Asset Pricing

John H. Cochrane

Acknowledgments

This book owes an obvious intellectual debt to Lars Hansen and Gene Fama. I thank Dan Nelson, George Constantinides, Mike Stutzer, Owen Lamont, Anthony Lynch, Andrea Eisfeldt, Michael Roberts and several generations of Ph.D. students at the University of Chicago for comments on earlier drafts. I thank the NSF and the Graduate School of Business for research support.

Until publication, ever-improving drafts of this book are available at

http://www-gsb.uchicago.edu/fac/john.cochrane/research/papers/finbook.pdf

John H. Cochrane
Graduate School of Business
University of Chicago
1101 E. 58th St.
Chicago IL 60637
773 702 3059
john.cochrane@gsb.uchicago.edu
January 25, 1999

Contents

Acknowledgments 3

1 Preface **10**

Part I. Asset pricing theory **14**

2 Consumption-based model and overview **15**

2.1 Basic pricing equation 15

2.2 Marginal rate of substitution/stochastic discount factor 17

2.3 Prices, payoffs and notation 18

2.4 Intuition, implications, and classic issues in finance 20

3 Discount factors in continuous time **29**

3.1 Assumptions and applicability 33

3.2 Consumption-based model in practice 35

3.3 Alternative asset pricing models: Overview 36

4 The discount factor **39**

4.1 Contingent claims 39

4.2 Investors again 42

4.3 Risk sharing 44

4.4 State diagram and price function 44

4.5 Law of one price and existence of a discount factor 47

4.6 No-Arbitrage and positive discount factors 53

4.7 Existence in continuous time 57

5 Mean-variance frontier and beta representations **59**

5.1 Expected return - Beta representations 59

5.2 Mean-variance frontiers 61

5.3 Relation between $p = E(mx)$, beta, and mean-variance frontiers 74

5.4 Testing for priced factors: lambdas or b's? 87

6 Implications of existence and equivalence theorems **89**

6.1 Discount factors vs. mean, variance and beta. 94

7 Conditioning information **97**

7.1 Scaled payoffs 98

7.2 Sufficiency of adding scaled returns 100

7.3 Conditional and unconditional models 101

7.4 Scaled factors: a partial solution 109

7.5 Summary 111

8 Factor pricing models **112**

8.1 Capital Asset Pricing Model (CAPM) 114

8.2 Intertemporal Capital Asset Pricing Model (ICAPM) 124

8.3 Comments on the CAPM and ICAPM 126

8.4 Arbitrage Pricing Theory (APT) 129

8.5 APT vs. ICAPM 139

Part II. Estimating and evaluating asset pricing models **141**

9 GMM estimation and testing of asset pricing models **142**

9.1 GMM in explicit discount factor models. 142

9.2 Interpreting GMM 145

9.3 Estimating the spectral density matrix 151

10 General formulas, other uses of GMM **157**

10.1 General GMM formulas 157

10.2 Standard errors of anything by delta method 159

10.3 Using GMM for regressions 160

10.4 Problems 162

11 GMM variations **163**

11.1 Horse Races 163

11.2 Prespecified weighting matrices 166

11.3 Testing moments 172

11.4 Applying GMM to linear factor models 173

12 Regression-based tests 177

12.1 Time-series regressions 177

12.2 Cross-sectional regressions 182

12.3 Fama-MacBeth Procedure 189

13 Maximum likelihood 196

13.1 Maximum likelihood 196

13.2 When factors are returns, ML prescribes a time-series regression. 198

13.3 When factors are not excess returns, ML prescribes a cross-sectional regression 201

13.4 Time series vs. cross-section 203

14 ML, GMM and Regression 208

14.1 ML is GMM on the scores 208

14.2 ML approach to a consumption-based model 210

14.3 ML vs. GMM 212

14.4 Problems 220

14.5 References 221

Part III. General equilbrium and derivative pricing 223

15 General Equilibrium 224

15.1 General Equilibrium Recipe 228

15.2 Endowment economies 228

15.3 Production economies 229

15.4 Justification for the procedure 229

15.5 Risk sharing and security design. 229

15.6 Incomplete markets economies. 230

15.7 Outlook 230

16 Continuous time and derivative pricing 231

16.1 Diffusion models 231

16.2 Ito’s lemma 234
16.3 Densities 236
16.4 Tricks 237
16.5 Tricks 237
16.6 Black Scholes with discount factors 237
16.7 Arbitrage bounds using discount factors 242

17 “Good-deal” pricing 243

18 Term structure of interest rates 244
18.1 Overview 244
18.2 Models based on a short rate process 244
18.3 Ho-Lee approach 244
18.4 Use to price options 244

19 Frictions 245

Part IV. Empirical survey 246

20 Return Predictability 247
20.1 Stocks 247
20.2 Term structure 247
20.3 Comparison to continuous-time models 247
20.4 Foreign exchange 247
20.5 Econometric issues 247
20.6 Conditional variance 247
20.7 Conditional Sharpe ratios and portfolio implications 248

21 Present value tests 249

22 Factor pricing models 250
22.1 Capm 250
22.2 Chen Roll Ross model 250
22.3 Investment and macro factors 250

22.4 Book to market 250

22.5 Momentum and more 250

22.6 Digesting the tests 250

23 Consumption based model tests 251

23.1 CRRA utility 251

23.2 Durable goods 251

23.3 Habits 251

23.4 State-nonseparabilities 251

24 Hansen-Jagannathan bounds and equity premium puzzle. 252

24.1 The basic HJ bound and equity premium 253

24.2 Many returns–formulas 254

24.3 Results 264

24.4 Beyond mean and variance. 264

24.5 What do we know about discount factors: a summary 266

24.6 Comments on the Hansen-Jagannathan bound. 266

25 Investment, q and asset pricing 268

Part V. Appendix 269

26 Notation 270

27 Utility functions 274

28 Probability and statistics 275

28.1 Probability 275

28.2 Statistics 279

28.3 Regressions 282

28.4 Partitioned matrix inverse formulas 283

29 Bibliography 285

30 Problems 286

31 Additional algebra **290**

32 Notes: **297**

Chapter 1. Preface

Asset pricing theory tries to understand and predict the prices or values of claims to uncertain payments. Accounting for the time and risk of prospective payments makes this theory interesting and challenging. A low price implies a high rate of return, so one can also think of the theory as explaining why some assets pay higher average returns than others.

If there were no risk, asset pricing would be easy, and would simply consist of discounting future cash flows using present value formulas. *Uncertainty,* or *corrections for risk* make asset pricing interesting and challenging. The large size of risk corrections in real world asset markets make asset pricing theory challenging and relevant.

Asset pricing theory shares the positive vs. normative tension present in the rest of economics. Does it describe the way the world *does* work or the way the world *should* work? We observe the prices or returns of many assets. We can use the theory positively, to try to understand why prices or returns are what they are. If the world does not obey a model's predictions, we can decide that the model needs improvement. However, we can also decide that the *world* is wrong, that some assets are "mis-priced" and present trading opportunities for the shrewd investor. This latter use of asset pricing theory accounts for much of its popularity and practical application. Also, and perhaps most importantly, the prices of many assets or claims to uncertain cash flows are *not* observed, such as potential public or private investment projects, new financial securities, buyout prospects, and complex derivatives. We can apply the theory to establish what the prices of these claims should be as well; the answers are important guides to public and private decisions.

Asset pricing theory all stems from one simple equation, derived in the first page of Chapter 1 of this book: price equals expected discounted payoff. The rest is elaboration, special cases, and a closet full of tricks that make the central equation useful for one or another application. There are two polar approaches to this elaboration. I will call them *absolute pricing* and *relative pricing*.

In *absolute pricing,* we price each asset by reference to its exposure to fundamental sources of macroeconomic risk. The consumption-based and general equilibrium models described below are the purest examples of this approach. The absolute approach is most common in academic settings, in which we use asset pricing theory positively to give an economic explanation for why prices are what they are or in order to predict how prices might change if policy or economic structure changed. In *relative pricing,* we ask a less ambitious question. We ask what we can learn about an asset's value *given* the prices of some other assets. We do not ask where the price of the other set of assets came from, and we use as little information about fundamental risk factors as possible. Black-Scholes option pricing is the classic example of this approach. While limited in scope, this approach offers precision in many applications.

Asset pricing problems are solved by judiciously choosing how much absolute and how much relative pricing one will do, depending on the assets in question and the purpose of the calculation. Almost no problems are solved by the pure extremes. For example, the CAPM and its successor factor models price assets "relative" to the market or factors, without answering what determines the market or factor risk premia and betas. The latter are treated as free parameters. On the other end of the spectrum, most practical financial engineering questions involve assumptions beyond pure lack of arbitrage, about equilibrium "market prices of risk".

The central and unfinished task of absolute asset pricing is to understand and measure the sources of aggregate or macroeconomic risk that drive asset prices. Of course, this is also the central question of macroeconomics, and this is a particularly exciting time for researchers who want to answer these fundamental questions in macroeconomics and finance. A lot of empirical work has documented tantalizing stylized facts and links between macroeconomics and finance. For example, expected returns vary across time and across assets in ways that are linked to macroeconomic variables, or variables that also forecast macroeconomic events; a wide class of models suggests that a "recession" or "financial distress" factor lies behind many asset prices. Yet theory lags behind; we do not yet have a well-described model that explains these interesting correlations.

This book advocates a discount factor / generalized method of moments view of asset pricing theory and associated empirical procedures. I summarize asset pricing by two equations:

$$p_t = E(m_{t+1}x_{t+1})$$

$$m_{t+1} = f(\text{data, parameters}).$$

where p_t = asset price, x_{t+1} = asset payoff, m_{t+1} = stochastic discount factor.

The major advantage of the discount factor / moment condition approach are its simplicity and universality. Where once there were three apparently different theories for stocks, bonds, and options, now we see each as just special cases of the same theory. The common language also allows us to use insights from each field of application in other fields.

