = 1$ and in the down node $\frac{2}{3} \times 9.5 - 6\frac{1}{3} = 0$ which are the same as the call option.

20.2.5 Where is the Risk Premium?

We have used a no-arbitrage method to price the derivative. It works since the derivative is a redundant asset: it can be replicated by a portfolio of the underlying asset and a risk-free asset (see (20.13)) and therefore must have the same price as this portfolio. Clearly, this portfolio will incorporate a risk premium and so must the derivative.

It may seem as if the pricing formula (20.8) is free from the preference parameters that would determine the risk premium. Not correct. The pricing formula contains the current asset price (through f_u and f_d) which is indeed affected by preference parameters.

20.3 The Risk Neutral Probabilities

The relation between the true probabilities (q) and the risk neutral probabilities (p) depends whether the underlying asset has a risk premium or not.

From the spot-forward parity for an asset without dividends, we know that $F/S = e^{yh}$. Thus, an asset with a positive risk premium must obey

$$E_t S_{t+h}/S > F/S \text{ (with positive risk premium)}, \quad (20.14)$$

where we (for simplicity) focus on an asset without dividends. This expression implies that the expected return is higher than the risk-free rate.

In a binomial process, the expected value of the gross return is

$$E_t S_{t+h}/S = qu + (1 - q) d, \quad (20.15)$$

where q is the natural probability of the up state. At the same time, the risk neutral expected value equals the forward price (see (20.11)) divided by S

$$F/S = pu + (1 - p) d. \quad (20.16)$$

Combining (20.15) and (20.16) shows that for (20.14) to hold, we must have

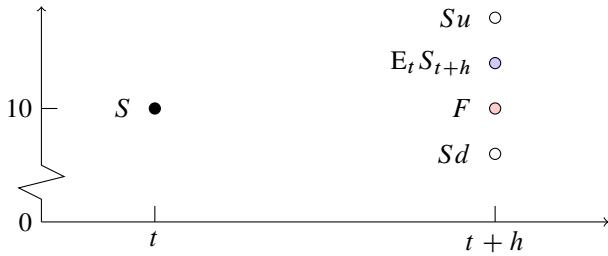
$$q > p \text{ (with positive risk premium)}. \quad (20.17)$$

To understand the intuition for this, consider an alternative case where the risk premium is zero ($E_t S_{t+h}/S = F/S$), then

$$q = p \text{ (with no risk premium)}. \quad (20.18)$$

The absence of a risk premium could either depend on (a) the asset has no systematic risk; or (b) that we have risk neutral investors. In either case, p equals the true probabilities, suggesting the name “risk neutral probability” for p .

Now, (20.17) is easier to interpret: both $E_t S_{t+h}/S$ and F/S are averages of the same values (d and u) and they only differ with respect to the probabilities. Clearly, for $E_t S_{t+h}/S$ to exceed F/S , the former must have a larger probability for the high value (u). See Figure 20.6 for an illustration. One interpretation is that a risk averse investor requires a higher probability of the up state, and thus a higher expected return, than a risk neutral investor.



Calculations:

$$\begin{aligned} S &= 10, S_d = 9.5, S_u = 11 \\ E_t S_{t+h} &= 0.6 \times 11 + 0.4 \times 9.5 = 10.4 \\ F &= 1/3 \times 11 + 2/3 \times 9.5 = 10 \end{aligned}$$

Figure 20.6: Risk premium and risk neutral probabilities for an asset with a positive risk premium

Example 20.7 (*Natural versus risk neutral probability*) With the parameters in Example 20.1

$$E_t S_{t+h}/S = 0.6 \times 1.1 + (1 - 0.6) \times 0.95 = 1.04.$$

With $y = 0$, $F = S = 10$, so $F/S = 1$. In this case, the underlying asset indeed has a positive risk premium (see (20.14)), and $q = 0.6$ while $p = 1/3$. See Figure 20.6 for an illustration.

Proof (of (20.17)) For (20.14) to hold, we need $qu + (1 - q)d > pu + (1 - p)d$. Subtract d from both sides to get $q(u - d) > p(u - d)$ and notice that $u - d > 0$ to conclude that $q > p$ is required. \square

20.4 Multi-Period Trees I: Basic Setup

20.4.1 The Binomial Tree for the Underlying Asset

In numerical applications, we chain a large number of up/down movement to get more realistic model properties of the underlying asset. This means that the (fixed) time to expiration is divided into many small time steps and that we can rebalance the portfolio at each of them.

Figure 20.7 is an illustration of a binomial tree with two subintervals and Figure 20.8 gives a numerical example. This tree has only three final nodes since $S_{ud} = S_{du}$: it is “recombining,” which is very useful to keep the number of nodes manageable. This would

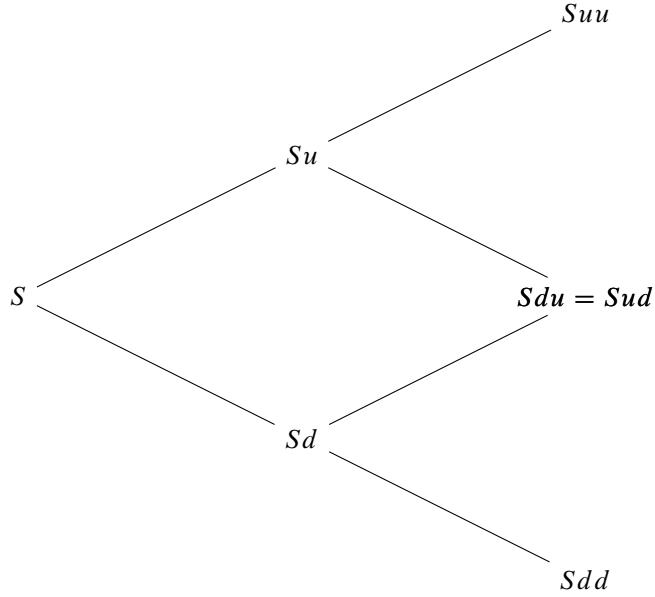


Figure 20.7: Binomial tree for underlying asset ($n = 2$)

not be the case if the up and down moves were different for different periods (non-iid price process). See Remark 20.9 for details.

Let m be the time to expiration of the derivative, typically measured in years. With n short time intervals, the length of each interval is $h = m/n$, see Figure 20.9. Clearly, if we use more time steps, then each of them is shorter. The size of the up and down movements, as well as the discounting, must also be scaled by the number of time steps: compare Figures 20.2 and 20.8. Otherwise we cannot preserve/control the general properties of the underlying asset. Later sections will discuss this recalibration in detail.

Remark 20.8 (*Building the tree**) One way of building the tree is to calculate the value of the underlying asset in a node as $Su^{N_u} d^{N_d}$ where N_u is the number of up steps and N_d the number of down steps since the beginning of the tree. For time step i , $(N_u, N_d) = (i, 0)$ for the top node, $(i-1, 1)$ for the second node and so forth until $(0, i)$ for the bottom node. In short, $N_u = [i, i-1, \dots, 0]$ and $N_d = i - N_u$.

Remark 20.9 (*Size of the binomial tree*) With n time steps, there are $n+1$ different prices at the end nodes. Also, there are a total of $(n+1)(n+2)/2$ nodes. There are $n!/(n-s)!s!$ different ways to reach the s th node below the top node (where $x! = x \times (x-1) \times \dots \times 1$). Summing across the nodes shows that the tree contains 2^n different paths. For instance,

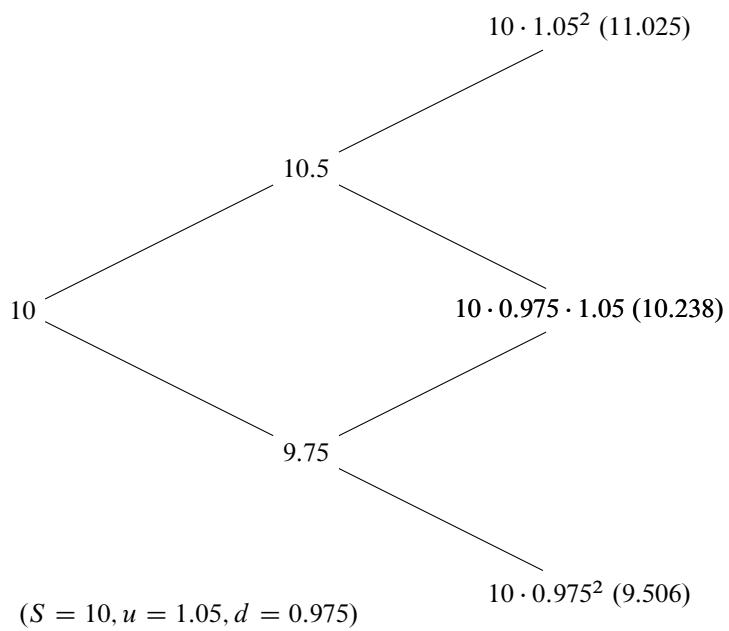
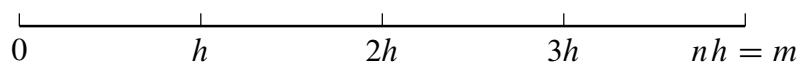


Figure 20.8: Numerical example of a binomial tree for underlying asset ($n = 2$)

our (recombining) tree has

n	no. end nodes	no. total nodes	no. paths
2	3	6	4
25	26	351	33, 554, 432
200	201	20, 301	1.6×10^{60}

In contrast, a non-recombining tree has 2^n end nodes, that is, as many as there are paths in the recombining tree.



Expiry: $m = 1/2$ years
 Steps: $n = 4$
 Step length: $h = m/n = 1/8$ years

Figure 20.9: Steps to reach time to expiration m

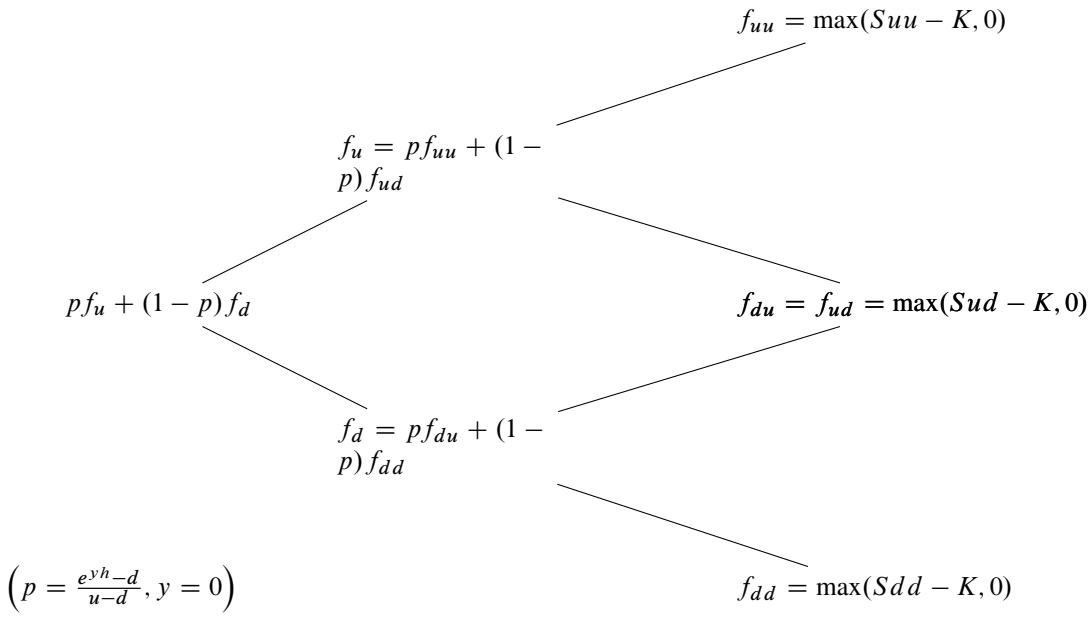


Figure 20.10: Binomial tree for European call option ($n = 2$), zero interest rate

20.4.2 Using a Binomial Tree for Pricing European Options

We can now apply the pricing formula (20.8) to each “subtree,” *starting at the end* of the tree (time step n) and *working backwards* towards the start of the tree (time step 0). Figure 20.10 illustrates the computations for a European call option with strike price K and two steps ($n = 2$) and Figure 20.11 gives a numerical example.

The structure of the tree for a European put option is the same as for a European call option, except that the payoff at the end nodes differ ($\max(0, S_m - K)$ for the call and $\max(0, K - S_m)$ for the put), see Figure 20.12.

Example 20.10 (*Tree for a European put*) For a put option with strike price $K = 10$, the values in Figure 20.11 would change to $f = 0.219$, $(f_u, f_d) = (0, 0.329)$ and $(f_{uu}, f_{ud}, f_{dd}) = (0, 0, 0.494)$.

This recursive calculation (using a tree with $n = 2$ as in Figure 20.10) gives the European option price

$$\begin{aligned} f &= e^{-y^2 h} [pf_u + (1 - p)f_d] \\ &= e^{-ym} [p^2 f_{uu} + 2p(1 - p)f_{ud} + (1 - p)^2 f_{dd}], \end{aligned} \quad (20.19)$$

since $e^{-y^2 h} = e^{-ym}$.

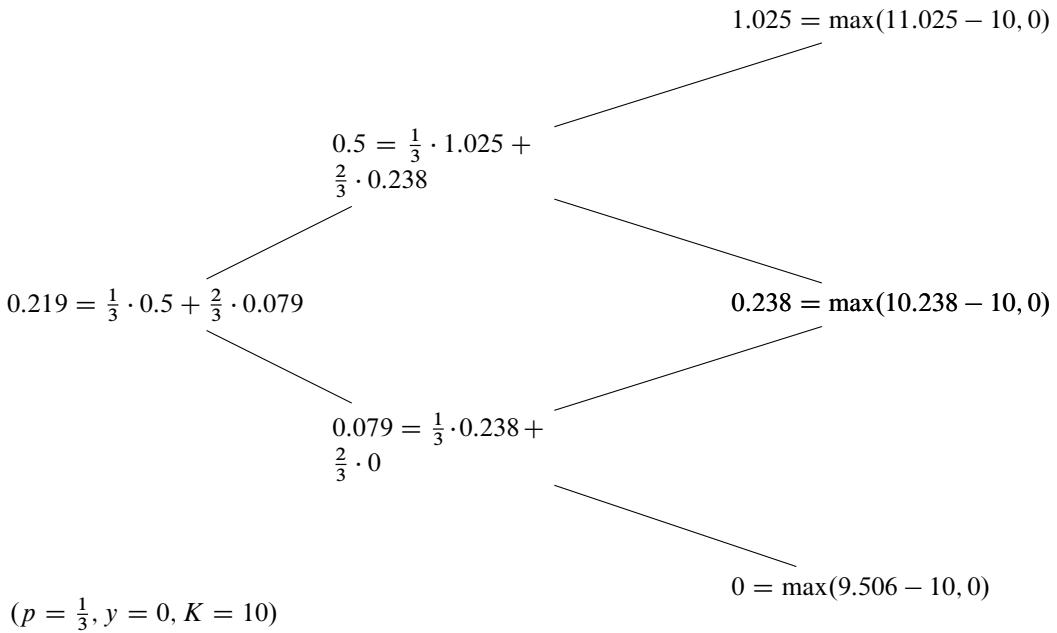


Figure 20.11: Numerical example of binomial tree for European call option ($n = 2$), zero interest rate. The underlying is described in Figure 20.8.

Notice that p^2 is the risk-neutral probability of the payoff f_{uu} , $2p(1-p)$ of the payoff f_{ud} and $(1-p)^2$ of the payoff f_{dd} , as illustrated in Figure 20.13. Therefore, (20.19) is a generalisation of (20.9):

$$f = e^{-ym} \mathbb{E}^*(\text{payoff of derivative at expiration}), \quad (20.20)$$

which says that the (European-style) derivative is the present value of the risk-neutral expected payoff at expiration. The distribution behind the risk neutral expectation is clearly more involved than before, but the same logic applies. See Figure 20.14 for an illustration of how the probabilities for different final outcomes change as the number of time steps (n) changes and the up and down movements are recalibrated to mimic the properties of the underlying asset (using the CRR approach, to be discussed later).

Remark 20.11 (*The binomial distribution**) After n independent draws, the number of up moves (k) has the binomial pdf, $n!/[k!(n-k)!] p^k (1-p)^{n-k}$ for $k = 0, 1, \dots, n$. For instance, with $n = 2$, we have p^2 for $k = 2$, $2p(1-p)$ for $k = 1$, and $(1-p)^2$ for $k = 0$.

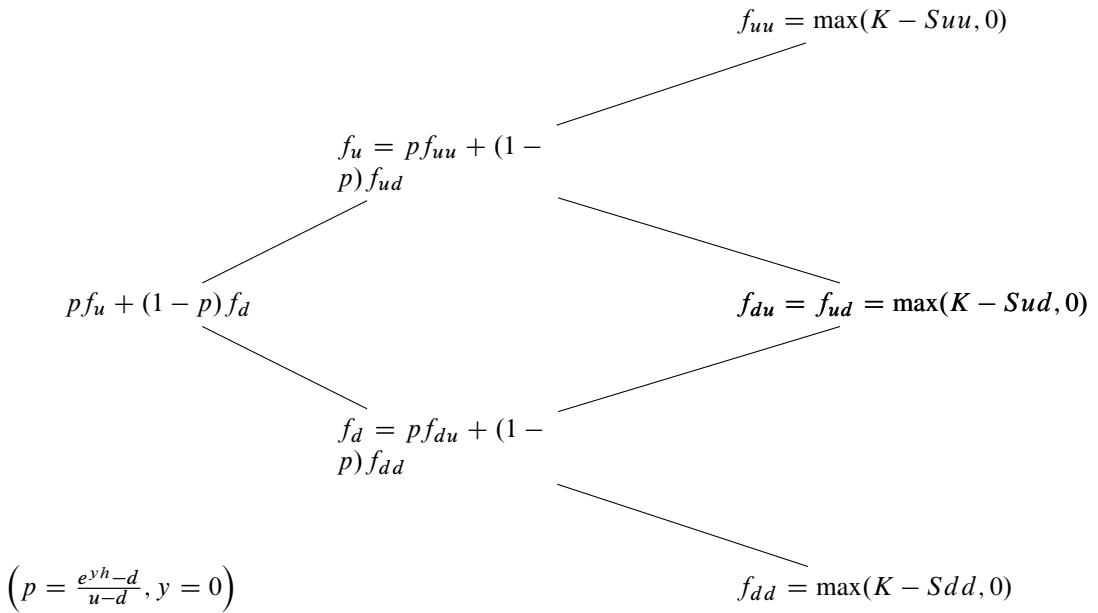


Figure 20.12: Binomial tree for a European put option ($n = 2$), zero interest rate

20.4.3 Using a Binomial Tree for Pricing American Options

The binomial tree we have used so far assumes that the derivative is “alive” until expiration. This is not necessarily the case for American options, so the approach needs to be modified to handle the possibility of early exercise.

Whenever you can exercise, the option value is the maximum of the exercise value and the value of keeping the option alive

$$\max(\text{value if exercised now}, \text{value of keeping an unexercised option}). \quad (20.21)$$

The value of an unexercised option is calculated as in (20.8): the present value of the risk neutral expected value in the next time step. This means that we solve this problem starting from the expiration date (just like for the European options), and calculate the value at each node, assuming, perhaps counter factually, that the option has not already been exercised at an earlier time step. See Figure 20.15 for an illustration. Also, see Figure 20.16 for a numerical example (the nodes where exercise is optimal are indicated by bold).

Figure 20.17 illustrates the solution for an American put option on an asset without dividends (the details of the calculations will be discussed later). Notice that the American put price exceeds the European put price, and more so at low asset prices and high interest

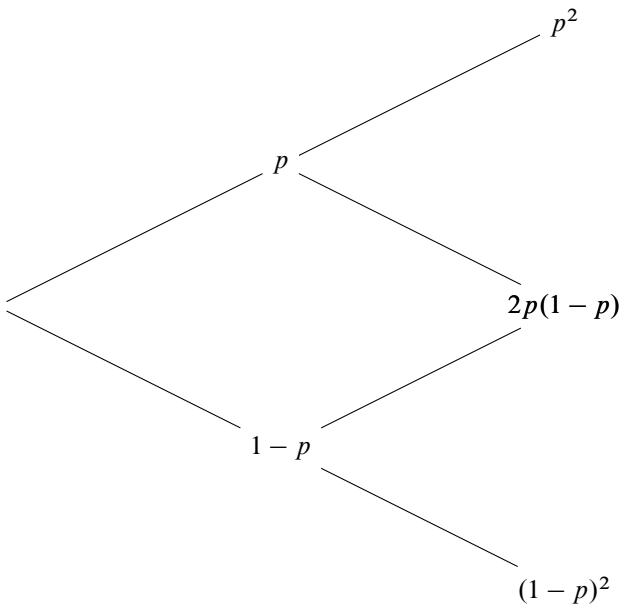


Figure 20.13: Probabilities of different nodes in a binomial tree

rates. The lower and upper limits on the put price are from the put-call “parity” (two inequalities) for American options. The call price C used in the figure is the same for European and American options (since there is no early exercise in this case).

20.5 Multi-Period Trees II: Calibrating the Tree

We now discuss how to construct a binomial tree (how to choose u and d) with many small time steps, so that it mimics the statistical properties of the underlying asset.

20.5.1 Mean and Variance of Data

Suppose you have a sample of log returns (r_τ for $\tau = 1$ to T) of the underlying asset, and that you are willing to assume that they are *iid*. Calculate the sample mean and variance and *annualize* them by dividing by the time length of the return period in the data (k)

$$\hat{\mu} = \frac{1}{k} \bar{r}_\tau \quad (20.22)$$

$$\hat{\sigma}^2 = \frac{1}{k} \widehat{\text{Var}}(r_\tau) \quad (20.23)$$

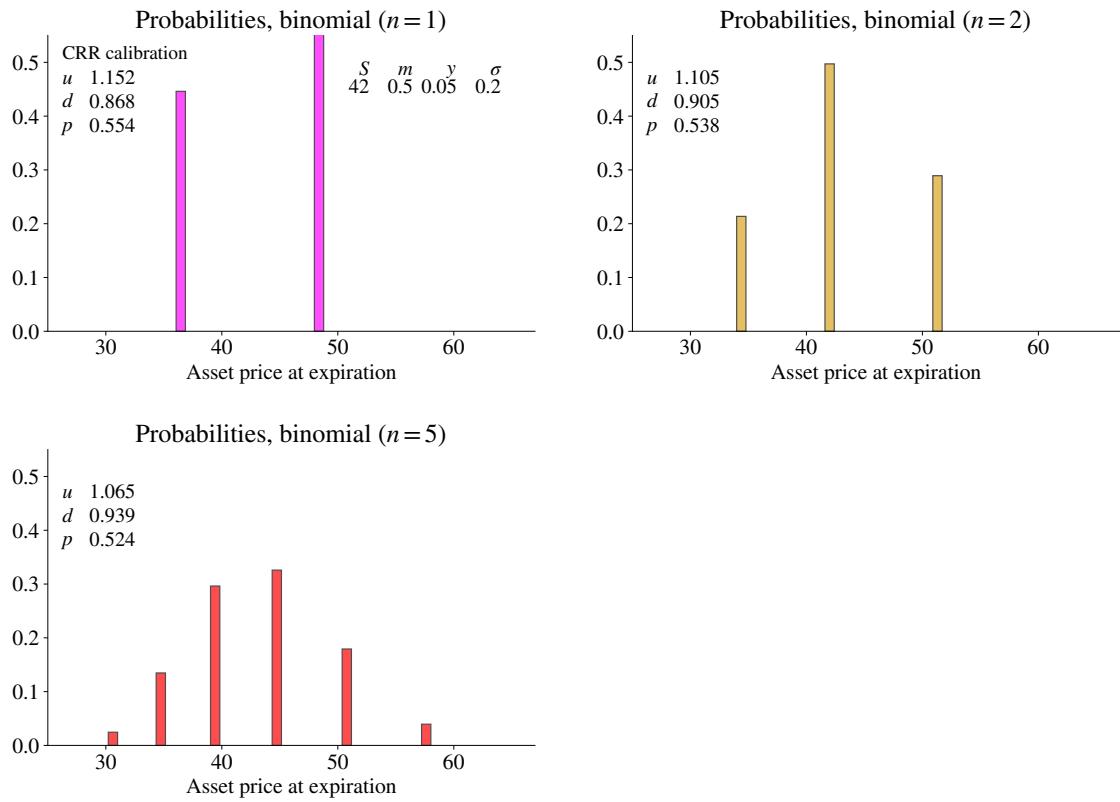


Figure 20.14: Probabilities of different final outcomes for different values of n

For instance, with daily return data $k = 1/252$ (only counting the trading days), so we multiply the moments in data by 252.

Expressing the moments in terms of annualised numbers ($\hat{\mu}$ and $\hat{\sigma}^2$) helps relating to the binomial model, and to compare results across different sampling intervals.

Example 20.12 (Variance for daily return) If the data is daily ($k = 1/252$) and the standard deviation is estimated to be 0.0126, then the annualised variance is $\hat{\sigma}^2 = 0.0126^2 \times 252 \approx 0.2^2$ and the annualized standard deviation is $\hat{\sigma} \approx 0.0126 \times \sqrt{252} = 0.2$.

20.5.2 Mean and Variance according to the Binomial Model

Recall the binomial process (for instance, in Figure 20.1)

$$S_{t+h} = \begin{cases} Su & \text{with probability } q \\ Sd & \text{with probability } 1 - q, \end{cases} \quad (20.24)$$

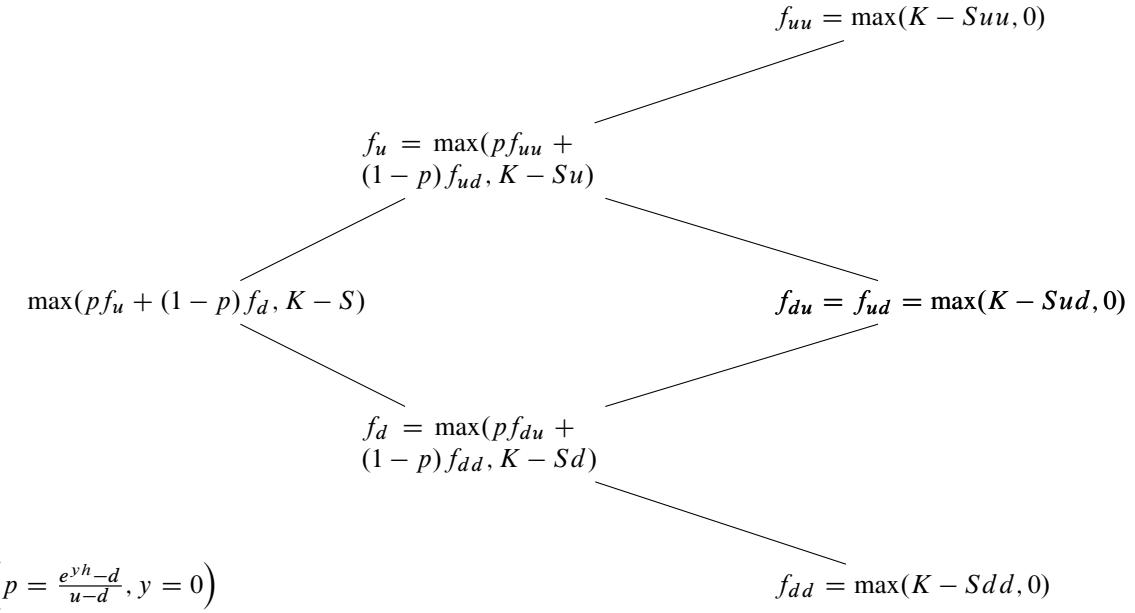


Figure 20.15: Binomial tree for an American put option ($n = 2$), zero interest rate

where S_{t+h} denotes the price in the next period (each period in the model is h years long). Clearly, this means that the log returns, $r_{t+h} = \ln(S_{t+h}/S_t)$, follow

$$r_{t+h} = \begin{cases} \ln u & \text{with probability } q \\ \ln d & \text{with probability } 1-q. \end{cases} \quad (20.25)$$

Remark 20.13 (*Mean and variance of a binomial process*) The mean of a (shifted) binomial process like (20.25) is $q \ln u + (1-q) \ln d$ and the variance is $q(1-q)(\ln u - \ln d)^2$.

This binomial process implies that the *annualized* mean and variance of the asset returns are (see Remark 20.13)

$$\text{annualized mean} = \frac{1}{h}[q \ln u + (1-q) \ln d], \quad (20.26)$$

$$\text{annualized variance} = \frac{1}{h}q(1-q)(\ln u - \ln d)^2. \quad (20.27)$$

The $1/h$ is the number of small time steps needed to get a full year. Notice that the k for the length of time periods in data and the h in the binomial tree need not be the same.

Example 20.14 (*Binomial process*) Suppose $S = 10, u = 1.1, d = 0.95$, and $q = 0.6$. This gives an expected value of $0.6 \times \ln 1.1 + 0.4 \times \ln 0.95 = 0.037$ and a variance

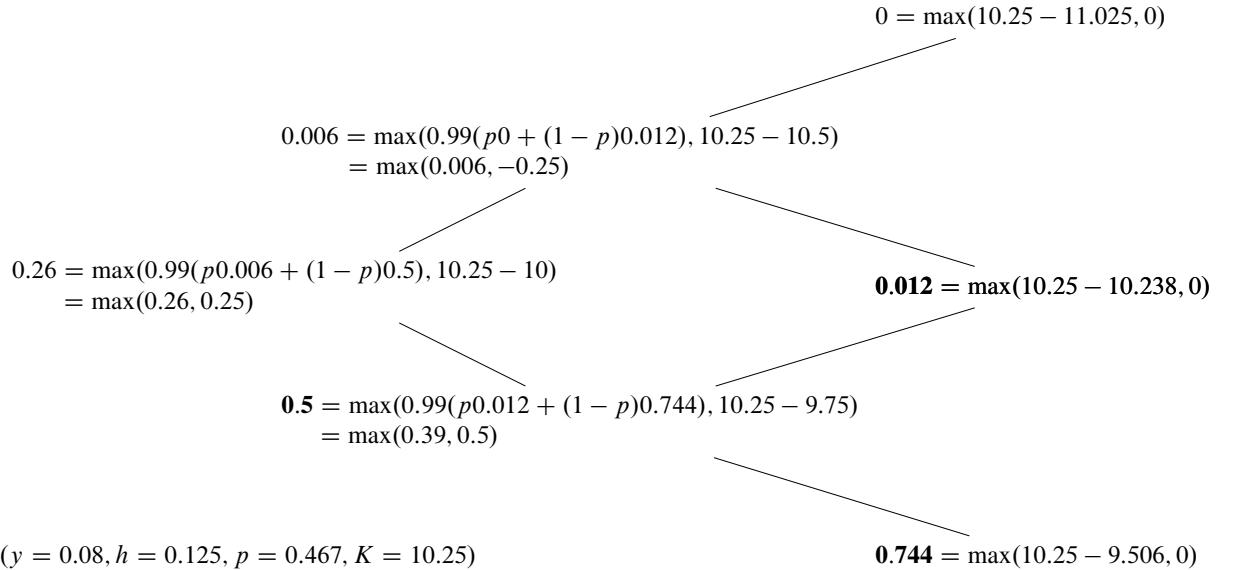


Figure 20.16: Numerical example of a binomial tree for an American put option ($n = 2$). Exercise is indicated by bold.

of $0.6 \times 0.4 \times (\ln 1.1 - \ln 0.95)^2 = 0.0052$. If the periods in the model are weeks ($h = 1/52$), then the annualized mean is $0.037 \times 52 \approx 1.9$ and the annualized variance is $0.0052 \times 52 \approx 0.27$.

20.5.3 Comparing Data and Model: The CRR Approach

There are three parameters (u , d , and q) which can be chosen to match the two moments, that is, to make the annualized mean and the variance from the model (20.26)–(20.27) equal to $(\hat{\mu}, \hat{\sigma}^2)$ from data in (20.22)–(20.23). We therefore have some degrees of freedom.

The most common approach is that of Cox, Ross, and Rubinstein (1979) where

$$u = e^{\hat{\sigma}\sqrt{h}} \text{ and } d = 1/u. \quad (20.28)$$

Recall that p needs to change when (u, d) change, since $p = (e^{yh} - d)/(u - d)$.

Example 20.15 (Parameters to binomial tree) With $h = 1/52$ and $\hat{\sigma} = 0.2$, (20.28) gives $u \approx 1.028$ and $d \approx 0.973$.

See Figure 20.18 for an illustration of how the parameters (p, u, d) converge as the number of time steps increases. Also, see Figure 20.19 for an illustration of CRR trees for different number of steps (n).

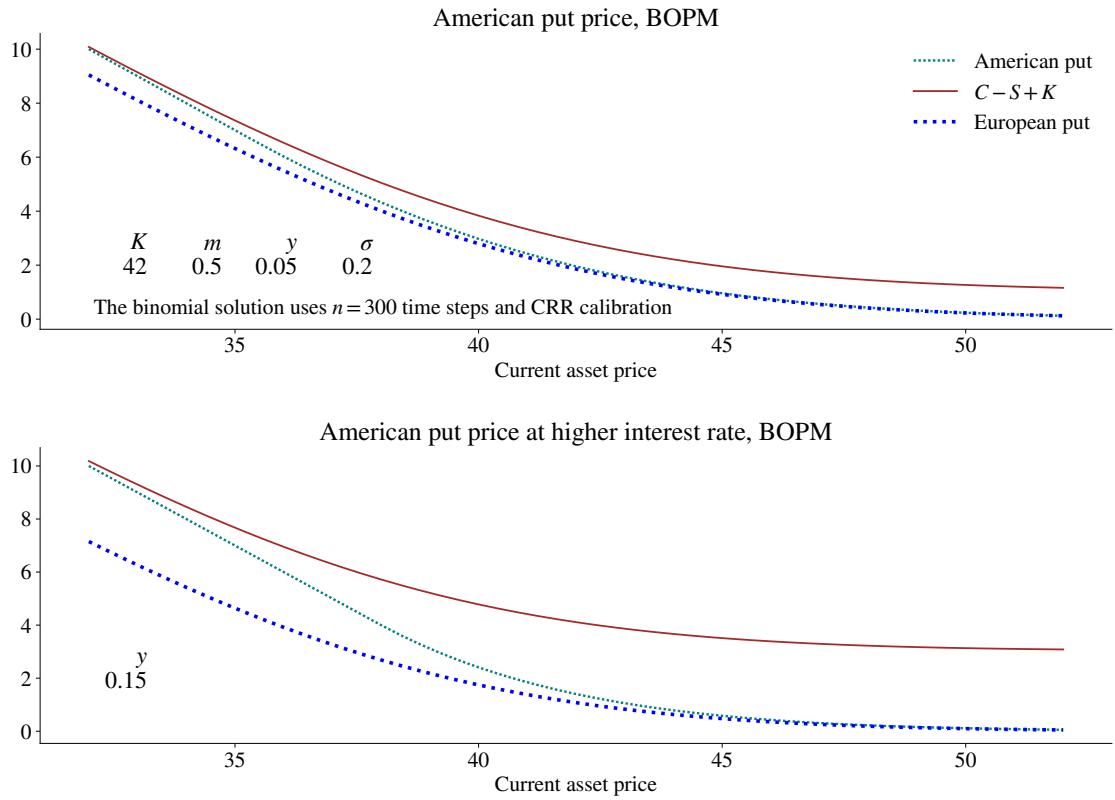


Figure 20.17: Numerical solution of an American put price

If the natural probability of an up move (q) is used to make the model implied mean equal to $\hat{\mu}$, then it can be shown (after some straightforward algebra) that

$$\text{Annualized Var (binomial process)} = \hat{\sigma}^2 - \hat{\mu}^2 h. \quad (20.29)$$

This does not fit the volatility in data exactly because of the $\hat{\mu}^2 h$ term, but the approximation improves as h decreases (the size of time steps decreases). However, once we have the values of u and d , the pricing of derivatives does not use the natural probability of the up state (q).