This approach also allows us to conveniently separate the step of specifying economic assumptions of the model (second equation) from the step of deciding which kind of empirical representation to pursue or understand. For a given model – choice of $f(\cdot)$ – we will see how the first equation can lead to predictions stated in terms of returns, price-dividend ratios, expected return-beta representations, moment conditions, continuous vs. discrete time implications and so forth. The ability to translate between such representations is also very helpful in digesting the results of empirical work, which uses a number of apparently distinct but fundamentally connected representations.

It also turns out to often be much simpler to think in terms of discount factors rather than portfolios. For example, it is easier to insist that there exists a positive discount factor than to check that every possible portfolio that dominates every other portfolio has a larger price, and the long arguments over the APT stated in terms of portfolios are easy to digest when stated in terms of discount factors.

For these reasons, the discount factor language is common in academic research and high-tech practice. It is not yet common in textbooks, and that is the niche that this book tries to fill.

I also diverge from the usual order of presentation. Most books are structured following the history of thought: portfolio theory, mean-variance frontiers, spanning theorems, CAPM, ICAPM, APT, and finally consumption-based model. Contingent claims are an esoteric extension of option-pricing theory. I go the other way around: contingent claims and the consumption-based model are the basic and simplest models around; the others are specializations. Just because they were discovered in the opposite order is no reason to present them that way.

I also try to unify the treatment of empirical methods. A wide variety of methods are popular, including time-series and cross-sectional regressions, and methods based on generalized method of moments (GMM) and maximum likelyhood. However, in the end all of these apparently different approaches do the same thing: they pick free parameters of the model to make it fit best, which usually means to minimize pricing errors; and they evaluate the model by a quadratic form in pricing errors.

As with the theory, I do not attempt an encyclopedic compilation of empirical procedures. As with the theory, the literature on econometric methods contains lots of methods and special cases (likelyhood ratio ways of doing common Wald tests; cases with and without riskfree assets and when factors do and don't span the mean variance frontier, etc.) that are seldom used used in practice. I try to focus on the basic ideas and on methods that are actually used in practice.

The accent in this book is on understanding statements of theory, and working with that theory to applications, rather than rigorous or general proofs. Also, I skip very lightly over many parts of asset pricing theory that have faded from current applications, although they occupied large amounts of the attention in the past. Some examples are portfolio separation theorems, properties of various distributions, or asymptotic APT. Since my focus is on the determinants of asset prices, I do not spend much time on portfolio theory either. While that theory is still interesting and useful theory for finding portfolios, it is no longer a cornerstone of pricing. Rather than use portfolio theory to find a demand curve for assets, which intersected with a supply curve gives prices, we now go to prices directly. One can then find optimal portfolios, but it is a side issue.

Again, my organizing principle is that everything can be traced back to specializations of the basic pricing equation $p = E(mx)$. Therefore, after reading the first chapter, one can pretty much skip around and read topics in as much depth or order as one likes. Each major subject always starts back at the same pricing equation.

The target audience for this book is economics and finance Ph.D. students, advanced MBA students or professionals with similar background. I hope the book will also be useful to fellow researchers and finance professionals, by clarifying, relating and simplifying the set of tools we have all learned in a hodgepodge manner. I presume some exposure to undergraduate economics and statistics. A reader should have seen a utility function, a random variable, a standard error, and a time series before, should have some basic calculus and should have solved a maximum problem by setting derivatives to zero. The hurdles in asset pricing are really conceptual rather than mathematical.

PART I
Asset pricing theory

Chapter 2. Consumption-based model and overview

2.1 Basic pricing equation

We derive the basic consumption-based model,

$$p_t = E_t \left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1} \right].$$

Our basic objective is to figure out the value of any stream of uncertain cash flows. We start with an apparently simple case, which turns out to capture very general situations.

Let us find the value at time t of a *payoff* x_{t+1}. For example, if one buys a stock today, the payoff next period is the stock price plus dividend, $x_{t+1} = p_{t+1} + d_{t+1}$. x_{t+1} is a random variable: an investor does not know exactly how much he will get from his investment, but he can assess the probability of various possible outcomes. Don't confuse the *payoff* x_{t+1} with the *profit* or *return*; x_{t+1} is the value of the investment at time $t+1$, without subtracting or dividing by the cost of the investment.

We find the value of this payoff by asking what it is worth to a typical investor. To do this, we need a convenient mathematical formalism to capture what an investor wants. We model investors by a *utility function* defined over current and future values of consumption,

$$U(c_t, c_{t+1}) = u(c_t) + \beta E_t \left[u(c_{t+1}) \right].$$

We will often use a convenient power utility form,

$$u(c_t) = \frac{1}{1-\gamma} c_t^{1-\gamma} \ \ \gamma \neq 1; \quad u(c_t) = \ln c_t \ \ \gamma = 1.$$

This formalism captures investors' impatience and their aversion to risk. Therefore, we will be able to quantitatively correct for the risk and delay of cash flows. The utility function captures the fundamental desire for more *consumption*, rather than posit a desire for intermediate objectives such as means and variance of portfolio returns. Consumption c_{t+1} is also random; the investor does not know his wealth tomorrow, and hence how much he will decide to consume. The period utility function $u(\cdot)$ is increasing, reflecting a desire for more consumption, and concave, reflecting the declining marginal value of additional consumption. The last bite is never as satisfying as the first. More importantly, the curvature of the utility function also generates aversion to risk and to intertemporal substitution: The consumer prefers a consumption stream that is steady over time and across states of nature. Discounting the future by β captures impatience, and β is called the *subjective discount factor*.

Now, assume that the investor can freely buy or sell as much of the payoff x_{t+1} as he wishes, at a price p_t. How much will he buy or sell? To find the answer, denote by e the original consumption level (if the investor bought none of the asset), and denote by ξ the amount of the asset he chooses to buy. Then, his problem is,

$$\max_{\{\xi\}} u(c_t) + E_t \beta u(c_{t+1}) \quad s.t.$$

$$\begin{aligned} c_t &= e_t - p_t \xi \\ c_{t+1} &= e_{t+1} + x_{t+1} \xi \end{aligned}$$

Substituting the constraints into the objective, and setting the derivative with respect to ξ equal to zero, we obtain the first-order condition for an optimal consumption and portfolio choice,

$$p_t u'(c_t) = E_t \left[\beta u'(c_{t+1}) x_{t+1} \right] \tag{1}$$

or,

$$p_t = E_t \left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1} \right]. \tag{2}$$

The consumer buys more or less of the asset until this first order condition holds.

Equation (1) expresses the standard marginal condition for an optimum: $pu'(c_t)$ is the loss in utility if the consumer buys another unit of the asset; $E_t \left[\beta u'(c_{t+1}) x_{t+1} \right]$ is the increase in (discounted, expected) utility obtained from the payoff corresponding to an additional unit of the asset at $t+1$. The consumer continues to buy or sell the asset until the marginal loss equals the marginal gain.

Equation (2) is *the* central asset-pricing formula. Given the payoff x_{t+1} and given the investor's consumption choice c_t, c_{t+1}, it tells you what market price p_t to expect. Its economic content is simply the first order conditions for optimal consumption and portfolio formation. Most of the theory of asset pricing just consists of specializations and manipulations of this formula.

Notice that we have stopped short of a complete solution to the model, i.e. an expression with exogenous items on the right hand side. We relate one endogenous variable, price, to two other endogenous variables, consumption and payoffs. One can continue to solve this model and derive the optimal consumption choice c_t, c_{t+1} in terms of the givens of the model. In this case, those givens are initial wealth, the income sequence e_t, e_{t+1} and a specification of the full set of assets that the consumer may buy and sell. We will in fact study such fuller solutions below. However, for many purposes one can stop short of specifying (possibly wrongly) all this extra structure, and obtain very useful predictions about asset prices from (2), even though consumption is an endogenous variable.

2.2 Marginal rate of substitution/stochastic discount factor

We break up the basic consumption-based pricing equation into

$$p = E(mx)$$

$$m = \beta \frac{u'(c_{t+1})}{u'(c_t)}$$

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

A convenient way to break up the basic pricing equation (2) is to define the *stochastic discount factor* m_{t+1}

$$m_{t+1} \equiv \beta \frac{u'(c_{t+1})}{u'(c_t)} \tag{3}$$

Then, the basic pricing formula (2) can simply be expressed as

$$p_t = E_t(m_{t+1} x_{t+1}). \tag{4}$$

When it isn't necessary to be explicit about time subscripts, I'll suppress them and just write $p = E(mx)$. The price always comes at t, the payoff at $t + 1$, and the expectation is conditional on time t information.

The term *stochastic discount factor* refers to the way m generalizes standard discount factor ideas. If there is no uncertainty, we can express prices via the standard present value formula

$$p_t = \frac{1}{R^f} x_{t+1} \tag{5}$$

where R^f is the risk-free rate. $1/R^f$ is the *discount factor.* Since gross interest rates are typically greater than one, the payoff x_{t+1} sells "at a discount." Riskier assets have lower prices than equivalent risk-free assets, so they are often valued by using risk-adjusted discount factors,

$$p_t^i = \frac{1}{R^i} E_t(x_{t+1}^i).$$

Here, I have added the i superscript to emphasize that each risky asset i must be discounted by an asset-specific risk-adjusted discount factor $1/R^i$.

In this context, equation (4) is obviously a generalization, and it says something deep: one can incorporate all risk-corrections by defining a *single* stochastic discount factor – the same one for each asset – and putting it inside the expectation. m_{t+1} is *stochastic* or *random* because it is not known with certainty at time t. As we will see, the correlation between the random components of m and x^i generate asset-specific risk corrections.

m_{t+1} is also often called the *marginal rate of substitution* after (3). In that equation, m_{t+1} is the rate at which the investor is willing to substitute consumption at time $t+1$ for consumption at time t. $m_{t+1}$is sometimes also called the *pricing kernel.* If you know what a kernel is and express the expectation as an integral, you can see where the name comes from. It is sometimes called a *change of measure* or a *state-price density* for reasons that we will see below.