Proof (of (20.29)*) Use (20.28) in (20.26) and choose q so that the model implied annualized mean equals $\hat{\mu}$. Use the three parameters in (20.27) and simplify. \square

However, we must ensure that (20.1) holds ($u > e^{yh} > d$, to rule arbitrage opportunities), that is,

$$e^{\hat{\sigma}\sqrt{h}} > e^{yh} > e^{-\hat{\sigma}\sqrt{h}}, \quad (20.30)$$

which requires $\hat{\sigma} > y\sqrt{h} > -\hat{\sigma}$. In practice, this means that h must be small (the number

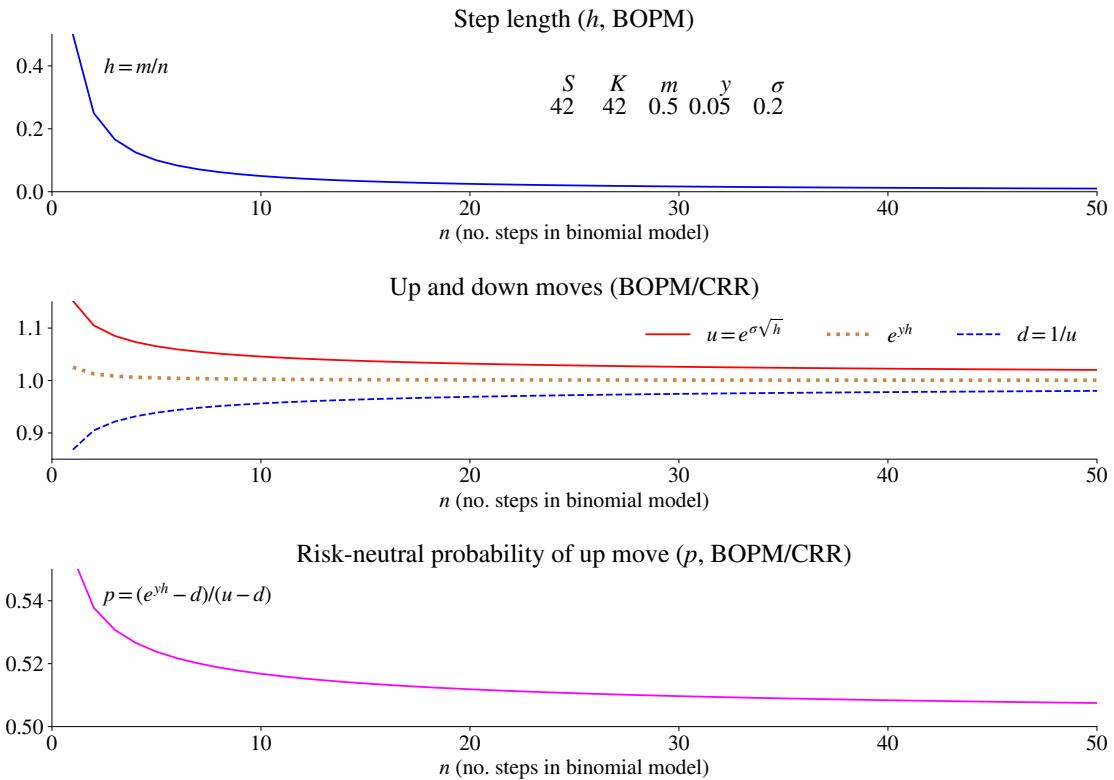


Figure 20.18: Convergence of the parameters in a binomial model

of steps, n , large). Always check that this condition is satisfied. Otherwise, the results of the calculations might be nonsense.

Example 20.16 (Checking parameters of binomial tree) With the parameters in Example 20.15 and assuming $y = 0.05$, we notice that $e^{yh} = e^{0.05/52} \approx 1.001$, so the requirement is fulfilled

$$1.028 > 1.001 > 0.973.$$

See Figures 20.20–20.21 for an illustration of how the resulting option price converge as the number of time steps increases. (The result from the Black-Scholes model will be discussed in detail in another chapter.) The zig-zag pattern suggests that some kind of average price, across $n - 1$ and n steps, may improve the performance (see Figure 20.20).

Figure 20.22 illustrates the calculations of the American put price for a single current value of the underlying asset (S). The shaded areas show the location of the nodes (possible prices of the underlying asset in the future) that are used in the calculation—and at which nodes that early exercise will happen. For comparison, Figure 20.23 shows that

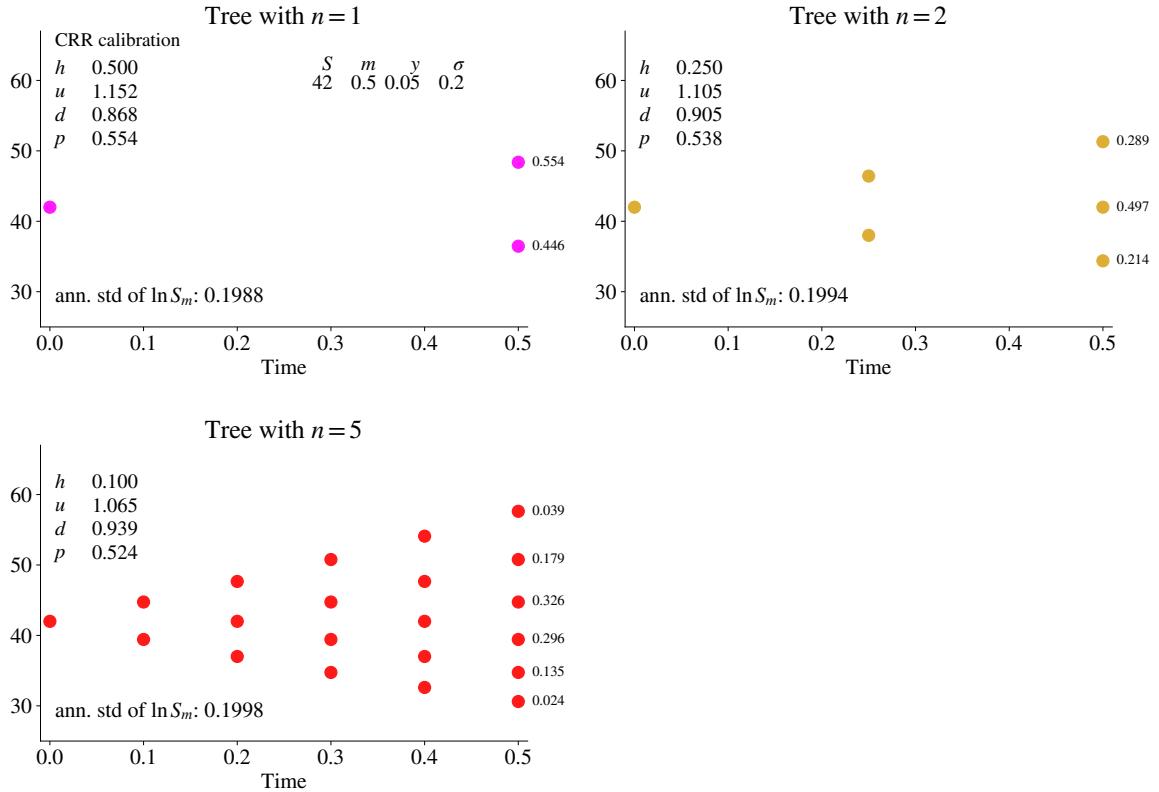


Figure 20.19: CRR trees for different values of n . The numbers at the end nodes are the probabilities of reaching that node.

the numerical calculations verify the theoretical result that an American call option is not exercised early.

20.6 Appendix – Continuous Dividends*

It is straightforward to construct another tree that allows for continuous dividends, provided they are proportional to the asset price.

Suppose dividends are paid at the known continuous *rate* δ and let the up and down movements in the asset price reflect the ex-dividend price (S in the initial period). Buying one unit of the underlying asset in the initial period costs S . If we move to the “up state” in the next period (h), then the owner first gets the dividend $Su(e^{\delta h} - 1)$ and can then sell the asset for the (ex-dividend) price Su : the total value is $Sue^{\delta h}$. Notice that the dividend is proportional to price in the same period. The “down state” is similar: just replace u by d .

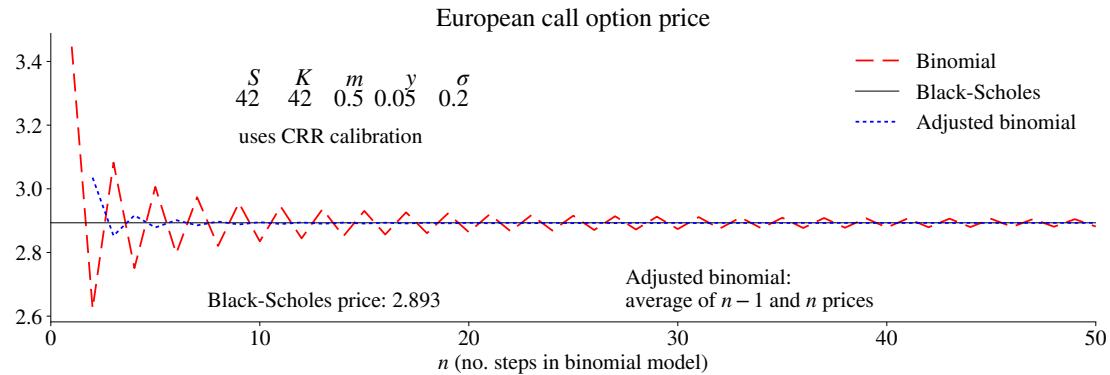


Figure 20.20: Convergence of the binomial price (call)

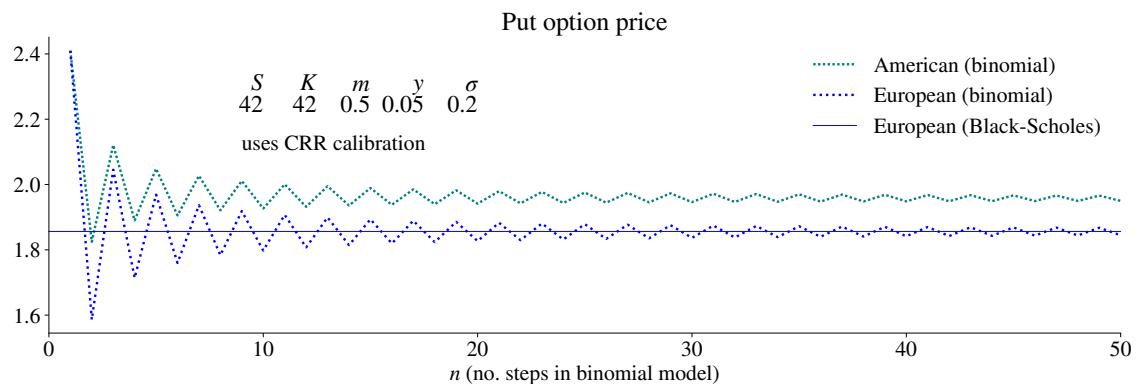


Figure 20.21: Convergence of the binomial price (put)

We now construct a risk-free portfolio to find out how a derivative is priced in the initial period. First, to construct a risk-free portfolio, hold Δ of the underlying asset and -1 of the derivative. The payoff of the portfolio at expiry is $\Delta S u e^{\delta h} - f_u$ in the “up” state and $\Delta S d e^{\delta h} - f_d$ in the “down” state. To make the portfolio risk-free the delta must be

$$\Delta = \frac{f_u - f_d}{S e^{\delta h} (u - d)}. \quad (20.31)$$

Second, to make the return of the portfolio equal to the risk-free rate, we set the present value of our risk-free portfolio equal to the cost of the portfolio

$$e^{-y h} [\Delta S e^{\delta h} u - f_u] = \Delta S - f. \quad (20.32)$$

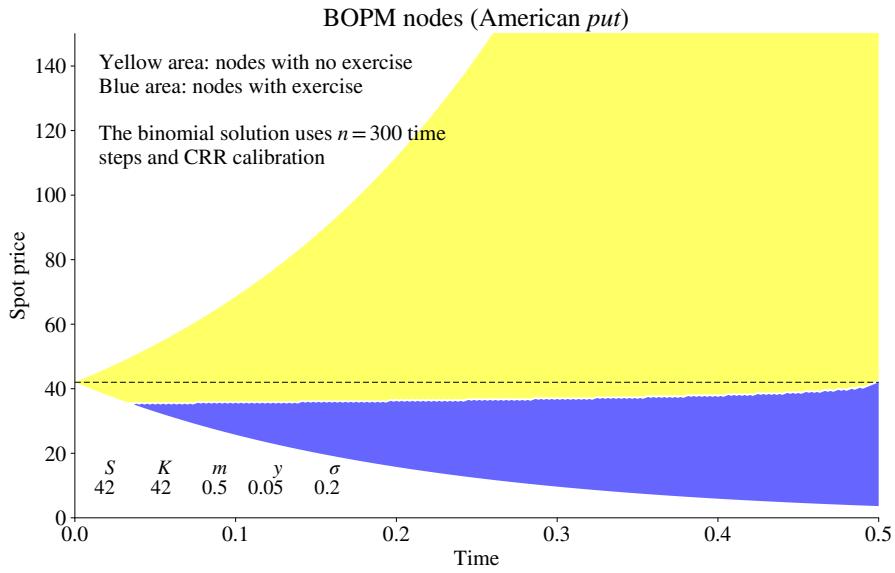


Figure 20.22: Numerical solution of an American put price

Use (20.31) and rearrange as

$$f = \Delta S [1 - e^{(\delta-y)h} u] + e^{-yh} f_u \quad (20.33)$$

$$= \frac{f_u - f_d}{e^{\delta h} (u - d)} [1 - e^{(\delta-y)h} u] + e^{-yh} f_u \quad (20.34)$$

$$= e^{-yh} [p f_u + (1 - p) f_d] \text{ with } p = \frac{e^{(y-\delta)h} - d}{u - d}. \quad (20.35)$$

With this new definition of p , the rest of the computations are as in the case without dividends. In particular, the drift of the asset price does not matter, so u and d can be chosen as before, for instance, as in (20.28).

Remark 20.17 (*Risk neutral drift with continuous dividends*) *With continuous dividends, the risk neutral expected value is $E_t^* S_{t+h}/S_t = e^{(y-\delta)h}$, so the drift is $(y - \delta)h$ over the short time interval h .*

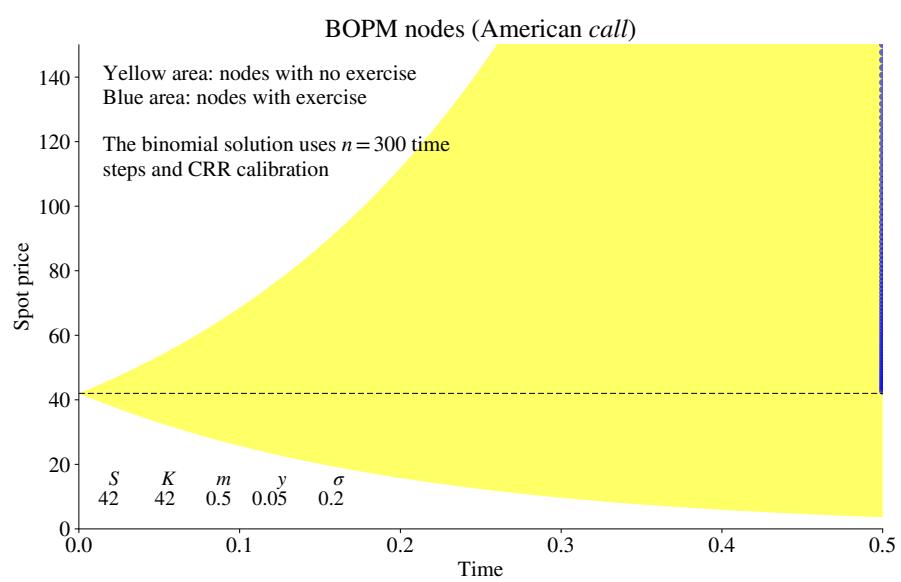


Figure 20.23: Numerical solution of an American call price

Chapter 21

The Black-Scholes Model

This chapter presents, derives and tests the Black-Scholes, also known as the Black-Scholes-Merton model, (see [Black and Scholes \(1973\)](#) and [Merton \(1973b\)](#)), which applies to European options. The basic model assumption is that the log price of the underlying asset is normally distributed, which turns out to be the limiting case for the binomial model as the number of time steps increases (for a fixed time to expiration).

21.1 The Black-Scholes Model

21.1.1 The Basic Black-Scholes Model (No Dividends)

The Black-Scholes (B-S) formula for the price of a *European call option* on an underlying asset *without dividends* is

$$C = S\Phi(d_1) - e^{-ym} K\Phi(d_2), \text{ where} \quad (21.1)$$

$$d_1 = \frac{\ln(S/K) + (y + \sigma^2/2)m}{\sigma\sqrt{m}} \text{ and } d_2 = d_1 - \sigma\sqrt{m}, \quad (21.2)$$

where m is the time to expiration, y the interest rate, S the current asset price, K the strike price and σ^2 the annualised variance of the return on the asset (to be discussed in detail further on). Also, $\Phi(d)$ denotes the probability of $x \leq d$ when x has an $N(0, 1)$ distribution, that is, the value of the standard normal distribution function at d . See Figure 21.1 and also an appendix for numerical values. The background (derivation) of the model is discussed below.

Figures 21.2–21.3 show that the Black-Scholes call option price at expiration coincides with the payoff functions (also demonstrated algebraically in an appendix), and how it differs from them as the time to expiration increases. In particular, the option price is

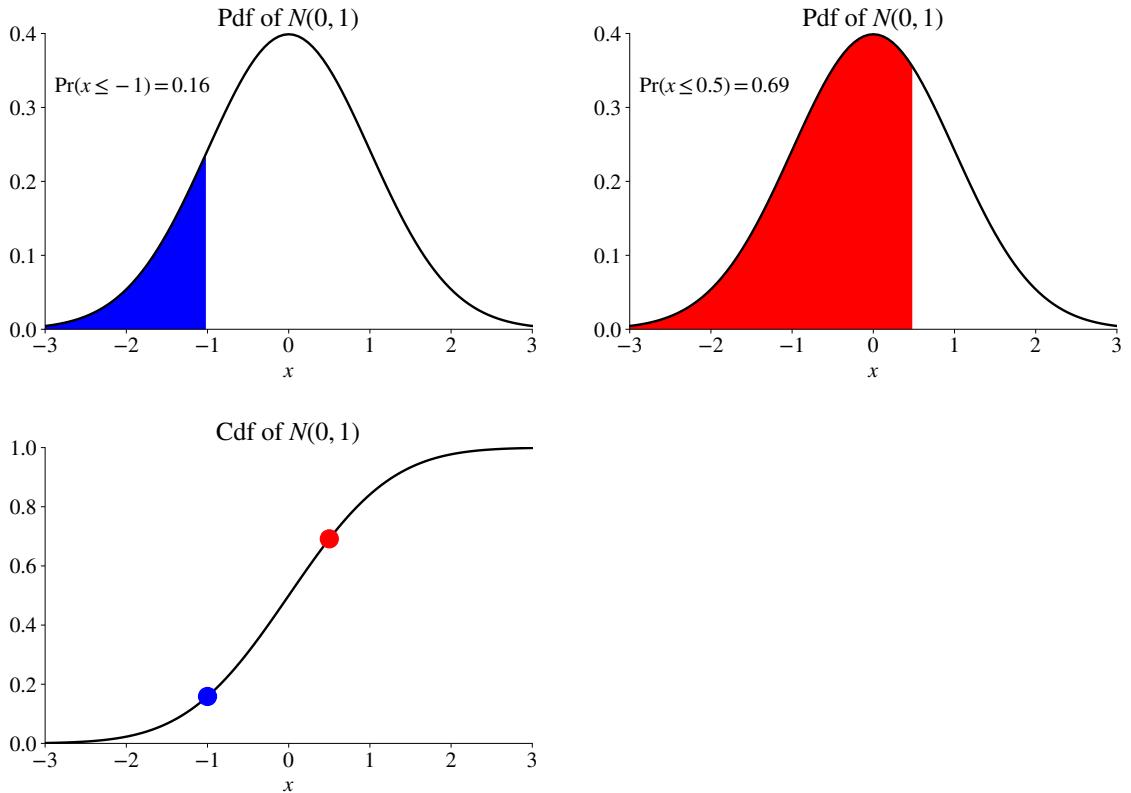


Figure 21.1: Pdf and cdf of $N(0, 1)$

increasing in the current asset price, volatility (σ), time to maturity and the interest rate, but decreasing in the strike price.

Example 21.1 (*Call option price*) With $(S, K, y, m, \sigma) = (42, 42, 0.05, 0.5, 0.2)$, (21.1)–(21.2) give $C = 2.893$.

Applying the put-call parity (21.1), the pricing formula for a put option is

$$P = e^{-ym} K \Phi(-d_2) - S \Phi(-d_1), \quad (21.3)$$

where d_1 and d_2 are defined in (21.2).

Proof (of (21.3)) Recall that the put-call parity for an asset without dividends is $C - P = S - e^{-my} K$. Use in (21.1) to get

$$P = S[\Phi(d_1) - 1] - e^{-ym} K[\Phi(d_2) - 1].$$

Since $\Phi(d) + \Phi(-d) = 1$, this can be written as (21.3). \square

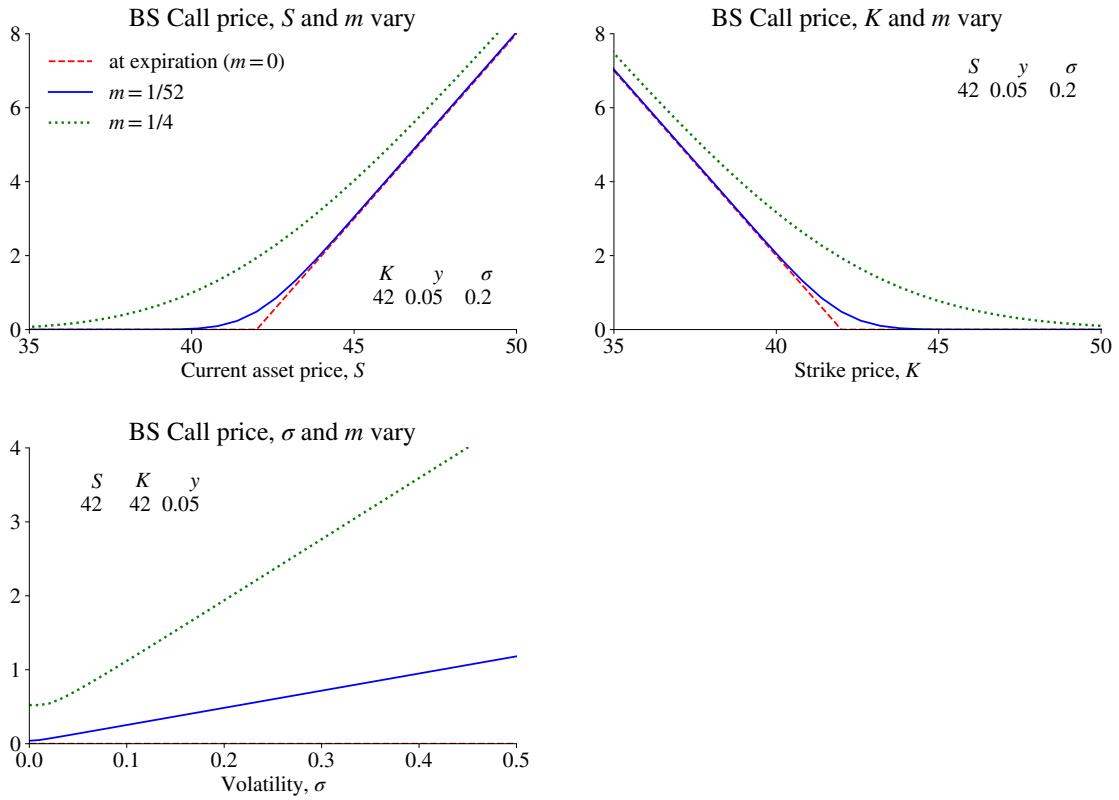


Figure 21.2: Call option price, Black-Scholes model

21.1.2 The Black-Scholes Model with Dividends

Consider a European option for an underlying asset that pays dividends before expiration. The Black-Scholes formula is then no longer valid. The basic reason is that the current price of the underlying embeds all future dividends, but the option will miss out on those dividends that are paid before the expiration.

To handle this, we could apply the B-S formula to a forward contract on the underlying, expiring on the same day as the option. The point is that the forward also misses out on the dividends. Let a prepaid forward contract, which is the present value of forward price, $e^{-ym}F$, substitute for the underlying asset price S in the B-S formula (21.1)–(21.2)

$$C = e^{-ym}F\Phi(d_1) - e^{-ym}K\Phi(d_2), \text{ where} \quad (21.4)$$

$$d_1 = \frac{\ln(F/K) + (\sigma^2/2)m}{\sigma\sqrt{m}} \text{ and } d_2 = d_1 - \sigma\sqrt{m}. \quad (21.5)$$

This is *Black's model* (see Black (1976)) which has many applications. The put option

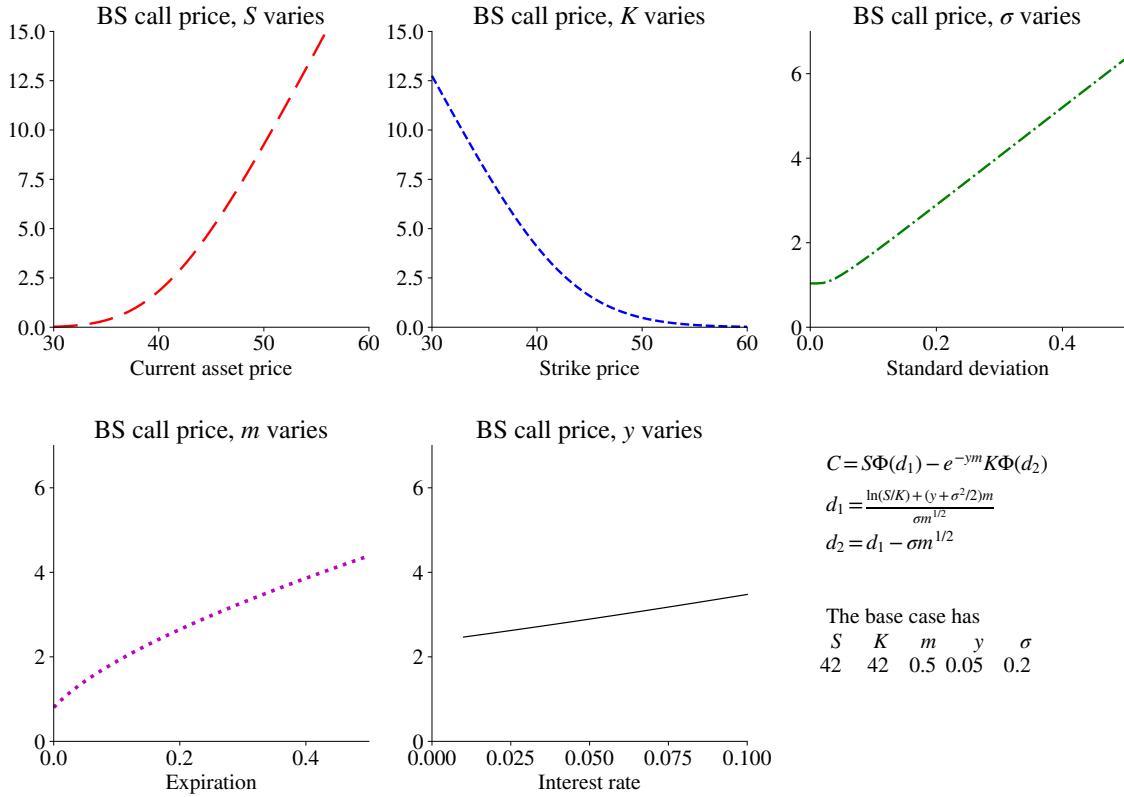


Figure 21.3: Call option price, Black-Scholes model

price is

$$P = e^{-ym} K\Phi(-d_2) - e^{-ym} F\Phi(-d_1). \quad (21.6)$$

For instance, for an asset with a continuous dividend rate of δ , the forward-spot parity says $F = Se^{(y-\delta)m}$. In this case (21.4)–(21.5) can also be written

$$C = e^{-\delta m} S\Phi(d_1) - e^{-ym} K\Phi(d_2), \text{ where} \quad (21.7)$$

$$d_1 = \frac{\ln(S/K) + (y - \delta + \sigma^2/2)m}{\sigma \sqrt{m}} \text{ and } d_2 = d_1 - \sigma \sqrt{m}. \quad (21.8)$$

Example 21.2 (Put price) Using the same parameters as in Example 21.1 and $\delta = 0$, we get $P = 1.856$. Instead, with $\delta = 0.05$, we get $P = 2.309$.

When the asset is a currency (read: foreign money market account) and δ is the foreign interest rate, then this is the Garman and Kohlhagen (1983) model. The put price is

$$P = e^{-ym} K\Phi(-d_2) - e^{-\delta m} S\Phi(-d_1). \quad (21.9)$$

See McDonald (2014) 15–16 and Hull (2022) 15–17 for more detailed discussions.

21.2 Deriving B-S I: Risk Neutral Pricing

We know that the risk neutral pricing of a European call option is

$$C = e^{-ym} \mathbb{E}^* \max(0, S_m - K), \quad (21.10)$$

where \mathbb{E}^* denotes the expectation according to the risk neutral distribution. We can express this as

$$C = e^{-ym} \int_K^\infty \max(0, S_m - K) f^*(S_m) dS_m, \quad (21.11)$$

where $f^*(S_m)$ is the risk neutral density function of the asset price at expiration (S_m). (Below K the value of the integrand is zero.)

We obtain the Black-Scholes price (21.1)–(21.2) if we solve the integral (21.11), assuming the following risk neutral distribution of $\ln S_m$

$$\ln S_m \sim^* N(\ln S + my - m\sigma^2/2, m\sigma^2), \quad (21.12)$$

where S is the current asset price. The probability density function $f^*(S_m)$ is obtained by a change-of-variable, from $\ln S_m$ to S_m . (A proof of is in an Appendix.)

We can alternatively calculate (21.11) by numerical integration to verify that we get the same value as from the Black-Scholes formula. See Figure 21.4 for an illustration.

Remark 21.3 (*Background to the risk neutral distribution in (21.12)**) If the (risk neutral) process for the log asset price is

$$\ln S_{t+h} - \ln S_t = h(y - \sigma^2/2) + \sqrt{h}\sigma\varepsilon_{t+h}, \text{ with } \varepsilon_{t+h} \sim \text{iid } N(0, 1),$$

then the distribution of $\ln S_m = \ln S_0 + \sum_{i=1}^n (\ln S_{ih} - \ln S_{(i-1)h})$ is as in (21.12). See Figure 21.5 for an illustration.

Remark 21.4 (*B-S from a stochastic discount factor**) Let M be a stochastic discount factor that satisfies $P = \mathbb{E} M x$ for every asset, where P is the asset price and x the payoff of the asset. Then, $C = \mathbb{E} M \max(0, S_m - K)$ gives the Black-Scholes formula if $(\ln M, \ln S_m)$ has a joint normal distribution. (See, for instance, Söderlind and Svensson (1997) for a proof.)

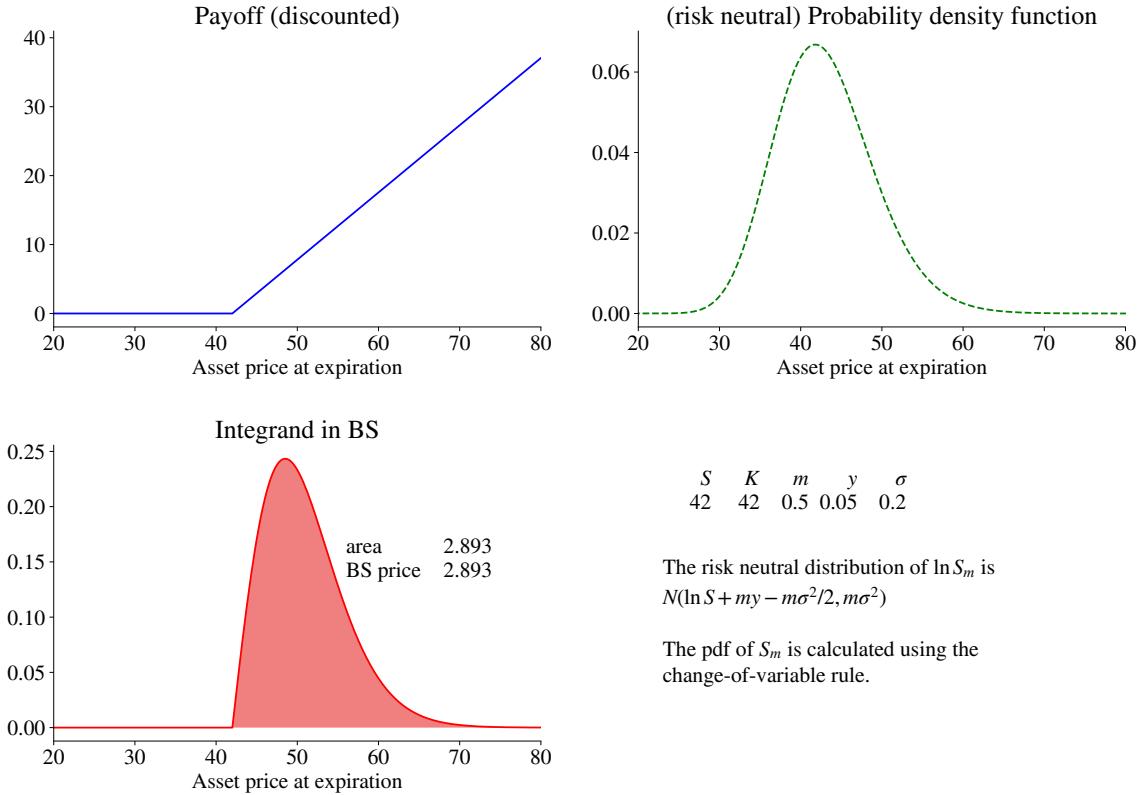


Figure 21.4: Numerical integration to the B-S call price

21.3 Deriving B-S II: Convergence of the BOPM

21.3.1 The Main Result

This section demonstrates that the option price from the binomial option pricing model (BOPM) converges to the price from the Black-Scholes model as we take more (but shorter) time steps to reach a fixed time to expiration m . See Figures 21.6–21.7 for an illustration of how the parameters (p, u, d) from the CRR approach and the resulting option price converge.

In the binomial option pricing model (BOPM), the risk neutral binomial process for the asset price gives the following binomial process for the *log returns* (changes of the log asset price)

$$r_{t+h} = \ln(S_{t+h}/S_t) = \begin{cases} \ln u & \text{with probability } p \\ \ln d & \text{with probability } 1-p. \end{cases} \quad (21.13)$$

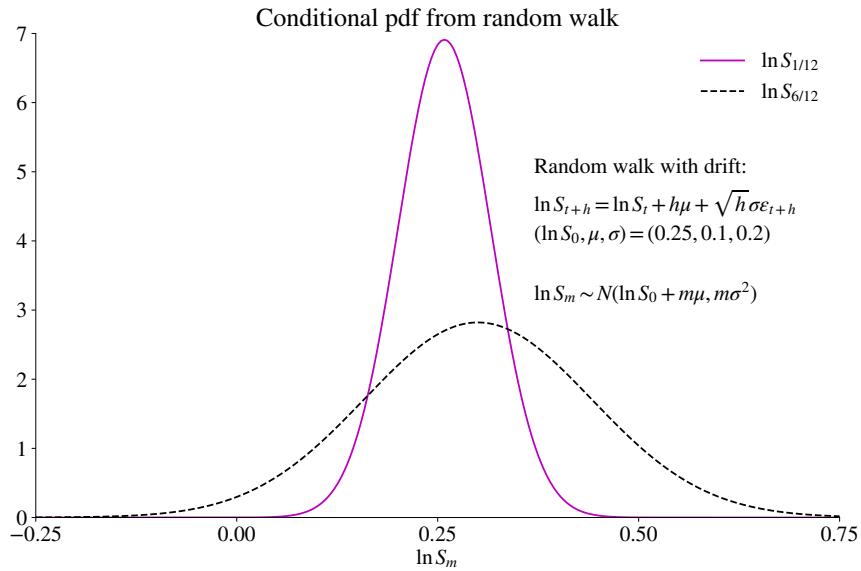


Figure 21.5: Conditional distribution from random walk with drift

The parameters u , d and p all depend on the time step length h . With the CRR approach, they are chosen so that the mean and variance of r_{t+h} are (at least in the limit) proportional to h .

Clearly, the binomial tree means that we reach $\ln S_m$ by starting at S_0 and adding n steps of the kind in (21.13)

$$\ln S_m = \ln S_0 + \ln(S_h/S_0) + \dots + \ln(S_{nh}/S_{(n-1)h}) \quad (21.14)$$

$$= \ln S_0 + \sum_{i=1}^n r_i, \quad (21.15)$$

where r_i is the log return between $(i-1)h$ and ih . Notice that the r_i are iid (same distribution for each i , and r_i and r_j are independent).