For the moment, introducing the discount factor m and breaking the basic pricing equation (2) into (3) and (4) is just a notational convenience. As we will see, however, it represents a much deeper and more useful separation. For example, notice that $p = E(mx)$ would still be valid if we changed the utility function, but we would have a different function connecting m to data. This turns out to be quite generally true: $p = E(mx)$ is a convenient accounting identity with almost no content. *All* asset pricing models amount to alternative models connecting the stochastic discount factor to data. Therefore, we can conveniently break up our vision of asset pricing into different expressions of $p = E(mx)$ and the effects of different models connecting m to data.

2.3 Prices, payoffs and notation

The *price* p_t gives rights to a *payoff* x_{t+1}. In practice, this notation covers a variety of cases, including the following:

	Price p_t	Payoff x_{t+1}
Stock	p_t	$p_{t+1} + d_{t+1}$
Return	1	R_{t+1}
Price-dividend ratio	$\frac{p_t}{d_t}$	$\left(\frac{p_{t+1}}{d_{t+1}} + 1\right)\frac{d_{t+1}}{d_t}$
Excess returns	0	$R^e_{t+1} = R^a_{t+1} - R^b_{t+1}$
Managed portfolio	z_t	$z_t R_{t+1}$
Moment condition	$E(p_t z_t)$	$x_{t+1} z_t$
One-period bond	p_t	1
Risk free rate	1	R^f
Option	C	$\max(S_T - K, 0)$

The price p_t and payoff x_{t+1} seem like a very restrictive kind of security. In fact, this notation is quite general and allows us easily to accommodate many different asset pricing questions. In particular, we can cover stocks, bonds and options and make clear that there is one theory for all asset pricing.

For stocks, the one period payoff is of course the next price plus dividend, $x_{t+1} = p_{t+1} + d_{t+1}$. We frequently divide the payoff x_{t+1} by the price p_t to obtain a *gross return*

$$R_{t+1} \equiv \frac{x_{t+1}}{p_t}$$

A return is a payoff with price one. If you pay one dollar today, the return is how many dollars (units of consumption) you get tomorrow. Thus, returns obey

$$1 = E(mR)$$

which is by far the most important special case of the basic formula $p = E(mx)$. Confusing payoffs and returns is a common mistake. You "lose money" if the payoff is less than the price, but the payoff is still positive.

I use capital letters to denote *gross* returns R, which have a numerical value like 1.05. I use lowercase letters to denote *net* returns $r = R - 1$ or log (continuously compounded) returns $\ln(R)$, both of which have numerical values like 0.05. One may also quote *percent* returns $100 \times r$. Prices, payoffs, returns etc. may all be real—denominated in consumption goods—or nominal—denominated in dollars.

Returns are often used in empirical work because they are stationary (in the statistical sense, not constant) over time. However, thinking in terms of returns takes us away from the central task of finding asset *prices*. Dividing by dividends and creating a payoff $x_{t+1} = (1 + p_{t+1}/d_{t+1})\, d_{t+1}/d_t$ corresponding to a price p_t/d_t is a way to look at prices but still to examine stationary variables.

Not everything can be reduced to a return, however. If you borrow a dollar at the interest rate R^f and invest it in an asset with return R, you pay no money out-of-pocket today, and get the payoff $R - R^f$. This is a payoff with a *zero* price. Zero price does not imply zero payoff; just an “even bet” that is not worth paying extra to take. It is common to study equity strategies in which one short sells one stock or portfolio and invests the proceeds in another stock or portfolio, generating an excess return. I denote any such difference between returns as an *excess return*, R^e. It is also called a *zero-cost portfolio* or a *self-financing portfolio*.

In fact, much asset pricing focuses on excess returns. Our economic understanding of interest rate variation turns out to have little to do with our understanding of risk premia, so it is convenient to separate the two exercises by looking at interest rates and excess returns separately.

We also want to think about the *managed portfolios*, in which one invests more or less in an asset according to some signal. The “price” of such a strategy is the amount invested at time t, say z_t, and the payoff is $z_t R_{t+1}$. For example a market timing strategy might put a weight in stocks proportional to the price-dividend ratio, investing less when prices are higher. We could represent such a strategy as a payoff using. $z_t = a - b(p_t/d_t)$

When we think about conditioning information below, we will think of objects like z_t as *instruments*. Then we take an unconditional expectation of $p_t z_t = E_t(m_{t+1}x_{t+1})z_t$, yielding $E(p_t z_t) = E(m_{t+1}x_{t+1}z_t)$. We can think of this operation as creating a “security” with payoff $x_{t+1}z_{t+1}$, and “price” $E(p_t z_t)$ represented with unconditional expectations.

A one period bond is of course a claim to a unit payoff. Bonds, options, investment projects are all examples in which it is often more useful to think of prices and payoffs rather than returns.

To accommodate all these cases, we will simply use the notation price p_t and payoff x_{t+1}. These symbols can denote $0, 1,$ or z_t and $R^e_t,\ r_{t+1}$, or $z_t R_{t+1}$ respectively, according to the case. Lots of other definitions of p and x are useful as well.

2.4 Intuition, implications, and classic issues in finance

I use simple manipulations of the basic pricing equation to introduce classic issues in finance: the economics of interest rates, risk adjustments, the mean-variance frontier, the slope of the mean-variance frontier, a beta representation for expected returns, and time-varying expected returns.

Risk-corrections are driven by covariance of payoffs with the stochastic discount factor. Prices are driven down and returns up for assets that make consumption more volatile.

A few simple rearrangements and manipulations of the basic pricing equation $p = E(mx)$ give a lot of intuition and introduce some classic issues in finance, including determinants of the interest rate, risk corrections, idiosyncratic vs. systematic risk, beta pricing models, and mean variance frontiers.

2.4.1 Risk free rate.

The risk free rate is given by

$$R^f = 1/E(m). \tag{6}$$

The risk free rate is known ahead of time, so $p = E(mx)$ becomes $1 = E_t(m_{t+1}R^f_{t+1}) = E_t(m_{t+1})R^f_{t+1}$.

If a risk free security is not traded, we can define $R^f = 1/E(m)$ as the "shadow" risk-free rate. (In some models it is called the "zero-beta" rate.) If one introduced a risk free security with return $R^f = 1/E(m)$, consumers would be just indifferent to buying or selling it.

To think about the economics behind interest rates, consider the consumption-based discount factor model with power utility $u'(c) = c^{-\gamma}$, and assume that consumption growth is lognormally distributed. Then, the riskfree rate equation becomes

$$r^f_t = \delta + \gamma E_t \Delta \ln c_{t+1} - \frac{\gamma^2}{2}\sigma^2_t(\Delta \ln c_{t+1}) \tag{7}$$

where I have defined the log riskfree rate $r^f$and subjective discount rate δ by

$$\begin{aligned} r^f_t &= \ln R^f_t \\ \beta &= e^{-\delta} \end{aligned}$$

and Δ denotes the first difference operator,

$$\Delta \ln c_t = \ln c_t - \ln c_{t-1}.$$

To derive expression (7) for the riskfree rate, start with

$$R^f_t = 1/E_t\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right].$$

Using the fact that normal z means

$$E\left(e^z\right) = e^{E(z)+\frac{1}{2}\sigma^2(z)},$$

We have

$$R^f_t = 1/\left[e^{-\delta}e^{-\gamma E_t \Delta \ln c_{t+1} + \frac{\gamma^2}{2}\sigma^2_t \Delta \ln c_{t+1}}\right].$$

and then take logarithms.

Looking at (7), interest rates are high when impatience δ is high. If everyone wants to consume now, it takes a high interest rate to convince them to save.

Interest rates are high when consumption *growth* is high. In times of high interest rates, it pays investors to consume less now, invest more, and consume more in the future. Thus, high interest rates lower the *level* of consumption while raising its growth rate. The power parameter γ is the inverse of the *elasticity of intertemporal substitution*. For high γ, people are less willing to rearrange consumption over time in response to interest rate incentives. Such consumers are also less willing to rearrange consumption over states of nature; with this utility function γ controls risk aversion as well as intertemporal substitution.

Finally, the σ^2 term captures *precautionary savings*. When consumption is more volatile, people with this utility function are more worried about the low consumption states than they are pleased by the high consumption states. Therefore, people want to save more, driving down interest rates.

2.4.2 Risk corrections.

Using the definition of covariance $cov(m,x) = E(mx) - E(m)E(x)$, we can write equation (2) as

$$p = E(m)E(x) + cov(m,x). \tag{8}$$

Substituting the riskfree rate equation (6), we obtain

$$p = \frac{E(x)}{R^f} + cov(m,x) \tag{9}$$

The first term is the standard discounted present value formula. This is the asset's price in a risk-neutral world – if consumption is constant or if utility is linear. The second term is a *risk adjustment*. An asset whose payoff covaries positively with the discount factor has its price raised and vice-versa.

To understand the risk adjustment, substitute back for m in terms of consumption, to obtain

$$p = \frac{E(x)}{R^f} + \frac{cov\left[\beta u'(c_{t+1}), x_{t+1}\right]}{u'(c_t)} \tag{10}$$

Marginal utility $u'(c)$ declines as c rises. Thus, an asset's price is lowered if its payoff covaries positively with consumption. Conversely, an asset's price is raised if it covaries negatively with consumption.

Why? Investors do not like uncertainty about consumption. If you buy an asset whose payoff covaries positively with consumption, one that pays off well when you are already feeling wealthy and pays off badly when you are already feeling poor, that asset will make your consumption stream more volatile. You will require a low price or a good average return to induce you to buy such an asset. If you buy an asset whose payoff covaries negatively with consumption, it helps to smooth consumption and so is more valuable than its expected payoff might indicate.

Insurance is an extreme example of the latter effect. Insurance pays off exactly when wealth and consumption is low for other reasons–you get a check when your house burns down. For this reason, you are happy to hold insurance, even though you expect to lose money—even though the price of insurance is greater than its expected payoff discounted at the risk free rate.