We demonstrate the convergence of this to the Black-Scholes risk neutral distribution (21.12) in two steps: first, that the binomial distribution converges to a normal distribution; and second that both distributions have the same mean and variance in the limit.

21.3.2 The Central Limit Theorem at Work

The Black-Scholes model is based on normally distributed changes of log prices. In the binomial model, the log price changes can only take two values, but the sum of many such changes will converge to a normally distributed variable as the number of time steps

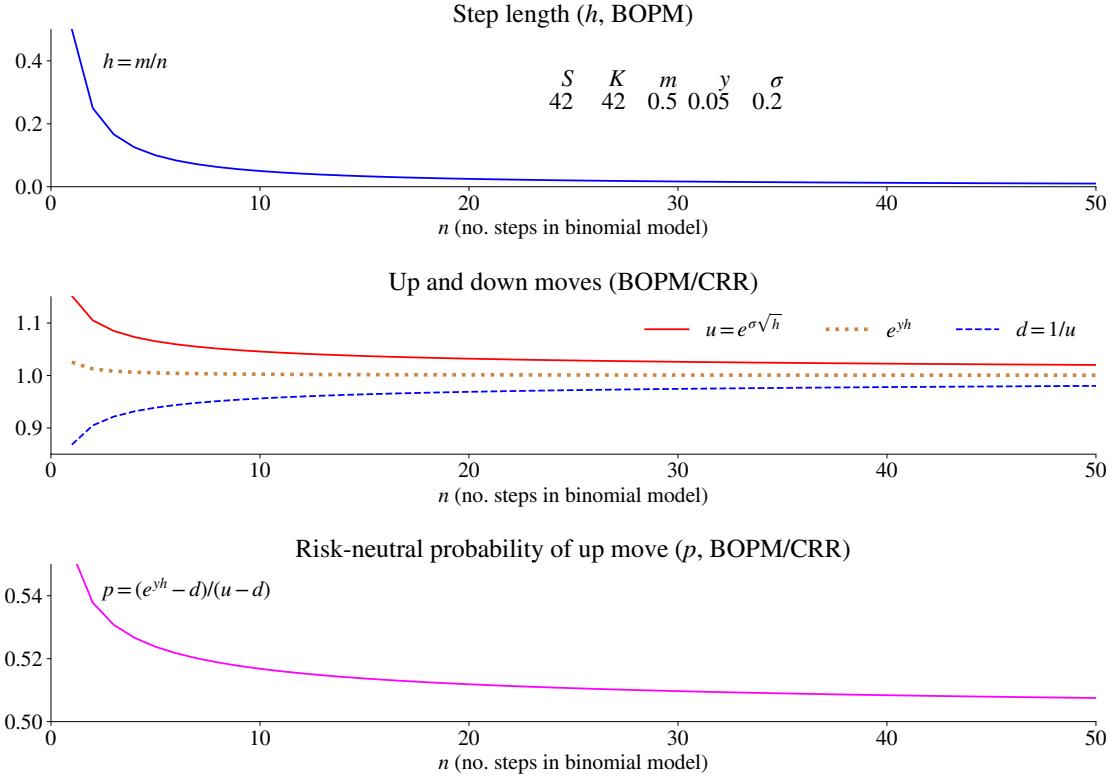


Figure 21.6: Convergence of the parameters in a binomial model

increases. This may seem counter intuitive since central limit theorems apply to sample averages times the square root of the sample size, not to sums. However, the rescaling of (u, d, p) as the number of time steps increases, implies that the sum is effectively a (scaled) sample average, so a CLT indeed applies.

See Figure 21.8 for an example of how the distribution converges. Notice that the figure shows the density functions for the *log* asset price (at expiration). Also, the discrete distribution from the binomial model is illustrated by bars centered on the outcome, normalised to have an area of one. The next proposition formalises this, and it applies in the limit to the CRR approach.

Proposition 21.5 *If u, d and p in the binomial process (21.13) are such that the mean and variance of $\ln S_{t+h} - \ln S_t$ are proportional to h , then the distribution of $\sum_{i=1}^n r_i$ converges to a normal distribution as the number of time steps n increases, keeping the maturity m constant (so $h = m/n$).*

Proof (*of Proposition 21.5) The binomial model (21.13)–(21.15) means that we can

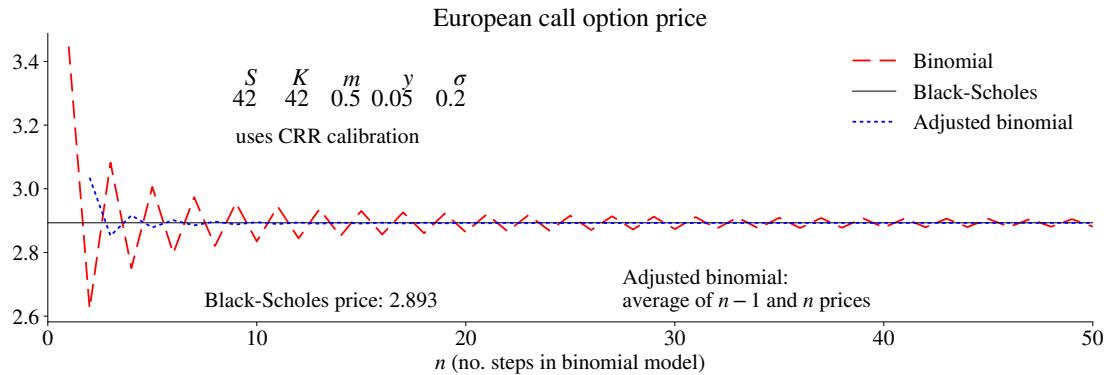


Figure 21.7: Convergence of the binomial price to the Black-Scholes price

write return $r_i = \varepsilon_i \sqrt{h} + \mu h$, where ε_i is an iid zero mean random variable with variance σ^2 . Notice that $E r_i = \mu h$ and $Var(r_i) = \sigma^2 h$, so both moments are proportional to h . Write (21.15) as

$$\sum_{i=1}^n r_i = \sqrt{h} \sum_{i=1}^n \varepsilon_i + nh\mu.$$

Since $h = m/n$, this can be written

$$\sum_{i=1}^n r_i = \sqrt{m} \underbrace{\sqrt{n} \frac{1}{n} \sum_{i=1}^n \varepsilon_i}_A + \mu m.$$

The term A is \sqrt{n} times the sample average of an iid random variable (ε_i) with $E \varepsilon_i = 0$ and $Var(\varepsilon_i) = \sigma^2 < \infty$. We can therefore apply the (Lindeberg-Lévy) central limit theorem to show that $A \xrightarrow{d} N(0, \sigma^2)$. The second term (μm) is just a constant. Together, we get that $\sum_{i=1}^n r_i \xrightarrow{d} N(\mu m, \sigma^2 m)$. \square

21.3.3 Convergence of the Mean and Variance

This section demonstrates that the mean and variance of the binomial distribution converge to the same values as in the risk neutral distribution of the Black-Scholes model (21.12). See Figure 21.9 for an illustration.

Proposition 21.6 (*Moments of CRR steps*) In the Cox, Ross, and Rubinstein (1979) tree, the parameters in (21.13) are

$$\ln u = \sigma \sqrt{h}, \ln d = -\sigma \sqrt{h} \text{ and } p = (e^{yh} - d)/(u - d).$$

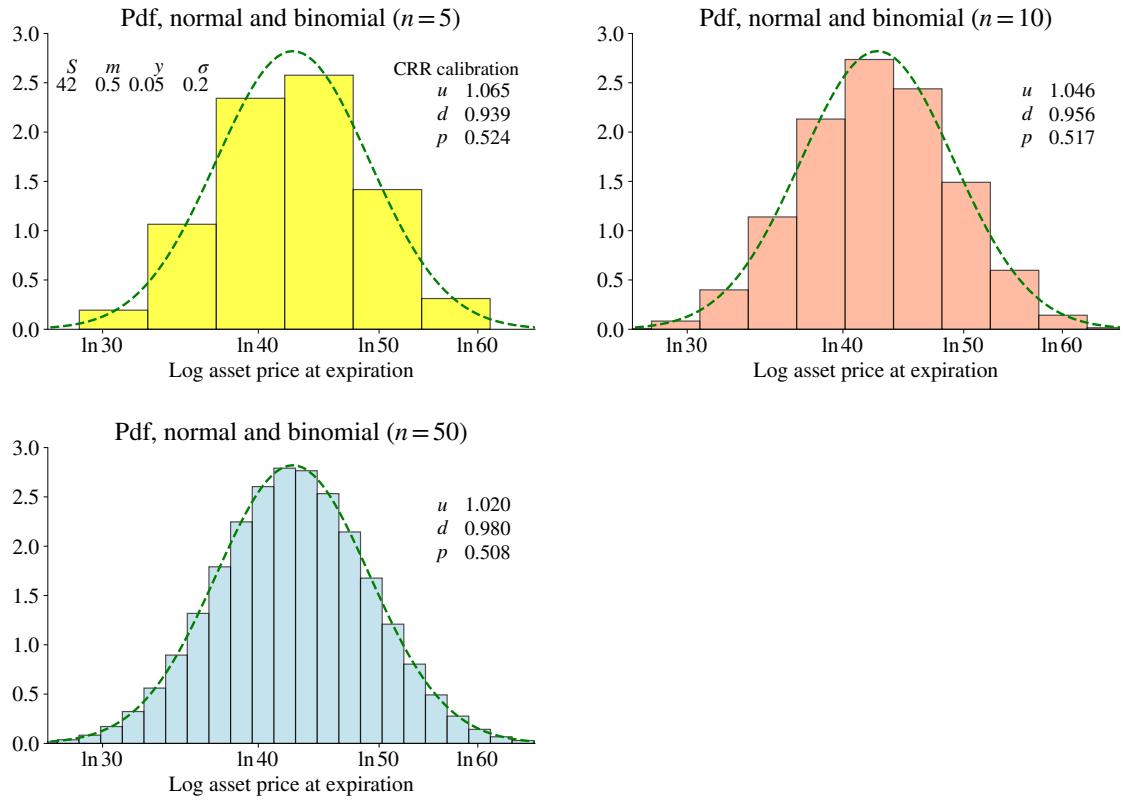


Figure 21.8: Convergence of the binomial model to the Black-Scholes model. The figure shows results for the log asset price. The (risk neutral) distribution from the binomial distribution is scaled so the area of the bars equals one.

As $n \rightarrow \infty$, but $h = m/n$ we have (since the price changes are independent)

$$E \sum_{i=1}^n r_i = m(y - \sigma^2/2) \text{ and } \text{Var}(\sum_{i=1}^n r_i) = m\sigma^2.$$

This is the same as in the risk neutral distribution of the Black-Scholes model.

Proof (*of Proposition 21.6) Recall that the mean and variance of r_i are $p \ln u + (1-p) \ln d$ and $p(1-p)(\ln u - \ln d)^2$ respectively. Since the terms in (21.15) are uncorrelated, the mean and the variance of the sum are $n E r_i$ and $n \text{Var}(r_i)$. Substitute for u, d and p and take the limits of as $n \rightarrow \infty$, but $h = m/n$. (This is straightforward, but slightly messy, calculation.) \square

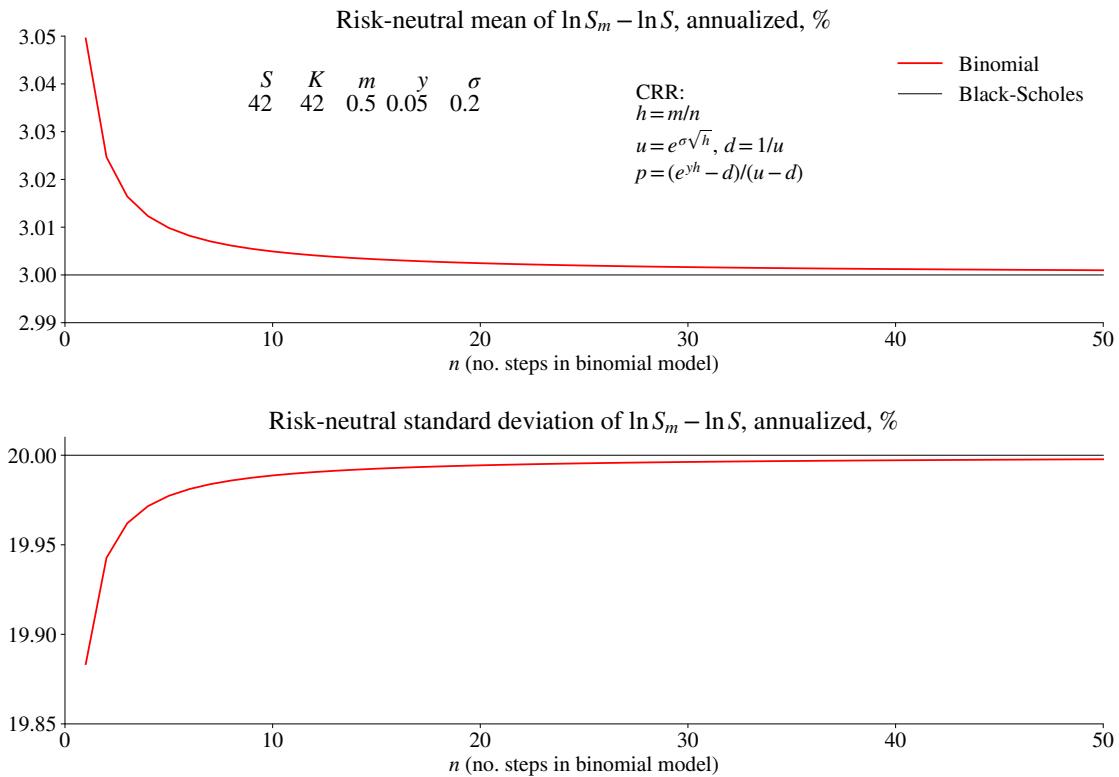


Figure 21.9: Convergence of the binomial mean and variance

21.4 Testing the B-S Model

The Black-Scholes formula (21.1)–(21.2) for a European call option contains only one unknown parameter: the standard deviation σ of the distribution of $\ln S_m$. With data on the option price, spot and forward prices, the interest rate, and the strike price, we can solve for σ (see from Figure 21.3 that the option price and the volatility have a monotonic relation).

The σ calculated in this way is called the *implied volatility* and it is often used as an indicator of market uncertainty about the future asset price, S_m . It can be thought of as an annualized standard deviation. You can also calculate the implied volatility from a put option, since the put-call parity shows that a call and a put with the same strike price have the same implied volatility.

Empirical Example 21.7 Figure 21.10 shows how the VIX has changed since it was first introduced. It is an average of implied volatilities of 30-day S&P 500 (close to) atm options.

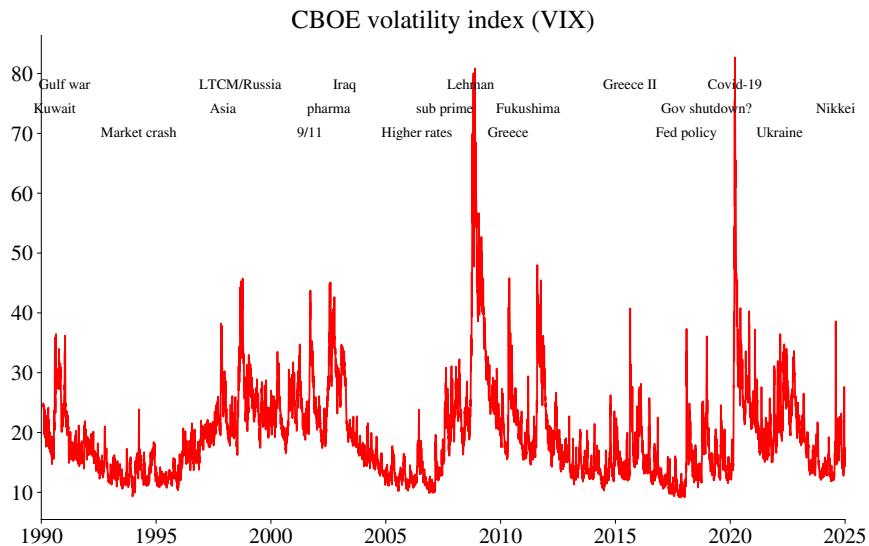


Figure 21.10: CBOE VIX, summary measure of implied volatilities (30 days) on US stock markets

Note that we can solve for one implied volatility for each available strike price. If the Black-Scholes formula is correct, then these volatilities should be the same across strike prices.

However, on currency markets, we often find a volatility “smile” (volatility is a U-shaped function of the strike price). One possible explanation is that the (perceived) distribution of the future asset price has relatively more probability mass in the tails (“fat tails”) than a normal distribution has. (Recall, Black-Scholes is built on the assumption of a normal distribution.)

On equity markets, we often find a volatility “smirk” instead, where the volatility is very high for very low strike prices. This is often interpreted as meaning that investors are willing to pay a premium for put options that protect them from a dramatic fall in the stock price. One possible explanation is thus that the distribution has more probability mass than a normal distribution at very low stock prices (negative skewness).

Empirical Example 21.8 See Figures 21.11–21.12 show implied volatility and strike prices for S&P 500 options.