2.4.3 Risk corrections to expected returns.

We use returns so often that it is worth restating the same intuition in terms of returns. Start with the basic pricing equation for returns,

$$1 = E(mR).$$

Apply the covariance decomposition,

$$1 = E(m)E(R) + cov(m, R) \tag{11}$$

$$E(R) = \frac{1}{E(m)} - \frac{cov(m, R)}{E(m)}$$

$$E(R) = R^f - \frac{cov[u'(c_{t+1}), R_{t+1}]}{E[u'(c_{t+1})]}. \tag{12}$$

All assets have an expected return equal to the risk-free rate, plus a risk adjustment. Assets whose returns covary positively with consumption make consumption more volatile, and so must promise higher expected returns to induce investors to hold them. Conversely, assets that covary negatively with consumption, such as insurance, can offer expected rates of return that are lower than the risk-free rate, or even negative (net) expected returns.

Much of finance focuses on expected returns. We think of expected returns increasing or decreasing to clear markets; we offer intuition that "riskier" securities must offer higher expected returns to get investors to hold them, rather than saying "riskier" securities trade for lower prices so that investors will hold them. Of course, a low price for a given payoff corresponds to a high expected return, so this is no more than a different language for the same phenomenon.

2.4.4 Idiosyncratic risk does not affect prices.

You might think that an asset with a high payoff variance is "risky" and thus should have a large risk correction. However, if the payoff is uncorrelated with the discount factor m, the asset receives *no* risk-correction to its price, and pays an expected return equal to the risk-free rate! In equations, if

$$cov(m,x)=0$$

then

$$p=\frac{E(x)}{R^f}.$$

This prediction holds even if the payoff x is highly volatile and investors are highly risk averse. The reason is simple: if you buy a little bit of such an asset, it has no first-order effect on the variance of your consumption stream.

Another way of saying the same thing is that one gets no compensation or risk adjustment for holding *idiosyncratic* risk. Only *systematic* risk generates a risk correction. To give meaning to these words, we can decompose any payoff x into a part correlated with the discount factor and an idiosyncratic part uncorrelated with the discount factor by running a regression,

$$x=proj(x|m)+\varepsilon.$$

The price of x is the same as the price of its projection on m, and the residual has zero price:

$$\begin{aligned} p(\varepsilon) &= E(m\varepsilon)=0 \\ p(x) &= E(mx)=E[m\ (proj(x|m)+\varepsilon)]=E[m\ proj(x|m)] \end{aligned}$$

(I use projection to mean linear regression,

$$proj(x|m)=\frac{E(mx)}{E(m^2)}m.$$

You can verify that $E(m\varepsilon)=0$ follows from this definition.) The projection of x on m is of course that part of x which is perfectly correlated with m. The *idiosyncratic* component of any payoff is that part uncorrelated with m. Thus only the systematic *part* of a payoff accounts for its price.

2.4.5 Expected return-beta representation.

We can rewrite equation (14) as

$$E(R^i)=R^f+\left(\frac{cov(R^i,m)}{var(m)}\right)\left(-\frac{var(m)}{E(m)}\right)$$

or

$$E(R^i)=R^f+\beta_{i,m}\lambda_m$$

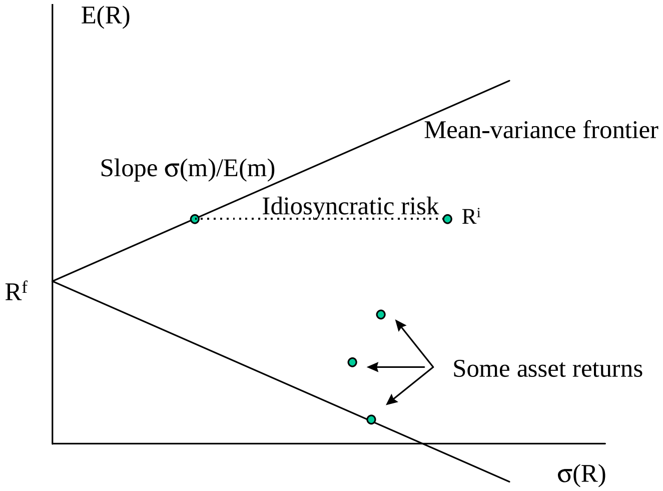


Figure 1. Mean-variance frontier. The mean and standard deviation of all assets priced by a discount factor m must line in the wedge-shaped region

where β_{im} is the regression coefficient of the return R^i on m. This is the first instance of a *beta pricing model*, which we will look at in more detail below. It says that expected returns on assets $i = 1, 2, ...N$ should be proportional to their betas in a regression of returns on the discount factor. Notice that the coefficient λ_m is the same for all assets i,while the $\beta_{i,m}$ varies from asset to asset. The λ_m is often interpreted as the *price of* β *risk* and the β as the quantity of risk in each asset.

Obviously, there is nothing deep about saying that expected returns are proportional to betas rather than to covariances. There is a long historical tradition and some minor convenience in favor of betas. The betas of course refer to the projection of R on m that we studied above, so you see again how only the systematic component of risk matters.

2.4.6 Mean-variance frontier

Asset pricing theory has focused a lot on the means and variances of asset returns. Interestingly, the set of means and variances of returns is limited. All assets priced by the discount factor m must obey

$$\left|E(R^i) - R^f\right| \leq \frac{\sigma(m)}{E(m)}\sigma(R^i). \qquad (13)$$

Means and variances of asset returns therefore must lie in the wedge-shaped region illustrated in Figure 1.

To derive (13) write for a given asset return R^i

$$1 = E(mR^i) = E(m)E(R^i) + \rho_{m,R^i}\sigma(R^i)\sigma(m)$$

and hence

$$E(R^i) = R^f - \rho_{m,R^i}\frac{\sigma(m)}{E(m)}\sigma(R^i). \tag{14}$$

Correlation coefficients can't be greater than one in magnitude, leading to (13).

The boundary of the mean-variance region in which assets can lie is called the *mean-variance frontier*. It answers a naturally interesting question, "how much mean return can you get for a given level of variance?" It also plays a central role in asset pricing, which we'll see below.

All returns on the frontier are perfectly correlated with the discount factor: the frontier is generated by $\left|\rho_{m,R^i}\right| = 1$. Returns on the upper part of the frontier are perfectly negatively correlated with the discount factor and hence positively correlated with consumption. They are "maximally risky" and thus get the highest expected returns. Returns on the lower part of the frontier are perfectly positively correlated with the discount factor and hence perfectly negatively with consumption. They thus provide the best insurance against consumption fluctuations.

All frontier returns are also perfectly correlated with each other, since they are all perfectly correlated with the discount factor. This implies that we can *span* or *synthesize* any frontier return from two such returns. For example if you pick any single frontier return R^m then all frontier returns R^{mv} must be expressible as

$$R^{mv} = R^f + a\left(R^m - R^f\right)$$

for some number a.

Since each point on the mean-variance frontier is perfectly correlated with the discount factor, we must be able to pick constants a, b, d, e such that

$$\begin{aligned} m &= a + bR^{mv} \\ R^{mv} &= d + em. \end{aligned}$$

Thus, *any mean-variance efficient return* (except the riskfree rate) *carries all pricing information.* Given a mean-variance efficient return, we can find a discount factor that prices all assets and vice versa. Given a discount factor, we can construct a single-beta representation, so *Expected returns can be described in a single - beta representation using any mean-variance efficient return* (except the riskfree rate),

$$E(R^i) = R^f + \beta_{i,mv}\left[E(R^{mv}) - R^f\right].$$

The essence of the β pricing model is that, even though the means and standard deviations of returns fill out the space inside the mean-variance frontier, a graph of mean returns versus *betas* should yield a straight line.

We can plot the decomposition in point 4 above of any return into a "priced part" and an residual as shown in Figure 1. The priced part is perfectly correlated with the discount factor, and hence perfectly correlated with any frontier asset. The residual part generates no expected return, and is uncorrelated with the discount factor or any frontier asset. For the latter reason, it is often referred to as the *idiosyncratic* component of risk.

2.4.7 Slope of the mean-standard deviation frontier.

The slope of the mean-standard deviation frontier is naturally interesting. It answers "how much more mean return can I get by shouldering a bit more variance?" Let R^p denote the return of a portfolio on the frontier. From equation (13), the slope of the frontier is

$$\left|\frac{E(R^p) - R^f}{\sigma(R^p)}\right| = \frac{\sigma(m)}{E(m)} = \sigma(m)R^f$$

Thus, the slope of the frontier, also known as the *price of risk* or maximal *Sharpe ratio* is governed by the volatility of the discount factor.

For an economic interpretation, again consider the power utility function, $u'(c) = c^{-\gamma}$,

$$\left|\frac{E(R^i) - R^f}{\sigma(R^i)}\right| = \frac{\sigma\left[(c_{t+1}/c_t)^{-\gamma}\right]}{E\left[(c_{t+1}/c_t)^{-\gamma}\right]}. \tag{15}$$

The standard deviation is large if consumption is volatile or if γ is large. We can state this approximation again using the lognormal assumption. If consumption growth is lognormal,

$$\left|\frac{E(R^i) - R^f}{\sigma(R^i)}\right| = \sqrt{e^{\gamma^2\sigma^2(\Delta \ln c_{t+1})} - 1} \approx \gamma\sigma(\Delta \ln c).$$

Reading the equation, *the slope of the mean-standard deviation frontier is higher if the economy is riskier – if consumption is more volatile – or if consumers are more risk averse.* Both situations naturally make consumers more reluctant to take on the extra risk of holding risky assets. We will come back to this slope in detail in the Hansen-Jagannathan bound and equity premium discussion below.

2.4.8 Time-varying expected returns and random walks.

Since all the moments above can be conditional and can vary as conditioning information varies, we can talk about variation of prices and returns over time as well as across assets.

$1 = E_t(m_{t+1}R_{t+1})$ implies

$$E_t(R_{t+1}) = R_t^f - \frac{cov_t(m_{t+1}, R_{t+1})}{E_t(m_{t+1})} \tag{16}$$

Expected excess returns were once thought to be constant over time. This idea makes intuitive sense and is still thought to hold quite well for short time horizons. A high return (or other news) today shouldn't signal a high return tomorrow, or mechanical strategies could make a lot of money.

Examining equation (16), however, we see that expected returns can be predictable–expected returns can vary over time. However, such predictability has to be explained by changing mean consumption growth, changing conditional covariance of return with consumption growth, or changing risk aversion (the function relating consumption to the discount factor). Matching the observed predictability of returns with these economic determinants is an empirical challenge, which we take up below.

The constant expected return idea is often expressed as "prices follow a random walk." (A *random walk* is a process $p_t = p_{t-1} + \varepsilon_t$. It is a special case of a *martingale* which has the property $p_t = E_t(p_{t+1})$.) Going back to the basic first order condition, or just multiply the now familiar consumption-based pricing equation by $u'(c_t)$,

$$p_t u'(c_t) = E_t[\beta u'(c_{t+1})(p_{t+1} + d_{t+1})].$$

Prices adjusted for dividend payments and scaled by utility *do* follow a martingale. Actual prices do not follow a martingale when something interesting is happening to the $u'(c)$ terms.

2.4.9 Present value statement.

It is convenient to use only the two period valuation, thinking of a price p_t and a payoff x_{t+1}. But there are times when we want to relate price to the entire cash flow stream. To do this, either maximize the entire expected utility $E_t \sum_j \beta^j u(c_{t+j})$ by purchasing a stream $\{d_{t+j}\}$ at price p_t, or just chain together the two period formula $p_t = E_t[m_{t+1}(p_{t+1}+d_{t+1})]$ (plus the "transversality condition" $\lim_{t\to\infty} E_t m_{t+j} p_{t+j} = 0$, which we will discuss in a lot of detail below), to express price as a stochastically discounted present value of the entire dividend stream,

$$p_t = E_t \sum_{j=0}^{\infty} \beta^j \frac{u'(c_{t+j})}{u'(c_t)} d_{t+j} = E_t \sum_{j=0}^{\infty} m_{t,t+j} d_{t+j}. \tag{17}$$

Remember, *everything* derived in this section just comes from manipulation of the consumer's first order condition for purchase of an asset. We have just rewritten those first order conditions in a lot of interesting ways.

Chapter 3. Discount factors in continuous time

Continuous time analogies to the basic pricing equations.

Discrete	Continuous
$p_t = E_t \sum_{j=0}^{\infty} \delta^t \frac{u'(c_{t+j})}{u'(c_t)} D_{t+j}$	$p_t u'(c_t) = E_t \int_{s=0}^{\infty} e^{-\delta s} u'(c_{t+s}) D_{t+s} ds$
$m_{t+1} = \delta \frac{u'(c_{t+1})}{u'(c_t)}$	$\Lambda_t = e^{-\delta t} u'(c_t)$
$p = E_t(mx)$	$0 = \Lambda D\ dt + E_t[d(\Lambda p)]$
$E(R) = R^f - R^f cov(m, R)$	$E_t \left(\frac{dp}{p}\right) + \frac{D}{p}\ dt = r_t^f dt - E_t \left[\frac{d\Lambda}{\Lambda} \frac{dp}{p}\right]$

It is often convenient to express asset pricing ideas in the language of continuous time stochastic differential equations rather than discrete time stochastic difference equations as I have done so far. The appendix contains a brief introduction to continuous time processes that covers what you need to know for this book. Even if one wants to end up with a discrete time representation, manipulations are often easier in continuous time.

First, we need to think about how to model securities, in place of price p_t and one-period payoff x_{t+1}. Let a generic security have price p_t at any moment in time, and let it pay dividends at the rate $D_t dt$. (I will continue to denote functions of time as p_t rather than $p(t)$ to maintain continuity with the discrete-time treatment, and I will drop the time subscripts where they are obvious, e.g. dp in place of dp_t. In an interval dt, the security pays dividends $D_t dt$. I use capital D for dividends to distinguish them from the differential operator d.)

The instantaneous total return is

$$\frac{dp_t}{p_t} + \frac{D_t}{p_t} dt.$$

Risky securities will in general have price processes that follow diffusions, for example

$$\frac{dp_t}{p_t} = \mu(\cdot)dt + \sigma(\cdot)dz.$$

(I will reserve dz for increments to a standard Brownian motion, e.g. $z_{t+\Delta} - z_t$ ˜ $\mathcal{N}(0, \Delta)$. I use the notation $(\cdot)$ to indicate that the drift and diffusions can be functions of state variables.)

We can think of a riskfree security as one that has a constant price equal to one and pays the riskfree rate as a dividend,

$$p = 1;\ D_t = r_t^f, \tag{18}$$

or as a security that pays no dividend but whose price climbs deterministically at a rate

$$\frac{dp_t}{p_t} = r_t^f \, dt. \tag{19}$$

Next, we need to express the first order conditions in continuous time. The utility function is

$$E_0 \int_{t=0}^{\infty} e^{-\delta t} u(c_t) dt$$

Suppose the consumer can buy a security whose price is p_t and that pays a dividend stream D_t. Then, the first order condition for buying the security at t and selling it at $t + \Delta$ is

$$p_t u'(c_t) = E_t \int_{s=0}^{\Delta} e^{-\delta s} u'(c_{t+s}) D_{t+s} ds + E_t \left[e^{-\delta \Delta} u'(c_{t+\Delta}) p_{t+\Delta} \right]$$

Right away, this first order condition gives us the infinite period version of the basic pricing equation,

$$p_t u'(c_t) = E_t \int_{s=0}^{\infty} e^{-\delta s} u'(c_{t+s}) D_{t+s} ds$$

This equation is an obvious continuous time analogue to

$$p_t = E_t \sum_{j=0}^{\infty} \delta^t \frac{u'(c_{t+j})}{u'(c_t)} D_{t+j}$$

It turns out that dividing by $u'(c_t)$ is not a good idea in continuous time, since the ratio $u'(c_{t+\Delta})/u'(c_t)$ isn't well behaved for small time intervals. Instead, we keep track of the *level* of marginal utility. Therefore, define the discount factor in continuous time as

$$\Lambda_t \equiv e^{-\delta t} u'(c_t).$$

Then we can write the first order condition as

$$p_t \Lambda_t = E_t \int_{s=0}^{\Delta t} \Lambda_{t+s} D_{t+s} ds + E_t \left[\Lambda_{t+\Delta t} p_{t+\Delta t} \right]$$

The analogue to $p = E(mx)$ is

$$0 = \Lambda D \, dt + E_t \left[d(\Lambda p) \right]. \tag{20}$$

Let's derive this fundamental equation. "One period" must mean dt in continuous time. Thus, for Δt small,

$$p_t \Lambda_t = E_t \Lambda_t D_t \Delta t + E_t \left[\Lambda_{t+\Delta t} p_{t+\Delta t} \right] \tag{21}$$

Introduce differences by

$$p_t \Lambda_t = \Lambda_t D_t \Delta t + E_t \left[\Lambda_t p_t + \left(\Lambda_{t+\Delta t} p_{t+\Delta t} - \Lambda_t p_t \right) \right] \tag{22}$$

canceling $p_t \Lambda_t$ and using the notation $\Delta x = x_{t+\Delta t} - x_t$,

$$0 = \Lambda_t D_t \Delta t + E_t \left[\Delta(\Lambda_t p_t) \right]$$

Taking the obvious limit at $\Delta t \rightarrow 0$,

$$0 = \Lambda_t D_t dt + E_t \left[d(\Lambda_t p_t) \right]$$

or, dropping time subscripts, (20)

Equation (20) looks different than $p = E(mx)$ because there is no price on the left hand side; we are used to thinking of the one period pricing equation as determining price at t given other things, including possibly price at $t+1$. But price at t is really here, or course, as you can see from equation (21) or (22). It is just easier to express the *difference* in price over time.

The object $d(\Lambda p)$ also looks a bit mysterious. It isn't: it is just the change (increment) in marginal utility weighted price. Since we will write down price processes for dp and discount factor processes for $d\Lambda$, it is often convenient to break up this term using Ito's lemma:

$$d(\Lambda p) = p d\Lambda + \Lambda dp + dp d\Lambda.$$

(If keeping the second order terms is still mysterious, go back to discrete time in equation 22.

$$\Lambda_{t+\Delta t} p_{t+\Delta t} - \Lambda_t p_t = p_t(\Lambda_{t+\Delta} - \Lambda_t) + \Lambda_t(p_{t+\Delta t} - p_t) + (\Lambda_{t+\Delta t} - \Lambda_t)(p_{t+\Delta t} - p_t).$$

Now you see where the $dp d\Lambda$ came from.) Using this expansion in the basic equation (20), and dividing by $p\Lambda$ to make it pretty, we obtain an equivalent, slightly less compact but slightly more intuitive version,

$$0 = \frac{D}{p} dt + E_t \left[\frac{d\Lambda}{\Lambda} + \frac{dp}{p} + \frac{d\Lambda}{\Lambda} \frac{dp}{p} \right]. \tag{23}$$

(This formula only works when both Λ and p can never be zero. That is often enough the case that this formula is useful. If not, multiply through by Λ and p and keep them in numerators.)

Applying the basic pricing equations (20) or (23) to a riskfree rate, defined as (18) or (19), we obtain

$$r_t^f dt = -E_t \frac{d\Lambda}{\Lambda}$$

This equation is the obvious continuous time equivalent to

$$R_t^f = \frac{1}{E_t(m_{t+1})}.$$

If a riskfree rate is not traded, we can define a shadow riskfree rate or zero-beta rate by

$$\alpha_t dt = -E_t \frac{d\Lambda}{\Lambda}.$$

With this interpretation, (23) can be rearranged as

$$E_t\left(\frac{dp_t}{p_t}\right) + \frac{D_t}{p_t}\, dt = \alpha_t dt - E_t\left[\frac{d\Lambda_t}{\Lambda_t}\frac{dp_t}{p_t}\right]. \tag{24}$$

This is the obvious continuous-time analogue to

$$E(R) = R^f - R^f cov(m, R).$$

The last term in 24 is the covariance of the return with the discount factor or marginal utility.