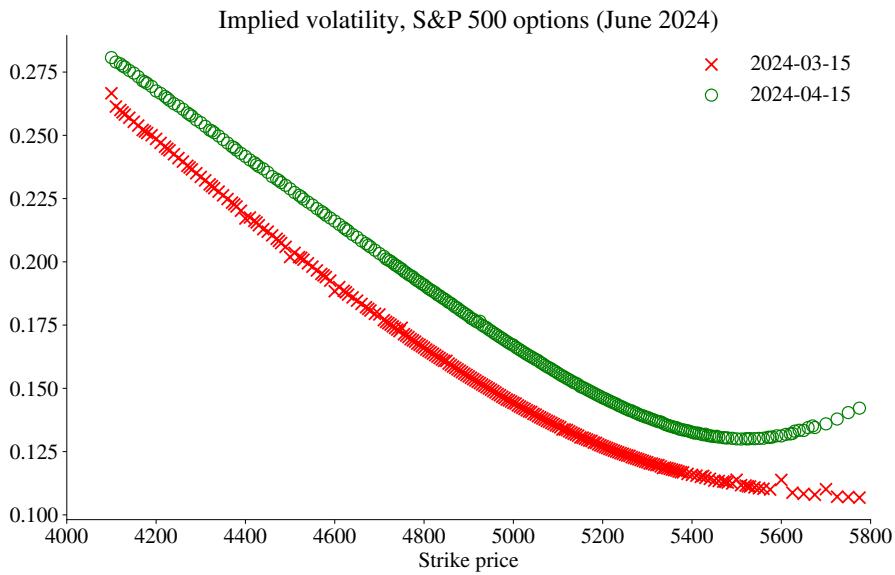


Figure 21.11: Implied volatilities of S&P 500 options, selected dates

21.5 Appendix – Details on the B-S Model*

21.5.1 Limits of the Black-Scholes Formula when $\sigma = 0$ or $m = 0$

Remark 21.9 (*Black-Scholes formula when $\sigma = 0$ **) From (21.2) $\lim_{\sigma \rightarrow 0} d_1 = \lim_{\sigma \rightarrow 0} d_2 = \infty$ if $e^{ym}S \geq K$ and $-\infty$ otherwise. Therefore, $\lim_{\sigma \rightarrow 0} \Phi(d_1) = \lim_{\sigma \rightarrow 0} \Phi(d_2) = 1$ if $e^{ym}S \geq K$ and 0 otherwise. The Black-Scholes call option price at $\sigma = 0$ is therefore $\max(S - e^{-ym}K, 0)$.

Remark 21.10 (*Call option price when $\sigma = 0$, version 2**) When the underlying asset is risk-free ($\sigma = 0$), then its return must equal the risk-free rate y , so the value of the underlying asset is $e^{ym}S$ at expiration. The present value of the known call payoff is $e^{-ym} \max(e^{ym}S - K, 0)$, which is the same as in the previous remark.

Remark 21.11 (*Black-Scholes formula when $m = 0$ **) From (21.2) $\lim_{m \rightarrow 0} d_1 = \lim_{m \rightarrow 0} d_2 = \infty$ if $S \geq K$ and $-\infty$ otherwise. Therefore, $\lim_{m \rightarrow 0} \Phi(d_1) = \lim_{m \rightarrow 0} \Phi(d_2) = 1$ if $S \geq K$ and 0 otherwise. The Black-Scholes call option price at $m = 0$ is therefore $\max(S - K, 0)$.

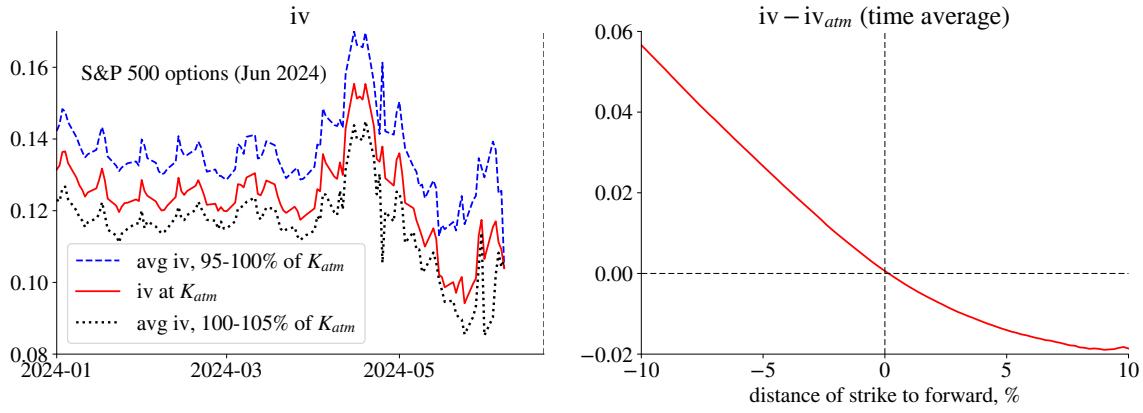


Figure 21.12: Implied volatilities over time

21.5.2 Calculating Black's model with Computer Code for the Black-Scholes Model

Remark 21.12 (*Coding Black's model with a forward price**) Suppose you have a computer code for the B-S model (21.1)–(21.2) which takes the inputs (S, K, y, m, σ) . To use that code for Black's model (21.4)–(21.5), substitute $(F, 0)$ for (S, y) and multiply the results by e^{-ym} .

Remark 21.13 (*Coding the B-S model with continuous dividends**) Suppose you have a computer code for the B-S model (21.1)–(21.2) which takes the inputs (S, K, y, m, σ) . To use that code for Black's model (21.4)–(21.5), substitute $e^{-\delta m} S$ for S .

Remark 21.14 (*Practical hint: finding the dividend rate**) If you don't know what the dividend rate is, use the forward-spot parity, $F = Se^{(y-\delta)m}$, to calculate it as $\delta = y - \ln(F/S)/m$.

21.6 Appendix – Probabilities in the BOPM and B-S Models*

The price of a European (call or put) option calculated by the binomial model converges to the Black-Scholes price as the number of subintervals increases (keeping the time to expiration constant, so the subintervals become shorter). This is illustrated in Figure 21.7.

Both the binomial option pricing model (BOPM) and the Black-Scholes model imply that the call option price can be written as the discounted risk neutral expected payoff (21.10). We can clearly rewrite (21.10) as

$$C = e^{-ym} E^*(S_m - K | S_m > K) \Pr^*(S_m > K) \quad (21.16)$$

$$= e^{-ym} E^*(S_m | S_m > K) \Pr^*(S_m > K) - e^{-ym} K \Pr^*(S_m > K). \quad (21.17)$$

The first term is (the present value of) the risk neutral expected asset price conditional on exercise, times the risk neutral probability of exercise. The second term is (the present value of) the strike price times the risk neutral probability of exercise.

Example 21.15 (*Binomial model with $n = 2$*) The price of a European call option is

$$C = e^{-ym} [p^2 \max(S_{uu} - K, 0) + 2p(1-p) \max(S_{ud} - K, 0) + (1-p)^2 \max(S_{dd} - K, 0)].$$

Suppose we only exercise in the S_{uu} node ($S_{uu} > K$ but $S_{ud} < K$). The call price can then be written

$$\begin{aligned} C &= e^{-ym} p^2 (S_{uu} - K) \\ &= e^{-ym} \underbrace{\Pr^*(S_m > K)}_{E^*(S_m | S_m > K)} \underbrace{p^2}_{\Pr^*(S_{uu})} - e^{-ym} K \underbrace{p^2}_{\Pr^*(S_{uu})}. \end{aligned}$$

Remark 21.16 (*Properties of a lognormal distribution*) Let $x \sim N(\mu, s^2)$ and define $k_0 = (\ln K - \mu)/s$. First, $\Pr(e^x > K) = \Phi(-k_0)$. Second, $E(e^x | e^x > K) = e^{\mu+s^2/2} \Phi(s - k_0)/\Phi(-k_0)$. (To prove this, just integrate.)

Proposition 21.17 (*Riskneutral probability of $S_m > K$*) The $\Phi(d_2)$ term in the Black-Scholes formula (21.1)–(21.2) is the risk-neutral probability that $S_m > K$.

Proposition 21.18 (*$S\Phi(d_1)$ in Black-Scholes*) The $S\Phi(d_1)$ term in the Black-Scholes formula (21.1)–(21.2) is (the present value of) the expected asset price conditional on exercise, times the probability of exercise, that is, the first term in (21.17).

Proof (of Proposition 21.17) The risk neutral probability of $\ln S_m$ is $N(\mu, s^2)$ with $\mu = \ln S + ym - \sigma^2 m/2$ and $s^2 = \sigma^2 m$. Use Remark 21.16 to calculate the probability $\Pr(S_m > K)$ as $\Phi(-k_0)$ where $k_0 = (\ln K - \mu)/s$. Clearly, $-k_0$ is the same as d_2 in (21.2). \square

Proof (of Proposition 21.18) Using Remark 21.16, the first term in (21.17), here denoted A , can be written

$$A = e^{-ym} e^{\mu+s^2/2} \Phi(s - k_0),$$

since the two $\Phi(-k_0)$ terms cancel. Since $\ln S_m$ is $N(\mu, s^2)$ with $\mu = \ln S + ym - \sigma^2 m/2$ and $s^2 = \sigma^2 m$. we get

$$\mu + s^2/2 = \ln S + ym, \text{ and}$$

$$s - k_0 = \sigma \sqrt{m} - \frac{\ln K - (\ln S + ym - \sigma^2 m/2)}{\sigma \sqrt{m}} = d_1,$$

where the last line follows from comparing with (21.2). We can therefore write A as $S\Phi(d_1)$, since the $e^{-ym} e^{ym}$ term cancels. This is the same as in the Black-Scholes formula. \square

21.7 Appendix – Statistical Tables

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

Table 21.1: Values of the standard normal cumulative distribution function at x where x is the sum of the values in the first column and the first row.

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

Table 21.2: Values of the standard normal cumulative distribution function at x where x is the sum of the values in the first column and the first row.

Chapter 22

Hedging Options

This chapter shows how an option portfolio we can be hedged. The setting is that we have written (sold, issued) such an option portfolio, but we do not want to carry the risk.

22.1 Hedging an Option

A first order approximation suggests that the change (here indicated by d) in the option portfolio value (denoted L) due to a change in the underlying price is

$$dL \approx \frac{\partial L}{\partial S} dS. \quad (22.1)$$

When the option portfolio consists of a call option only, then $L = C$ (where C is the call option price) and the derivative is positive, see Figure 22.9.

Example 22.1 (*Option portfolios*) If we have issued one call option, then $L = C$ where C is the call option price. Instead, if we have issued 3 call options and bought 2 put options, then $L = 3C - 2P$, where P is the put price.

Remark 22.2 (*dX notation*) Warning: this section uses dX to indicate a change in variable X , mostly since Δ has another, and well established, interpretation in the option literature.

22.2 An Approximate Hedge

22.2.1 Basic Setup

Consider a portfolio which is long v units of the underlying asset (the hedging portfolio) and short one option portfolio (with value L). The value of the overall position is

$$V = vS + M - L, \quad (22.2)$$

where M is a money market account. The idea is to find v so that vS and L are equally sensitive to changes in S . (A long option position can be handled by $L < 0$.)

For now, we focus on movements of the price of the underlying, disregarding, for instance, movements in volatility and the value of the money market account. Use (22.1) to approximate the change (indicated by d) of the value of the overall portfolio as

$$\begin{aligned} dV &\approx vds - \frac{\partial L}{\partial S}ds \\ &\approx 0 \text{ if } v = \frac{\partial L}{\partial S} = \Delta, \end{aligned} \quad (22.3)$$

where the second line uses Δ as a symbol for the derivative $\partial L/\partial S$ (as is standard in the option literature). This approach makes the overall portfolio *delta neutral*, $\partial V/\partial S = 0$, and is therefore called a *delta hedge*. See See Hull (2022) 13 and McDonald (2014) 15–16 for more detailed treatments.

See Figure 22.1 for how the Black-Scholes option prices and their derivatives depend on the underlying asset price. Note that the derivative is positive for a call option and negative for a put option.

Example 22.3 (*Delta hedging a call or a put*) Suppose $\partial C/\partial S = 0.6$ and $\partial P/\partial S = -0.4$. If we have issued a call option, we buy $v = 0.6$ units of the underlying asset to be hedged, and if we have issued a put option, then we short-sell $v = -0.4$. Instead, if we have bought (not issued) those options, then we short-sell $v = -0.6$ to hedge the long call option and buy $v = 0.4$ to hedge the long put option.

Example 22.4 (*Delta hedging option portfolios*) Continuing Example 22.3, if we have issued 3 call options and bought 2 put options, then $\Delta = 3 \times 0.6 + (-2) \times (-0.4) = 2.6$, so we need to buy $v = 2.6$ units of the underlying.

The delta will change over time, necessitating portfolio rebalancing. In practice, the overall portfolio includes a position in a short-term money market account to make the

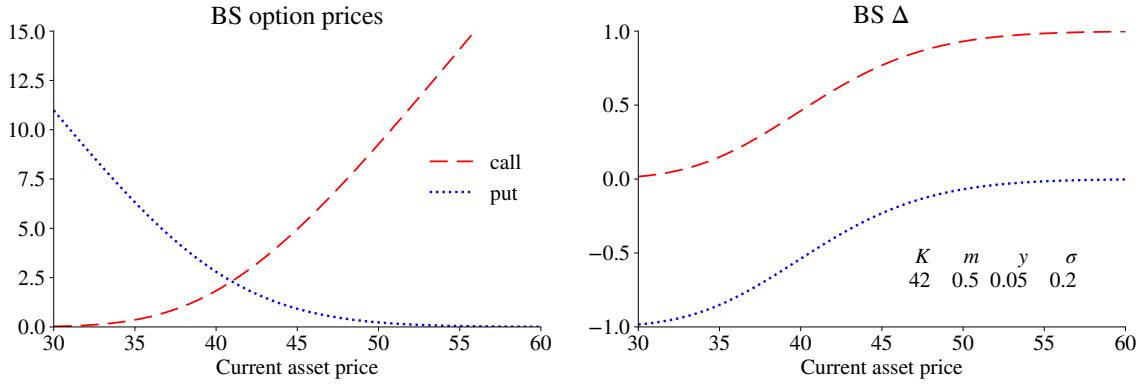


Figure 22.1: Option prices and deltas from the Black-Scholes model

initial portfolio value zero, so

$$M_0 = L_0 - \Delta_0 S_0. \quad (22.4)$$

This position is typically negative for a call option, which means that we finance the purchase of the underlying asset with the proceeds from selling the option and from borrowing.

Remark 22.5 (*Overall portfolio value over several subperiods**) Start by creating a hedge portfolio with a zero initial value as in (22.4). In $t + h$ (say, after one day so $h = 1/252$), this portfolio is worth (this is the marking-to-market)

$$V_{t+h} = \Delta_t (D_{t+h} + S_{t+h}) + M_t e^{y_t h} - L_{t+h},$$

where the underlying pays a dividend ($D_{t+h} = 0$ if no dividends), the prices are measured after dividends and y_t is the interest rate. In $t + h$ we need Δ_{t+h} units of the underlying asset (value $\Delta_{t+h} S_{t+h}$), which we finance with dividend payments and by adding/withdrawing funds from the money market account. See Figure 22.2 for an illustration. In that figure, “m-to-m” stands for the marking-to-market stage (first equation in this remark) and “rebalancing” for the stage after rebalancing the portfolio (second equation in this remark).

22.2.2 Deltas from the Black-Scholes Model

The following remark gives details of the Δ in the Black-Scholes model. (The other derivatives are presented in an appendix.)

Remark 22.6 (*Deltas in Black-Scholes*) The Black-Scholes formula for a European call option on an asset paying continuous dividends (δ) is

$$C = e^{-\delta m} S \Phi(d_1) - e^{-ym} K \Phi(d_2), \text{ where}$$

$$d_1 = \frac{\ln(S/K) + (y - \delta + \sigma^2/2)m}{\sigma \sqrt{m}} \text{ and } d_2 = d_1 - \sigma \sqrt{m}.$$

(Warning: d_1 and d_2 indicate the usual terms in the Black-Scholes formula. Do not confuse with the d used to indicate a change.) The derivatives of the call and put price equations are

$$\Delta = \frac{\partial C}{\partial S} = e^{-\delta m} \Phi(d_1)$$

$$\Delta_p = \frac{\partial P}{\partial S} = \Delta - e^{-\delta m}.$$

where $\phi()$ is the standard normal probability density function (the derivative of $\Phi()$). The result for the put follows from the put-call parity which says $P = C - Se^{-\delta m} + e^{-ym} K$. It is also useful to notice that the sensitivity to a forward price ($F = Se^{(y-\delta)m}$) is $\partial C / \partial F = e^{-ym} \Phi(d_1)$, where d_1 is as above, or $[\ln(F/K) + (\sigma^2/2)m] / (\sigma \sqrt{m})$.

See Figure 22.1 for an illustration of how the Black-Scholes Δ depends on the underlying price . In particular, notice that $0 \leq \Delta \leq 1$ for a call and $-1 \leq \Delta \leq 0$ for a put. In both cases, Δ is increasing with the price of the underlying asset. Intuitively, an option that is deep out of the money will not be very sensitive to the asset price—since the chance of exercising is low. Conversely, a option that is deep in the money moves almost 1:1 in same direction if it is a call option and in the opposite direction if it is a put option.

Example 22.7 (Δ and Δ_p) With $(S, K, m, y, \sigma) = (42, 42, 0.5, 0.05, 0.2)$ and $\delta = 0$, we have $\partial C / \partial S \approx 0.60$ and $\partial P / \partial S \approx -0.40$. The difference is equal to one (since $\delta = 0$).

Example 22.8 (*Delta hedging of a call option*) Using the same parameters as in Example 22.7 and $\delta = 0$, Figure 22.2 illustrates the initial positions (day 0), and two snap shots of the day after (day 1: after marking to market, day 1: after rebalancing). On day 0, the overall portfolio includes $\Delta = 0.6$ of the underlying asset (at a value of $0.6 \times 42 = 25.10$), -1 of the call option (at the value -2.89) and the balance on a money market account ($-25.10 + 2.89 = -22.21$) so the total portfolio is worth zero. This clearly means that the investor has borrowed.

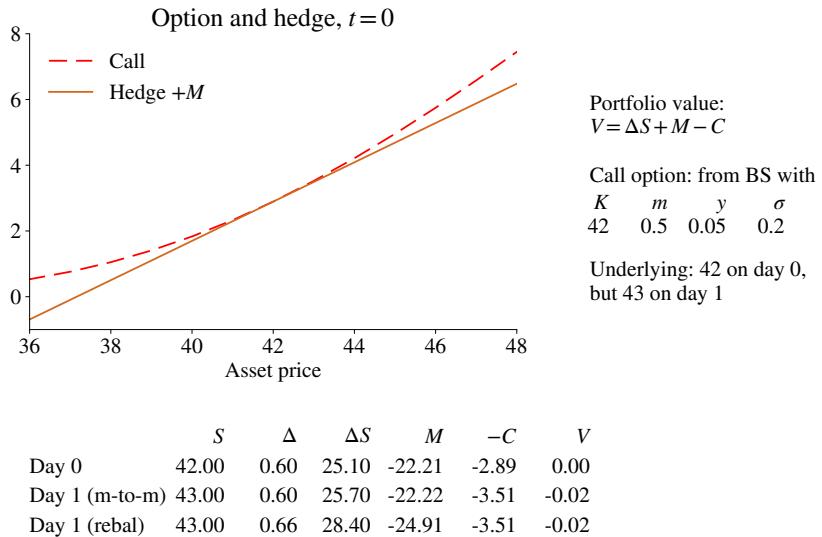


Figure 22.2: Delta hedging over time

Empirical Example 22.9 Figure 22.3 illustrates the hedging of a particular S&P 500 option over 5 months. The overall portfolio is much more stable than the option itself, but there are still some movements left. This suggests that the hedging strategy is largely effective, but remains imperfect.