Ito's lemma makes many transformations simple in continuous time. For example, the transformation between consumption itself and the discount factor is easy in continuous time. With $\Lambda_t = e^{-\delta t}u'(c_t)$ we have

$$d\Lambda_t = -\delta e^{-\delta t}u'(c_t)dt + e^{-\delta t}u''(c_t)dc_t + \frac{1}{2}e^{-\delta t}u'''(c_t)dc_t^2$$

$$\frac{d\Lambda_t}{\Lambda_t} = -\delta dt + \frac{c_t u''(c_t)}{u'(c_t)}\frac{dc_t}{c_t} + \frac{1}{2}\frac{c_t^2 u'''(c_t)}{u'(c_t)}\frac{dc_t^2}{c_t^2}$$

The quantity

$$\gamma = -\frac{c_t u''(c_t)}{u'(c_t)}$$

is known as the *local curvature* of the utility function. It is also called the local *coefficient of risk aversion*. I prefer not to use this term: in a dynamic model the coefficient of risk aversion is really aversion to wealth bets, measured by the second partial derivative of the value function. Only in certain very restrictive cases is the value function curvature the same as the utility function curvature. This quantity is equal to the power in the power utility model $u'(c) = c^{-\gamma}$.

Using this formula we can quickly find the riskfree interest rate in terms of consumption growth,

$$r_t^f dt = -E_t\left(\frac{d\Lambda_t}{\Lambda_t}\right) = \delta dt + \gamma E_t\left(\frac{dc_t}{c_t}\right) - \frac{1}{2}\frac{c_t^2 u'''(c_t)}{u'(c_t)}E_t\left(\frac{dc_t^2}{c_t^2}\right).$$

We see the same economics at work as with the discrete time representation: interest rates are higher when consumers are more impatient (δ).Interest rates are higher if expected consumption growth is higher $E_t\left(dc_t/c_t\right)$, and interest rates are more sensitive to consumption growth if utility curvature is higher γ. Reading the same terms backwards, consumption growth is higher when interest rates are higher, since people save more now and spend it in the future, and consumption is less sensitive to interest rates as the desire for a smooth consumption stream, captured by γ, rises. In this role, γ plays the role of the *intertemporal substitution elasticity*. The final term is a *precautionary savings* term. If consumption is more volatile, consumers would like to "save for a rainy day", driving interest rates down.

We can also express asset prices in terms of consumption risk rather than discount factor risk. From the basic pricing equation

$$E_t\left(\frac{dp_t}{p_t}\right)+\frac{D_t}{p_t}\,dt-r_t^f\,dt=\gamma E_t\left[\frac{dc_t}{c_t}\frac{dp_t}{p_t}\right].$$

Thus, assets whose returns covary more strongly with consumption get higher mean returns, and the constant relating covariance to mean return is the utility curvature coefficient (coefficient of risk aversion) γ.

3.1 Assumptions and applicability

In deriving the basic pricing equation (2),

$$p_t=E_t\left[\beta\frac{u'(c_{t+1})}{u'(c_t)}x_{t+1}\right]$$

we have *not* assumed complete markets or a representative investor. This equation applies to each individual investor, for each asset to which he has access, independently of the presence or absence of other investors or other assets. Complete markets/representative agent assumptions are used if one wants to use aggregate consumption data in $u'(c_t)$, or other specializations and simplifications of the model.

We have *not* said anything about payoff or return distributions, multivariate normality, the form of the utility function etc. This basic pricing equation should also hold for *any* asset, stock, bond, option, real investment opportunity, etc. Such assumptions can be added for some special cases, but they aren't here yet. In particular, it is often thought that mean-variance analysis and beta pricing models require this kind of limiting assumptions, but that is not the case here. A mean-variance efficient return carries all pricing information no matter what the distribution of payoffs, utility function, etc.

This is *not* a "two-period model." The fundamental pricing equation holds for any two periods of a multi-period model. Investors can live forever. For example, we can start with the utility function

$$E_0 \sum_{t=0}^{\infty} \beta^t u(c_t). \tag{25}$$

The basic pricing equation is still the first order condition for buying an asset p_t with payoff x_{t+1}. In a multiperiod context it is more important to distinguish conditional from unconditional moments. Equation (17) results directly from the first order condition for paying p_t to receive $\{d_t, d_{t+1}...\}$.

I have written things down in terms of a time- and state-separable utility function as in (25), and I have extensively used the convenient power utility example. Nothing important lies in either choice. Just replace $u'(c_t)$ the partial derivative of a general utility function with respect to consumption at time t. We will look at several examples below.

We do *not* assume that investors have no non-marketable human capital, or no outside sources of income. The first order conditions for purchase of an asset relative to consumption hold no matter what else is in the budget constraint. By contrast, the portfolio approach to asset pricing as in the CAPM and ICAPM relies heavily on the assumption that the investor has no non-asset income, and we will study these special cases below.

We don't even really need the assumption (yet) that the market is "in equilibrium," that consumer has bought all of the asset that he wants to, or even that he can buy the asset at all. We can interpret the formula as giving us the value, or willingness to pay for, a small amount of a payoff x_{t+1} that the consumer does not yet have. Here's why: If the investor had a little ξ more of the payoff x_{t+1} time $t+1,$ his utility $u(c_t) + \beta E_t u(c_{t+1})$ would increase by

$$\beta E_t \left[u(c_{t+1} + \xi x_{t+1}) - u(c_{t+1}) \right] \approx \beta E_t \left[u'(c_{t+1}) x_{t+1} \xi + \frac{1}{2} u''(c_{t+1}) \left(x_{t+1} \xi \right)^2 + ... \right]$$

If ξ is small, only the first term on the right matters. If the investor has to give up a small amount of money $v_t \xi$ at time t, that loss lowers his utility by

$$u(c_t - v_t \xi) \approx u'(c_t) p_t \xi + \frac{1}{2} u''(c_t) \left(p_t \xi \right)^2 .$$

Again, for small ξ, only the first term matters. Therefore, in order to receive the small amount ξx_{t+1}, the consumer is willing to pay the small amount $v_t \xi$ where

$$v_t = E_t \left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1} \right] .$$

If this private valuation is higher than the market value p_t, and if the consumer can buy some more of the asset, he will. As he buys more, his consumption will change; it will be higher in states where x_{t+1} is higher, driving down $u'(c_{t+1})$ in those states, until the value to the investor has declined to equal the market value. Thus, *after* an investor has reached his optimal portfolio, the *market value* should obey the basic pricing equation as well, using post-trade or equilibrium consumption. But the formula can also be applied to generate the marginal *private* valuation, using pre-trade consumption, or to value a *potential*, not yet traded security.

We *have* calculated the value of a "small" or marginal portfolio change for the investor. For some investment projects, an investor cannot take a small position. Then the value of a project not already taken, $E\sum\beta^j u(c_{t+j}+x_{t+j})$ might be substantially different from its marginal counterpart, $E\sum\beta^j u'(c_{t+j})x_{t+j}$. Once taken of course, $c_{t+j}+x_{t+j}$ becomes c_{t+j}, so the marginal valuation still applies to the ex-post consumption stream.

3.2 Consumption-based model in practice

The consumption-based model is, in principle, a complete answer to all asset pricing questions, but works poorly in practice. This observation motivates other asset pricing models.

The model we have sketched so far can, in principle, give a compete answer to all the questions of the theory of valuation. It can be applied to *any* security—bonds, stocks, options, futures, etc.—or to any uncertain cash flow. All we need is a functional form for utility, numerical values for the parameters, and a statistical model for the conditional distribution of consumption.

To be specific, consider the power utility function

$$u'(c) = c^{-\gamma}. \tag{26}$$

Then, excess returns should obey

$$0 = E_t\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} R^e_{t+1}\right] \tag{27}$$

Taking unconditional expectations and applying the covariance decomposition, expected excess returns should follow

$$E(R^e_{t+1}) = -\frac{cov\left[\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}, R^e_{t+1}\right]}{E\left[\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right]}. \tag{28}$$

Given a value for γ, and data on consumption and returns, one can easily estimate the mean and covariance on the right hand side, and check whether actual expected returns are, in fact, in accordance with the formula.

Similarly, the present value formula is

$$p_t = E_t \sum_{j=1}^{\infty} \beta^j \left(\frac{c_{t+j}}{c_t}\right)^{-\gamma} d_{t+j}. \tag{29}$$

Given data on consumption and dividends or another stream of payoffs, we can estimate the right hand side and check it against prices on the left.

Bonds and options do *not* require separate valuation theories. For example, an N-period pure discount default-free bond is a claim to one dollar at time $t+N$. Its price should be

$$p_t = E_t \left(\beta^N \left(\frac{c_{t+N}}{c_t}\right)^{-\gamma} \frac{\Pi_t}{\Pi_{t+N}} 1 \right)$$

where Π = price level ($/good). A European option is a claim to $\max(S_{t+T} - K, 0)$, where S_{t+T} = stock price at time $t+T$, $K =$ strike price. The option price should be

$$p_t = E_t \left[\beta^T \left(\frac{c_{t+T}}{c_t}\right)^{-\gamma} \max(S_{t+T} - K, 0) \right]$$

again, we can use data on consumption, prices and payoffs to check these predictions.

Unfortunately, the above specification of the consumption-based model does not work very well. To give a flavor of some of the problems, Figure 2 presents the mean return on the ten size-ranked portfolios of NYSE stocks vs. the predictions (right hand side of 2) of the consumption-based model. I picked the parameter γ to make the picture look as good as possible (The section on GMM estimation below goes into detail on how to do this.) As you can see, the model isn't hopeless–there is some correlation between mean returns and predictions. But the model does not do very well. The pricing error (actual expected return - predicted expected return) for each portfolio is of the same order of magnitude as the spread in expected returns across the portfolios.

3.3 Alternative asset pricing models: Overview

I motivate exploration of different utility functions, general equilibrium models, and linear factor models such as the CAPM, APT and ICAPM as approaches to circumvent the empirical difficulties of the consumption-based model.

The poor empirical performance of the consumption-based model motivates a search for alternative asset pricing models – alternative functions $m = f(\text{data})$. *All* asset pricing models amount to different functions for m. I give here a very bare sketch of some of the different approaches.