Remark 22.10 (Hedging with a forward contract*) Consider using a forward contract as hedging instrument. Recall that $W_t = e^{-ym}(F_t - F_\tau)$ is the value of an old forward contract (written in $\tau < t$). The hedge portfolio is $V = vW + M - C$. This portfolio is almost stable if $v = e^{ym}\partial C/\partial F$ (see Remark 22.6 for an expression). To see this, notice that $dV = vdW - dC \approx ve^{-ym}dF - \frac{\partial C_t}{\partial F}dF$.

22.2.3 Deltas from Other Models

The Δ (the derivative in (22.3)) could also be computed from other option pricing models, for instance, the binomial model.

The basic approach is straightforward: consider two different values of the underlying asset (S_a and S_b), use the model to compute the option price at each of them (get $L(S_a)$ and $L(S_b)$) and approximate the derivative with a finite difference ratio: $[L(S_a) - L(S_b)]/(S_a - S_b)$. (Clearly, this crude approach can be improved by using other numerical methods for approximating derivatives.)

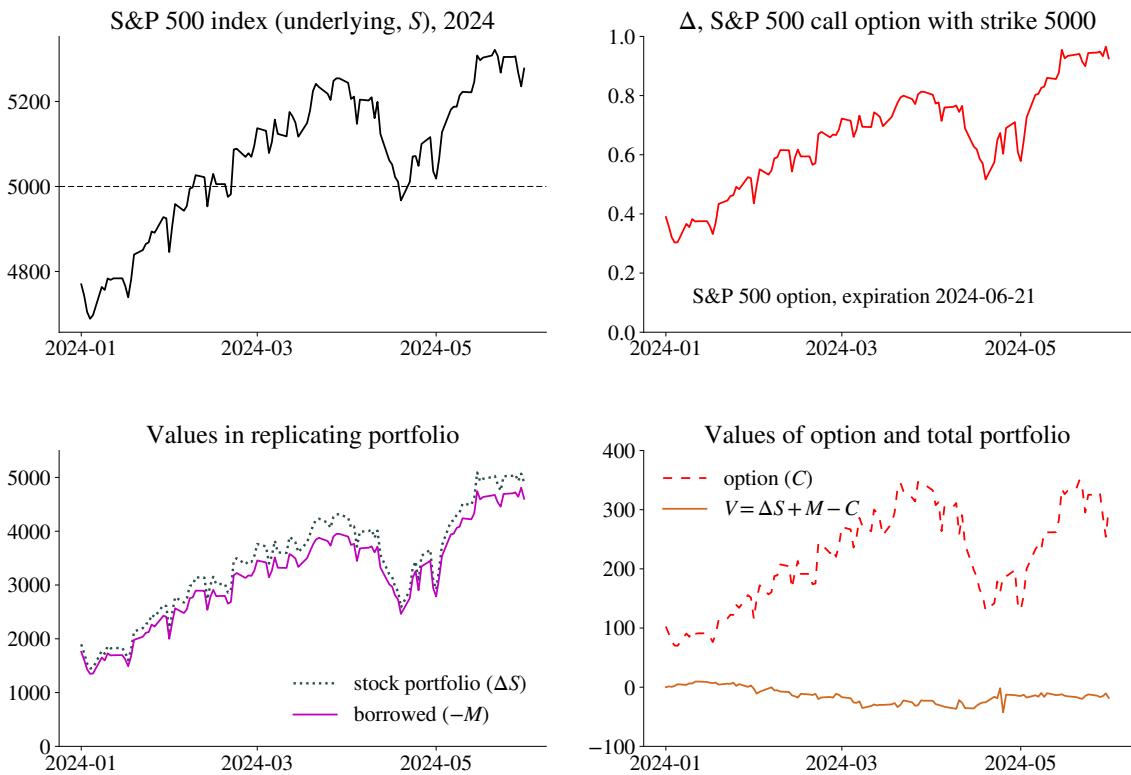


Figure 22.3: Delta hedging an S&P 500 call option

In particular, the binomial model has the advantage that it allows us to handle also American-style options. See Figure 22.4 and notice that the delta of an American put tends to be more negative than for a European put, especially at low prices of the underlying. Also, hedging in the binomial model can be made more precise, which is discussed in an appendix.

22.3 Higher-Order Hedging*

22.3.1 Delta-Gamma Hedging*

Delta hedging can be imprecise if the price of the underlying asset changes a lot or when we try to hedge an option portfolio whose value is a highly non-linear function of the underlying price. As an example of the latter, Figure 22.5 illustrates the price of a straddle (according to Black-Scholes). If the current price of the underlying is close to the strike price, then the (first-order) derivative is zero, but the straddle gains value as soon as the underlying price moves in either direction. In this case, using the underlying to hedge this

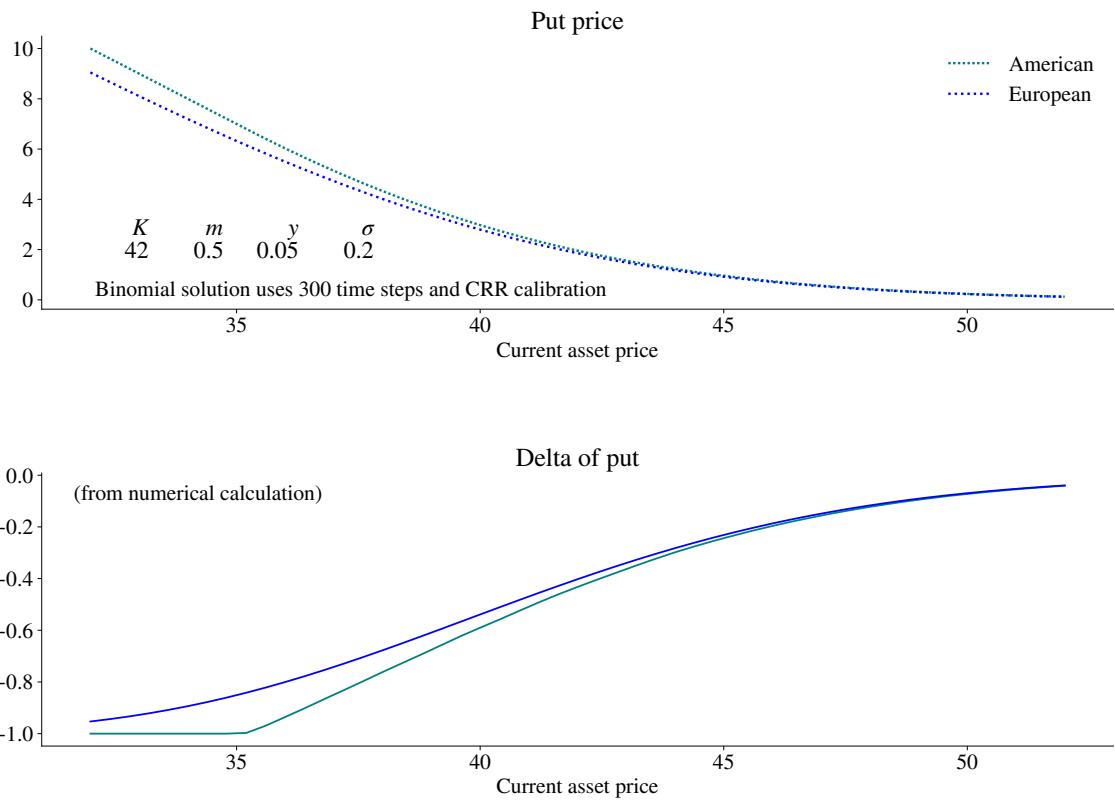


Figure 22.4: The deltas of American and European puts

straddle will not work.

We can improve the precision by using a second-order Taylor approximation of the option portfolio value

$$dL \approx \Delta dS + \frac{1}{2} \Gamma (dS)^2, \text{ where } \Gamma = \frac{\partial^2 L}{\partial S^2}. \quad (22.5)$$

The Γ (upper case gamma) of the Black-Scholes model is presented in an appendix.

To hedge, consider a portfolio with v of the underlying asset, w of another option (or another asset) with a price denoted L^* and short one option portfolio (with price L)

$$V = vS + wL^* - L. \quad (22.6)$$

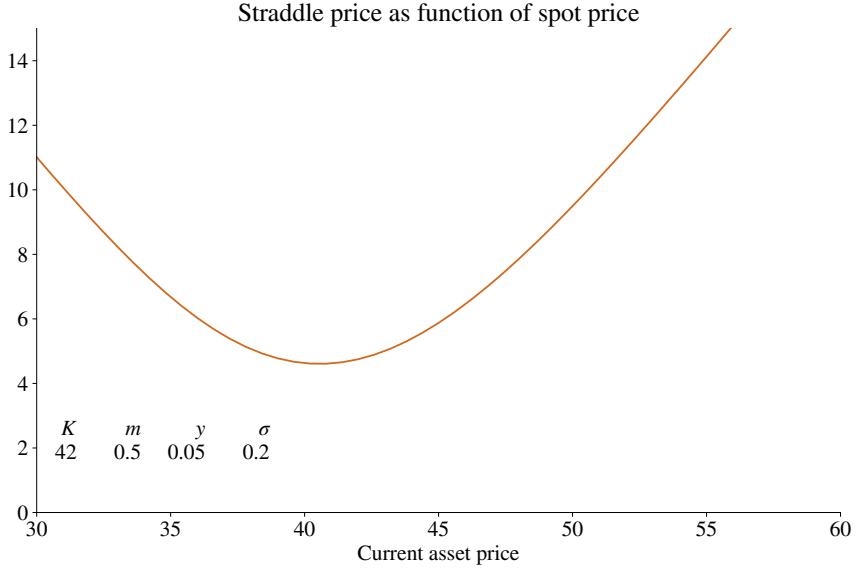


Figure 22.5: Price of a straddle (Black-Scholes)

We get $dV \approx 0$ by setting

$$w = \Gamma/\Gamma^*, \text{ and} \quad (22.7)$$

$$v = \Delta - w\Delta^*, \quad (22.8)$$

where Δ^* and Γ^* are the delta and gamma of L^* .

Proof (of (22.7)–(22.8)) A second-order Taylor approximation like (22.5) of the value of portfolio V gives

$$\begin{aligned} dV &\approx vdS + w[\Delta^*dS + \frac{1}{2}\Gamma^*(dS)^2] - [\Delta dS + \frac{1}{2}\Gamma(dS)^2] \\ &\approx (v + w\Delta^* - \Delta)dS + (w\Gamma^* - \Gamma)\frac{1}{2}(dS)^2. \end{aligned}$$

Using the values (w, v) in (22.7)–(22.8) makes this zero. \square

Example 22.11 (Delta-gamma hedging) Suppose $(\Delta, \Gamma) = (0.5, 0.07)$ and $(\Delta^*, \Gamma^*) = (0.3, 0.03)$, which requires $w = 2.33$ and $v = -0.2$. Clearly, this is quite different from a delta hedge (which has $v = 0.5$ and $w = 0$). Here, the lower sensitivity (gamma) of the second option to the quadratic term means that the hedge portfolio includes a lot of the second option. As a consequence, it becomes overexposed to the linear term, which is compensated for by a short position in the underlying asset.

22.3.2 Delta-Vega Hedging*

The volatility of financial markets fluctuates over time. To account for this, a first-order Taylor approximation of the call option price in terms of *both* the underlying and volatility is

$$dL \approx \Delta dS + \frac{\partial L}{\partial \sigma} d\sigma, \quad (22.9)$$

where $\partial L / \partial \sigma$ is the “vega” of the option portfolio (presented in an appendix). Notice that the Black-Scholes model is inconsistent with time-variation in volatility—so it can only be used as an approximation.

Consider hedging by holding the following portfolio

$$V = vS + wL^* - L, \quad (22.10)$$

where L^* is the price of some other option (or asset).

We get $dV \approx 0$ by setting

$$w = \frac{\partial L}{\partial \sigma} / \frac{\partial L^*}{\partial \sigma}, \text{ and} \quad (22.11)$$

$$v = \Delta - w\Delta^*, \quad (22.12)$$

where Δ^* and $\partial L^* / \partial \sigma$ are the delta and vega of L^* . For instance, if the L^* asset is directly linked to VIX, then $\Delta^* = 0$ and $\partial L^* / \partial \sigma = 1$.

Proof (of (22.11)–(22.12)) A first-order Taylor approximation like (22.9) of the value of portfolio V gives

$$\begin{aligned} dV &= vdS + w(\Delta^* dS + \frac{\partial L^*}{\partial \sigma} d\sigma) - (\Delta dS + \frac{\partial L}{\partial \sigma} d\sigma) \\ &= (v + w\Delta^* - \Delta)dS + (w \frac{\partial L^*}{\partial \sigma} - \frac{\partial L}{\partial \sigma})d\sigma. \end{aligned}$$

Using the values (w, v) in (22.11)–(22.12) makes this zero. \square

22.4 Appendix – Hedging in the Binomial Model*

The binomial model can be used to calculate the derivatives used in the hedging above. If the binomial model is accurate, it should provide an *exact hedge* rather than an approximation, as in (22.3).

To see that, recall that in any node $(ij$, where i is time step i and j indicates different values of the underlying asset) of the binomial model, we can replicate the derivative by

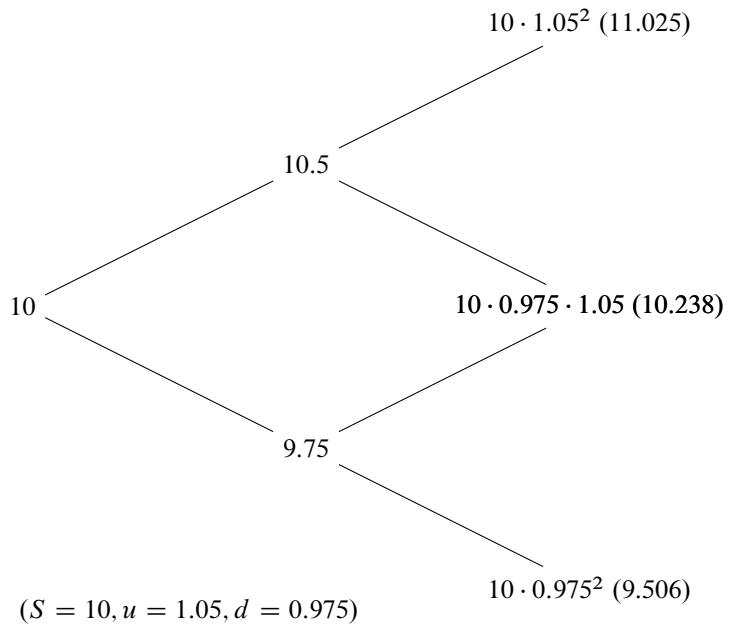


Figure 22.6: Numerical example of a binomial tree for underlying asset ($n = 2$)

the portfolio

$\Delta_{ij} S_{ij}$ in the underlying asset, and

$C_{ij} - \Delta_{ij} S_{ij}$ on a money market account, where

$$\Delta_{ij} = \frac{C_{u,ij} - C_{d,ij}}{S_{ij} (u - d)}. \quad (22.13)$$

where $(\Delta_{ij}, S_{ij}, C_{ij})$ are the values in the *current node* (time step i ,) and $(C_{u,ij}, C_{d,ij})$ are the values of the derivative in the *next time step* (depending on whether the underlying moves up to $S_{ij}u$ or down to $S_{ij}d$). The notation is a bit unconventional, but it is crucial to anchor it at the current price S_{ij} . See also Cox, Ross, and Rubinstein (1979).

Notice that Δ_{ij} is just the number of underlying assets that is needed to replicate the derivative. However, the right hand side of (22.13) shows that it actually is a finite difference ratio—essentially measuring how the derivative price reacts to changes in the underlying, that is, the analogue to $\partial C / \partial S$. Also notice that the amount on the money market account is the same as in (22.4).

See Figure 22.8 for an example of how the hedge portfolio is structured at each node. It is straightforward to show that the value of this portfolio (in the next time step) is the same as the value of the call option in Figure 22.7.

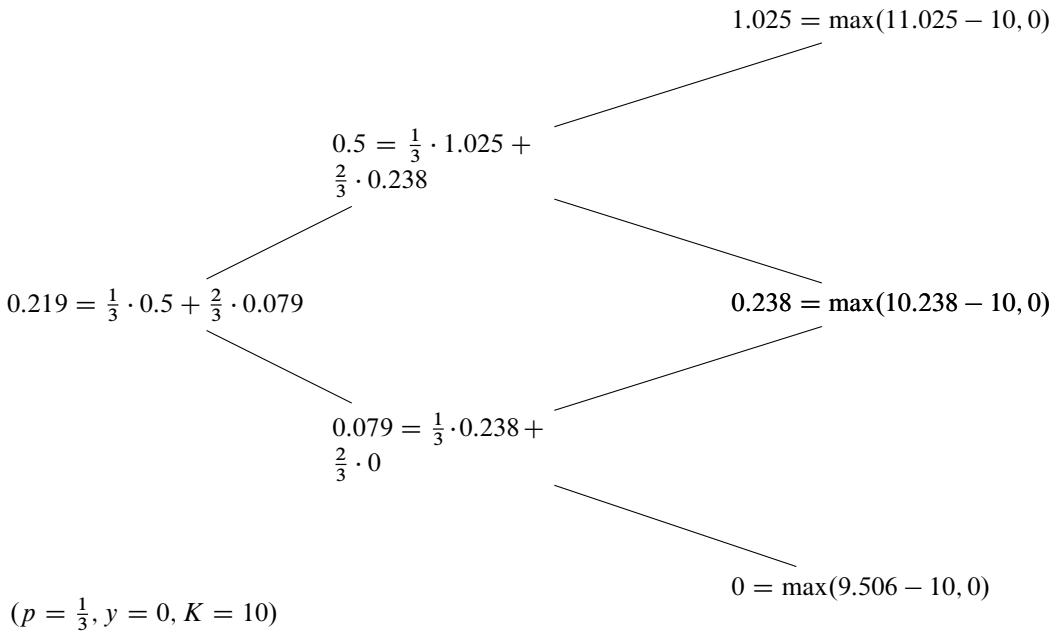


Figure 22.7: Numerical example of binomial tree for European call option ($n = 2$), zero interest rate. The underlying is described in Figure 22.6.

Example 22.12 (*Replicating portfolio in the binomial model*). In the initial node in Figure 22.8, we buy 0.561 underlying assets ($\Delta = 0.561$) and borrow 5.392 on the money market. If the underlying then moves up to 10.5, then this portfolio is worth $0.561 \times 10.5 - 5.392$ (since the interest rate is zero), that is, 0.5. This is the same as the value of the call option in 22.7.

Theoretically, the portfolio (22.13) should provide a perfect replication of the derivative irrespective of whether the underlying asset moves up or down. However, if the binomial model is just an approximation (the most likely case), then the hedge will also be approximate.

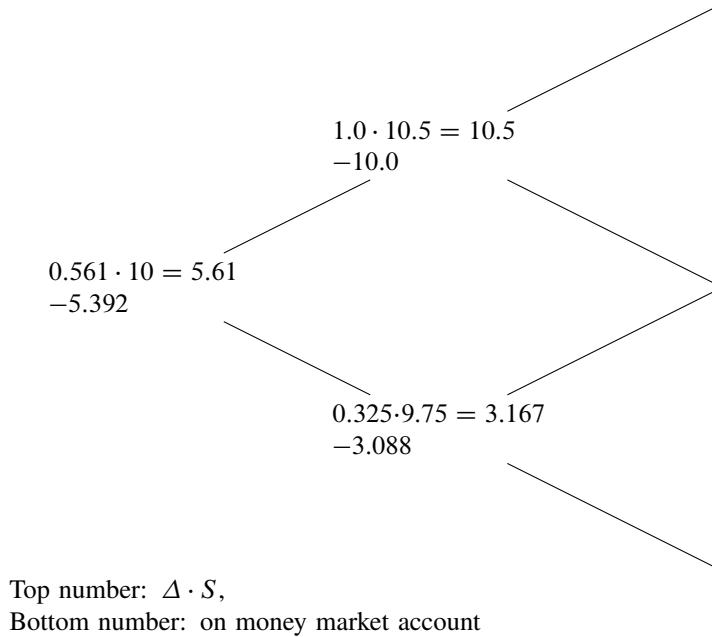


Figure 22.8: Numerical example of how to replicate a European call option ($n = 2$) in a binomial tree, zero interest rate. The underlying and call option are described in Figures 22.6 –22.7.

22.5 Appendix – More Greeks*

Remark 22.13 (*The “Greeks”*) In addition to the results in Remark 22.6, we have

$$\begin{aligned}\Gamma &= \frac{\partial^2 C}{\partial S^2} = \frac{e^{-\delta m} \phi(d_1)}{S \sigma \sqrt{m}} \\ \theta &= \frac{\partial C}{\partial t} = -\frac{\partial C}{\partial m} = \delta S e^{-\delta m} \Phi(d_1) - y K e^{-y m} \Phi(d_2) - \frac{1}{2\sqrt{m}} e^{-\delta m} S \phi(d_1) \sigma \\ (\text{vega}) &= \frac{\partial C}{\partial \sigma} = S e^{-\delta m} \phi(d_1) \sqrt{m} \\ \rho &= \frac{\partial C}{\partial y} = m K e^{-y m} \Phi(d_2).\end{aligned}$$

See Figures 22.9–22.10.

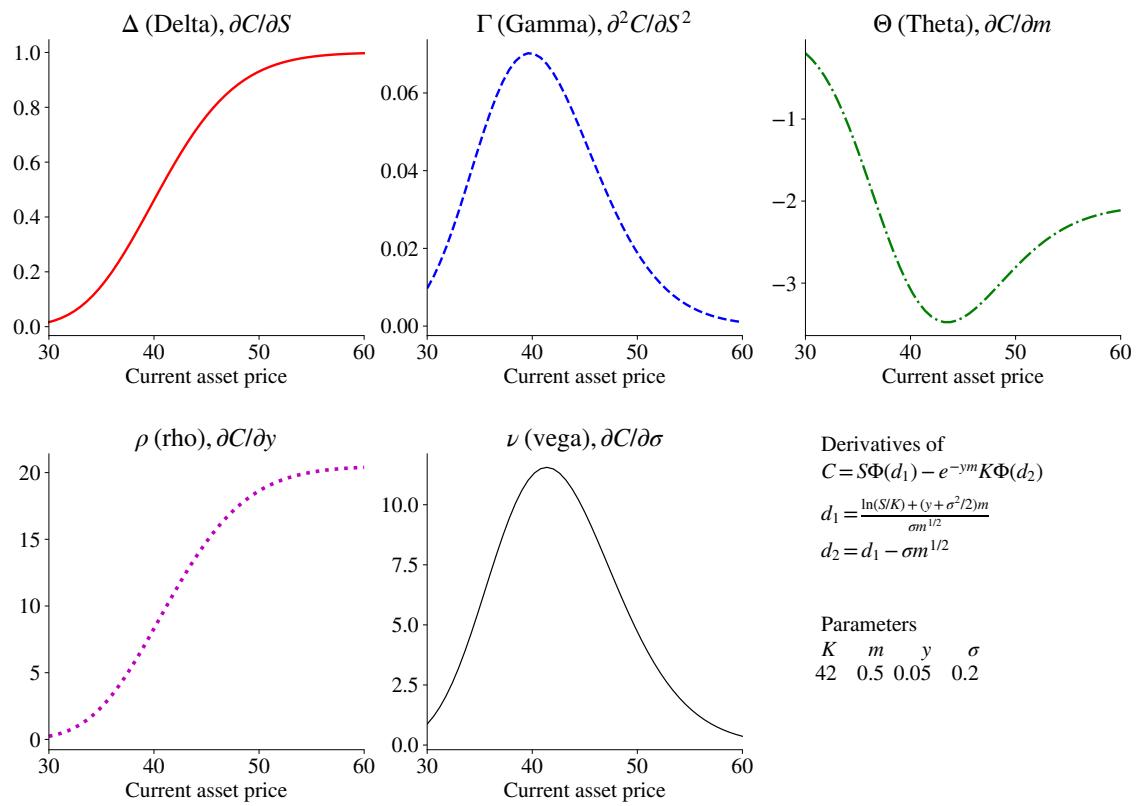


Figure 22.9: The Greeks in the Black-Scholes model as a function of the asset price

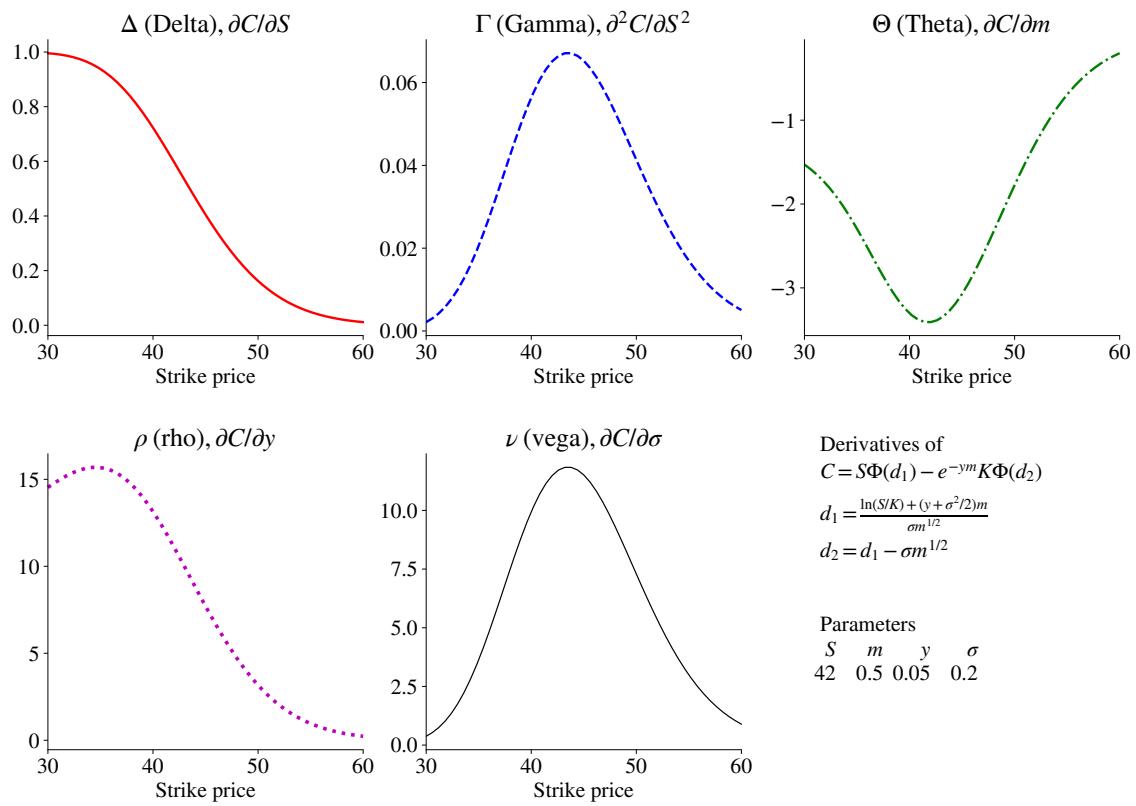


Figure 22.10: The Greeks in the Black-Scholes model as a function of the strike price

Bibliography

- Alexander, C., 2008, *Market Risk Analysis: Value at Risk Models*, Wiley.
- Back, K. E., 2010, *Asset Pricing and Portfolio Choice Theory*, Oxford University Press, Oxford.
- Barber, B., R. Lehavy, M. McNichols, and B. Trueman, 2001, “Can investors profit from the prophets? Security analyst recommendations and stock returns,” *Journal of Finance*, 56, 531–563.
- Bawa, V. S., and E. B. Lindenberg, 1977, “Capital market equilibrium in a mean-lower partial moment framework,” *Journal of Financial Economics*, 5, 189–200.
- Black, F., 1976, “The Pricing of Commodity Contracts,” *Journal of Financial Economics*, 3, 167–179.
- Black, F., and M. Scholes, 1973, “The pricing of options and corporate liabilities,” *Journal of Political Economy*, 81, 637–659.
- Blume, M. E., 1971, “On the Assessment of Risk,” *Journal of Finance*, 26, 1–10.
- Blume, M. E., 1975, “Betas and Their Regression Tendencies,” *Journal of Finance*, 30, 785–795.
- Bondt, W. F. M. D., 1991, “What do economists know about the stock market?,” *Journal of Portfolio Management*, 17, 84–91.
- Bondt, W. F. M. D., and R. H. Thaler, 1990, “Do security analysts overreact?,” *American Economic Review*, 80, 52–57.
- Boni, L., and K. L. Womack, 2006, “Analysts, industries, and price momentum,” *Journal of Financial and Quantitative Analysis*, 41, 85–109.

- Brandimarte, P., 2006, *Numerical Methods in Finance and Economics*, Wiley, Hoboken, NJ.
- Brock, W., J. Lakonishok, and B. LeBaron, 1992, “Simple technical trading rules and the stochastic properties of stock returns,” *Journal of Finance*, 47, 1731–1764.
- 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 L. M. Viceira, 1999, “Consumption and portfolio decisions when expected returns are time varying,” *Quarterly Journal of Economics*, 114, 433–495.
- Campbell, J. Y., and L. M. Viceira, 2002, *Strategic asset allocation: portfolio choice of long-term investors*, Oxford University Press.
- Chance, D. M., and M. L. Hemler, 2001, “The performance of professional market timers: daily evidence from executed strategies,” *Journal of Financial Economics*, 62, 377–411.
- Chen, N.-F., R. Roll, and S. A. Ross, 1986, “Economic forces and the stock market,” *Journal of Business*, 59, 383–403.
- Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.
- Cox, J. C., S. A. Ross, and M. Rubinstein, 1979, “Option pricing: a simplified approach,” *Journal of Financial Economics*, 7, 229–263.
- Damodaran, A., 2012, *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*, John Wiley & Sons, Hoboken, NJ, 3rd edn.
- Danthine, J.-P., and J. B. Donaldson, 2005, *Intermediate financial theory*, Elsevier Academic Press, 2nd edn.
- Diebold, F. X., and R. S. Mariano, 1995, “Comparing predictive accuracy,” *Journal of Business and Economic Statistics*, 13, 253–265.
- Elton, E. J., M. J. Gruber, S. J. Brown, and W. N. Goetzmann, 2014, *Modern portfolio theory and investment analysis*, John Wiley and Sons, 9th edn.
- Fabozzi, F. J., 2004, *Bond markets, analysis, and strategies*, Pearson Prentice Hall, 5th edn.

Fabozzi, F. J., E. H. Neave, and G. Zhou, 2012, *Financial Economics*, John Wiley & Sons, Hoboken, NJ, 2nd edn.

Fama, E. F., and K. R. French, 1993, “Common risk factors in the returns on stocks and bonds,” *Journal of Financial Economics*, 33, 3–56.

Fama, E. F., and K. R. French, 1996, “Multifactor explanations of asset pricing anomalies,” *Journal of Finance*, 51, 55–84.

Fama, E. F., and K. R. French, 2015, “A five-factor asset pricing model,” *Journal of Financial Economics*, 116, 1–22.

Forbes, W., 2009, *Behavioural finance*, Wiley.

Garman, M. B., and S. W. Kohlhagen, 1983, “Foreign currency option values,” *Journal of International Money and Finance*, 2, 231–237.

Goyal, A., and I. Welch, 2008, “A comprehensive look at the empirical performance of equity premium prediction,” *Review of Financial Studies* 2008, 21, 1455–1508.

Greene, W. H., 2018, *Econometric analysis*, Pearson Education Ltd, 8th edn.

Grinblatt, M., and S. Titman, 1993, “A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques,” *Journal of Financial and Quantitative Analysis*, 28, 419–444.

He, J., and L. Ng, 1994, “Economic forces and the stock market,” *Journal of Business*, 4, 599–609.

Huang, C.-F., and R. H. Litzenberger, 1988, *Foundations for financial economics*, Elsevier Science Publishing, New York.

Hull, J. C., 2022, *Options, futures, and other derivatives*, Pearson, 11th edn.

Ibbotson, R. G., and P. D. Kaplan, 2000, “Does Asset Allocation Policy Explain 40, 90 or 100 Percent of Performance?,” *Financial Analysts Journal*, 65, 26–33.

Ingersoll, J. E., 1987, *Theory of financial decision making*, Rowman and Littlefield.

JP Morgan, 1996, “RiskMetrics Technical Document,” Discussion paper, JP Morgan, New York, NY.

- Kochenderfer, M. J., and T. A. Wheeler, 2019, *Algorithms for Optimization*, The MIT Press.
- Ledoit, O., and M. Wolf, 2003, “Improved estimation of the covariance matrix with an application to portfolio selection,” *Journal of Empirical Finance*, 10, 603–621.
- Ledoit, O., and M. Wolf, 2004, “Honey, I shrunk the sample covariance matrix,” *Journal of Portfolio Management*, 30, 110–119.
- Lintner, J., 1965, “The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets,” *Review of Economics and Statistics*, 47, 13–37.
- Lo, A. W., H. Mamaysky, and J. Wang, 2000, “Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation,” *Journal of Finance*, 55, 1705–1765.
- Makridakis, S., S. C. Wheelwright, and R. J. Hyndman, 1998, *Forecasting: methods and applications*, Wiley, New York, 3rd edn.
- Markowitz, H., 1952, “Portfolio Selection,” *The Journal of Finance*, 7, 77–91.
- Mayers, D., 1972, “Nonmarketable Assets, Market Segmentation, and the Level of Asset Prices,” *Journal of Financial and Quantitative Analysis*, 7, 1–12.
- McCulloch, J., 1975, “The tax-adjusted yield curve,” *Journal of Finance*, 30, 811–830.
- McDonald, R. L., 2014, *Derivatives markets*, Pearson, 3rd edn.
- McNeil, A. J., R. Frey, and P. Embrechts, 2005, *Quantitative risk management*, Princeton University Press.
- Merton, R. C., 1973a, “An intertemporal capital asset pricing model,” *Econometrica*, 41, 867–887.
- Merton, R. C., 1973b, “Rational theory of option pricing,” *Bell Journal of Economics and Management Science*, 4, 141–183.
- Mossin, J., 1966, “Equilibrium in a capital asset market,” *Econometrica*, 34, 768–783.
- MSCI Inc., 2024, *MSCI Global Equity Factor Model*.

- Nantell, T. J., and B. Price, 1979, “An analytical comparison of variance and semivariance capital market theories,” *Journal of Financial and Quantitative Analysis*, 14, 221–242.
- Nelson, C., and A. Siegel, 1987, “Parsimonious modeling of yield curves,” *Journal of Business*, 60, 473–489.
- Owen, J., and R. Rabinovitch, 1983, “On the Class of Elliptically Contoured Distributions with Mean–Variance Portfolio Evaluation,” *Journal of Finance*, 38, 745–752.
- Pennacchi, G., 2008, *Theory of Asset Pricing*, Pearson Education.
- Rendleman, R. J., and B. J. Bartter, 1979, “Two-State Option Pricing,” *The Journal of Finance*, 34, 1093–1110.
- Rosenberg, B., and J. Guy, 1976, “Prediction of Beta from Accounting Data,” *Financial Analysts Journal*, 32, 60–72.
- Sercu, P., 2009, *International Finance*, Princeton University Press.
- Sharpe, W. F., 1964, “Capital asset prices: a theory of market equilibrium under conditions of risk,” *Journal of Finance*, 19, 425–442.
- Sharpe, W. F., 1992, “Asset allocation: management style and performance measurement,” *Journal of Portfolio Management*, 39, 119–138.
- Shefrin, H., 2005, *A behavioral approach to asset pricing*, Elsevier Academic Press, Burlington, MA.
- Söderlind, P., 2010, “Predicting stock price movements: regressions versus economists,” *Applied Economics Letters*, 17, 869–874.
- Söderlind, P., and L. E. O. Svensson, 1997, “New techniques to extract market expectations from financial instruments,” *Journal of Monetary Economics*, 40, 383–429.
- Stefek, D., 2002, “The Barra integrated model,” *Barra Research Insight*.
- Svensson, L., 1995, “Estimating forward interest rates with the extended Nelson&Siegel method,” *Quarterly Review, Sveriges Riksbank*, 1995:3, 13–26.
- 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.