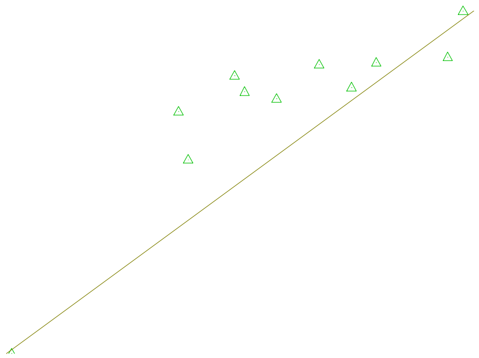


Figure 2. Mean excess returns of 10 CRSP size portfolios vs. predictions of the power utility consumption-based model.

1) *Different utility functions.* Perhaps the problem with the consumption-based model is simply the functional form we chose for utility. The natural response is to try different utility functions. Which variables determine marginal utility is a far more important question than the functional form. Perhaps the stock of durable goods influences the marginal utility of nondurable goods; perhaps leisure or yesterday's consumption affect today's marginal utility. These possibilities are all instances of *nonseparabilities.* One can also try to use micro data on individual consumption of stockholders rather than aggregate consumption. Aggregation of heterogenous consumers can make variables such as the cross-sectional variance of income appear in aggregate marginal utility.

2) *General equilibrium models.* Perhaps the problem is simply with the consumption *data.* General equilibrium models deliver equilibrium decision rules linking consumption to other variables, such as income, investment, etc. Substituting the decision rules $c_t = f(y_t, i_t, \ldots)$ in the consumption-based model, we can link asset prices to other, hopefully better-measured macroeconomic aggregates.

In addition, true general equilibrium models completely describe the economy, including the stochastic process followed by all variables. They can answer questions such as *why* is the covariance (beta) of an asset payoff x with the discount factor m the value that it is, rather than take this covariance as a primitive. They can in principle answer structural questions, such as how asset prices might be affected by different government policies. Neither kind of question can be answered by just manipulating consumer first order conditions.

3) *Factor pricing models.* We don't have satisfactory general equilibrium models. An-

other sensible response to bad consumption data is to model marginal utility in terms of other variables directly. *Factor pricing* models follow this approach. They just specify that the discount factor is a linear function of a set of proxies,

$$m_{t+1} = a + b_1 f^1_{t+1} + b_2 f^2_{t+1} + \ldots . \tag{30}$$

where f^i are *factors* and a, b_i are parameters. (This is a different sense of the use of the word "factor" than "discount factor." I didn't invent the confusing terminology.) Among others, the Capital Asset Pricing Model (CAPM) is the model

$$m_{t+1} = a + bR^W_{t+1}$$

where R^W is the rate of return on a claim to total wealth, often proxied by a broad-based portfolio such as the value-weighted NYSE portfolio. The Arbitrage Pricing Theory (APT) uses returns on broad-based portfolios derived from a factor analysis of the return covariance matrix. The Intertemporal Capital Asset Pricing Model (ICAPM) suggests macroeconomic variables such as GNP and inflation and variables that forecast macroeconomic variables or asset returns as factors. Term structure models such as the Cox-Ingersoll-Ross model specify that the discount factor is a function of a few term structure variables, for example the short rate of interest. Many factor pricing models are derived as general equilibrium models with linear technologies and no labor income; thus they also fall into the general idea of using general equilibrium relations to substitute out for consumption.

4) *Arbitrage or near-arbitrage pricing.* The mere existence of a representation $p = E(mx)$ and the fact that marginal utility is positive $m \geq 0$ (these facts are discussed in the next chapter) can be used to deduce prices of one payoff in terms of the prices of other payoffs. The Black-Scholes option pricing model is the paradigm of this approach: Since the option payoff can be replicated by a portfolio of stock and bond, any m that prices the stock and bond gives the same price for the option. Recently, there have been several suggestions on how to use this idea in more general circumstances by using very weak further restrictions on m.

We return to a more detailed derivation and discussion of these alternative models of the discount factor m below. First, and with this brief overview in mind, we look at $p = E(mx)$ and what the discount factor m represents in a little more detail.

Chapter 4. The discount factor

Now we look more closely at the discount factor. Rather than derive a specific discount factor as with the consumption-based discount factor above, I work backwards. A discount factor is just some random variable that generates prices from payoffs, $p = E(mx)$. What does this expression mean? Can one always find such a discount factor? Can we use the above convenient representations without all the structure of the consumers, utility functions, and so forth? Along the way I introduce the inner product representation which allows an intuitive visual representation of most of the theorems, and the idea of contingent claims.

I start by deriving the fact that discount factors exist, are positive, and the pricing function is linear from a complete markets or contingent claim framework. Then I show that these properties can be built up, without investors, utility functions and the rest, even in incomplete markets.

The chapter ends with two famous theorems. The *law of one price* states that if two portfolios have the same payoffs (in every state of nature), then they must have the same price. A violation of this law would give rise to an immediate kind of arbitrage profit, as you could sell the expensive version and buy the cheap version of the same portfolio. The first theorem is that this law of one price holds if and only if there is a discount factor that prices all the payoffs by $p = E(mx)$.

In finance, we reserve the term *absence of arbitrage* for a stronger idea, that if payoff A always at least as good as payoff B, and sometimes A is better, then the price of A must be greater than the price of B. The second theorem is that there are no arbitrage opportunities of this type if and only if there is a *positive* discount factor that prices all the payoffs by $p = E(mx)$.

These theorems are seful as a justification for using discount factors without all the structure we have imposed so far. More importantly, they show how many aspects of a *payoff space* (such as absence of arbitrage) can be conveniently captured by restrictions on the *discount factor* (such as it exists, or it is positive). Later, it will be much more convenient to impose, check, and intersect restrictions on the discount factor rather than to do so for all possible portfolios priced by that discount factor.

4.1 Contingent claims

I describe contingent claims. I interpret the stochastic discount factor m as contingent claims prices divided by probabilities, and $p = E(mx)$ as a bundling of contingent claims. I also interpret the discount factor m as a transformation to risk-neutral probabilities such that $p = E^*(x)/R^f$.

Suppose that one of S possible *states of nature* can occur tomorrow, i.e. specialize to a finite-dimensional state space. Denote the individual states by s. For example, we might have $S = 2$ and $s =$ rain or $s =$ shine.

A *contingent claim* is a security that pays one dollar (or one unit of the consumption good) in one state s only tomorrow. $pc(s)$ is the price today of the contingent claim. I write pc to specify that it is the price of a contingent claim and (s) to denote in which state s the claim pays off.

In a *complete market* investors can buy any contingent claim. They don't necessarily have to be faced with explicit contingent claims; they just need enough other securities to *span* or *synthesize* all contingent claims. For example, if the possible states of nature are (rain, shine), securities that pay 2 dollars if it rains and one if it shines, or $x_1 = (2, 1)$ and a riskfree security whose payoff pattern is $x_2 = (1, 1)$ are enough to span or synthesize any portfolio achieved by contingent claims. More practically, we see below that European options with every possible strike price span all claims contingent on the underlying asset's price.

If there are complete contingent claims, a discount factor exists, and it is equal to the contingent claim price divided by probabilities.

Let $x(s)$ denote an asset's payoff in state of nature s. We can think of the asset as a bundle of contingent claims—$x(1)$ contingent claims to state 1, $x(2)$ claims to state 2, etc. The asset's price must then equal the value of the contingent claims of which it is a bundle,

$$p(x) = \sum_s pc(s)x(s). \tag{31}$$

I denote the price $p(x)$ to emphasize it is the price of the payoff x. Where the payoff in question is clear, I suppress the (x). I like to think of equation (31) as happy-meal logic: the price of a happy meal (in a frictionless market) should be the same as the price of one hamburger, one small fries, one small drink and the toy.

It is easier to take expectations rather than sum over states. To this end, multiply and divide the bundling equation (31) by probabilities,

$$p(x) = \sum_s \pi(s) \left(\frac{pc(s)}{\pi(s)} \right) x(s)$$

where $\pi(s)$ is the probability that state s occurs. Then define m as the ratio of contingent claim price to probability

$$m(s) = \frac{pc(s)}{\pi(s)}.$$

Now we can write the bundling equation as an expectation as

$$p=\sum_s \pi(s)m(s)x(s)=E(mx).$$

Thus, in a complete market, the stochastic discount factor m in $p=E(mx)$ exists, and it is just a set of contingent claims prices, scaled by probabilities. As a result of this interpretation, the discount factor is sometimes called a *state-price density*.

The multiplication and division by probabilities seems very artificial in this finite-state context. In general, we posit states of nature ω that can take continuous (uncountably infinite) values in a space Ω. In this case, the sums become integrals, and we have to use *some* measure to integrate over Ω. Thus, scaling contingent claims prices by some probability-like object is unavoidable.

Risk neutral probabilities.

Another common transformation of $p=E(mx)$ results in "risk-neutral" probabilities. Define

$$\pi^*(s)\equiv R^f m(s)\pi(s)=R^f pc(s)$$

where

$$R^f\equiv 1/\sum pc(s)=1/E(m).$$

The $\pi^*(s)$ are positive, less than or equal to one ($pc(s)\leq 1/R^f$, the price of a sure unit of consumption), and sum to one, so they are a legitimate set of probabilities. Then we can rewrite the asset pricing formula as

$$p(x)=\sum_s pc(s)x(s)=\frac{1}{R^f}\sum \pi^*(s)x(s)=\frac{E^*(x)}{R^f}.$$

I use the notation E^* to remind us that the expectation uses the *risk neutral probabilities* π^* instead of the real probabilities π.

Thus, we can think of asset pricing as if agents are all risk neutral, but with probabilities π^* in the place of the true probabilities π. The probabilities π^* gives greater weight to states with higher than average marginal utility m. There is something very deep here: risk aversion is equivalent to paying more attention to unpleasant states, relative to their actual probability of occurrence. People who report high subjective probabilities of unpleasant events like plane crashes may not have irrational expectations, they may simply be reporting the risk neutral probabilities or the product $m\times\pi$. This product is after all the most important piece of information for many decisions: pay a lot of attention to contingencies that are either highly probable or that are improbable but have disastrous consequences.

The transformation from actual to risk-neutral probabilities is given by

$$\pi^*(s) = \frac{m(s)}{E(m)}\pi(s).$$

We can also think of the discount factor m as the *derivative* or *change of measure* from the real probabilities π to the subjective probabilities π^*.

The risk-neutral probability representation of asset pricing is often used to simplify option pricing formulas. I avoid it, however, because it is far too easy to brush over the difference between risk-neutral and actual probabilities. *Everything* really interesting in asset pricing concerns how to make this transformation, or how to make risk-adjustments to expected present value formulas.

4.2 Investors again

Investor's first order conditions with contingent claims.

Marginal rate of substitution

It's worth looking at the investor's first order conditions again in the contingent claim context. The investor starts with a pile of initial wealth and a state-contingent income. He purchases contingent claims to each possible state in the second period. His problem is then

$$\max_{\{c,c(s)\}} \; u(c) + \sum_s \beta\pi(s)u[c(s)] \;\; s.t. \;\; c + \sum_s pc(s)c(s) = y + \sum_s pc(s)y(s).$$

Introducing a Lagrange multiplier λ on the budget constraint, the first order conditions are

$$u'(c) = \lambda$$

$$\beta\pi(s)u'[c(s)] = \lambda pc(s).$$

Eliminating the Lagrange multiplier λ,

$$pc(s) = \beta\pi(s)\frac{u'[c(s)]}{u'(c)}$$

or

$$m(s) = \frac{pc(s)}{\pi(s)} = \beta\frac{u'[c(s)]}{u'(c)}$$

Coupled with $p = E(mx)$, we obtain the consumption-based model again.

The investor's first order conditions say that marginal rates of substitution between *states* tomorrow equals the relevant price ratio,

$$\frac{m(s_1)}{m(s_2)} = \frac{u'[c(s_1)]}{u'[c(s_2)]}.$$

$m(s_1)/m(s_2)$ gives the rate at which the consumer can give up consumption in state 2 in return for consumption in state 1 through purchase and sales of contingent claims. $u'[c(s_1)]/u'[c(s_2)]$ gives the rate at which the consumer is willing to make this substitution. At an optimum, the two rates should be equal.

We learn that the discount factor m is the marginal rate of substitution between date *and state* contingent commodities. That's why it, like $c(s)$, is a random variable. Also, scaling contingent claims prices by probabilities gives marginal utility, and so is not so artificial as it may have seemed above.

Figure 3 gives the economics behind this approach to asset pricing.. We observe the consumer's choice of date or state-contingent consumption. Once we know his utility function, we can calculate the contingent claim prices that must have led to the observed consumption choice, from the derivatives of the utility function.

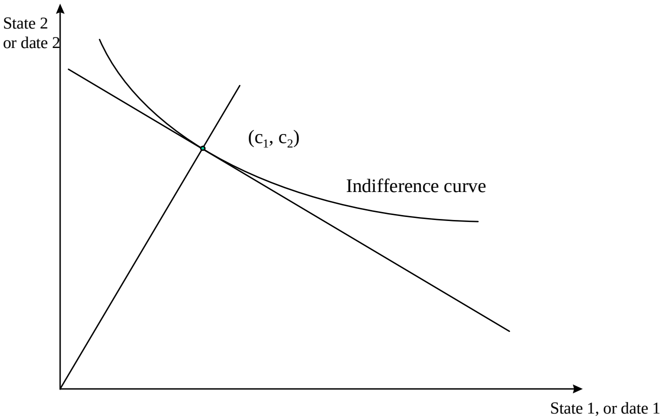


Figure 3. Indifference curve and contingent claim prices

The relevant probabilities are the consumer's *subjective* probabilities over the various states. Asset prices are set, after all, by consumer's demands for assets, and those demands are set by consumer's subjective evaluations of the probabilities of various events. We often assume *rational expectations*, namely that subjective probabilities are equal to objective frequencies. But this is an additional assumption that we may not always want to make.

4.3 Risk sharing

Risk sharing: In complete markets, consumption moves together. Only aggregate risk matters for security markets.

The above derivation holds for any consumer. But the prices are the same for all consumers. Therefore, *marginal utility growth should be the same for all consumers*

$$\beta^i \frac{u'(c^i_{t+1})}{u'(c^i_t)} = \beta^j \frac{u'(c^j_{t+1})}{u'(c^j_t)} \tag{32}$$

where i and j refer to different consumers. If consumers have the same utility function, then consumption itself should move in lockstep,

$$\frac{c^i_{t+1}}{c^i_t} = \frac{c^j_{t+1}}{c^j_t}.$$

These results mean that in a complete contingent claims market, all consumers share all risks. This risk sharing is *Pareto-optimal.* Suppose a social planner wished to maximize everyone's utility given the available resources. For example, with two consumers i and j, he would maximize

$$\max \sum \beta^{it} u(c^i_t) + \lambda \sum \beta^{jt} u(c^j_t) \;\; s.t. \; c^i_t + c^j_t = c^a_t$$

where c^a is the total amount available. The first order condition to this problem is

$$\beta^{it} u'(c^i_t) = -\lambda \beta^{jt} u'(c^j_t)$$

and hence the same risk sharing that we see in a complete market, equation (32).

This simple fact has profound implications. First, it shows you why *only aggregate shocks should matter for risk prices.* Any idiosyncratic income risk will be insured away through asset markets.

In addition, it highlights the function of security markets and much of the force behind financial innovation. Security markets – state-contingent claims – are what brings individual consumptions closer together by allowing people to share risks. Many successful new securities can be understood as devices to more widely share risks.

4.4 State diagram and price function

I introduce the state space diagram and inner product representation for prices, $p(x) = E(mx) = m \cdot x$.

$p(x) = E(mx)$ implies $p(x)$ is a linear function.

Think of the contingent claims price pc and asset payoffs x as vectors in $\mathcal{R}^S$, where each element gives the price or payoff to the corresponding state,

$$pc = \left[\begin{array}{cccc} pc(1) & pc(2) & \cdots & pc(S) \end{array}\right]',$$

$$x = \left[\begin{array}{cccc} x(1) & x(2) & \cdots & x(S) \end{array}\right]'.$$

Figure 4 is a graph of these vectors in $\mathcal{R}^S$. Next, I deduce the geometry of Figure 4.

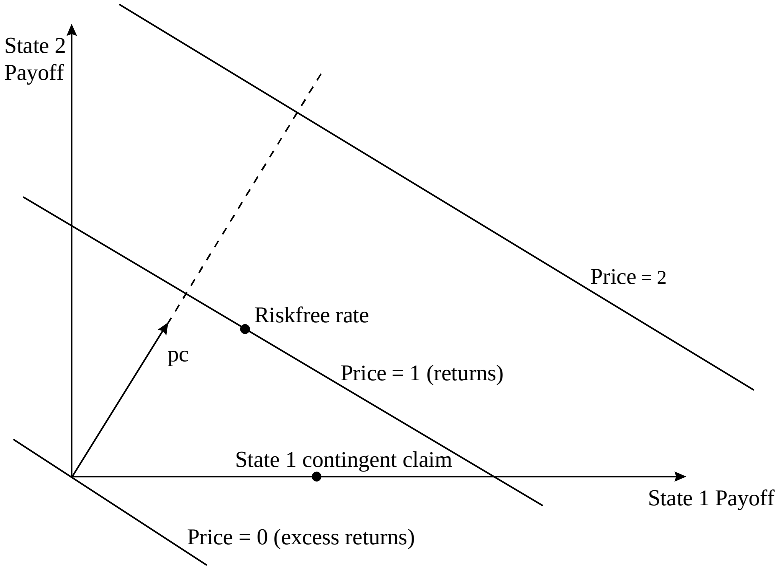


Figure 4. Contingent claims prices (pc) and payoffs.

The contingent claims price vector pc points in to the positive orthant. We saw in the last section that $m(s) = u'[c(s)]/u'(c)$. Now, marginal utility should always be positive (people always want more), so the marginal rate of substitution and discount factor are always nonnegative, $m > 0$ and $pc > 0$. This fact is important and related to the principle of no arbitrage below.

The set of payoffs with any given price lie on a (hyper)plane orthogonal to the contingent claim price vector. We reasoned above that the price of the payoff x must be given by its contingent claim value (31),

$$p(x) = \sum_s pc(s)x(s). \tag{33}$$

Interpreting pc and x as vectors, this means that the price is given by the *inner product* of the contingent claim price and the payoff. Recall that the inner product of two vectors x and pc equals the product of the magnitude of the projection of x onto pc times the magnitude of pc. Using a dot to denote inner product,

$$p(x) = \sum_s pc(s)x(s) = pc \cdot x = |pc||proj(x|pc)|$$

where $|x|$ means the length of the vector x. Since all payoffs on a plane orthogonal to pc have the same projection onto pc, they must have the same price.

Planes of constant price move out linearly, and the origin $x = 0$ must have a price of zero. If payoff $y = 2x$, then its price is twice the price of x,

$$p(y) = \sum_s pc(s)y(s) = \sum_s pc(s)2x(s) = 2\,pc(x).$$

Similarly, a payoff of zero must have a price of zero.

We can think of $p(x)$ as a pricing *function*, a map from the state space or payoff space in which x lies ($\mathcal{R}^S$ in this case) to the real line. We have just deduced from the definition (33) that $p(x)$ is a *linear function*, i.e. that

$$p(ax + by) = ap(x) + bp(y).$$

The constant price lines in Figure 4 are of course exactly what one expects from a linear function from $\mathcal{R}^S$ to $\mathcal{R}$. (One might draw the price on the z axis coming out of the page. Then the price function would be a plane going through the origin and sloping up with iso-price lines as given in Figure 4.)

Figure 4 also includes the payoffs to a contingent claim to the first state. This payoff is one in the relevant state and zero in other states and thus located on the axis. The plane of price = 1 payoffs is the plane of asset *returns*; the plane of price = 0 payoffs is the plane of *excess returns*. A riskfree unit payoff (the payoff to a risk-free pure discount bond) would lie on the $(1, 1)$ point in Figure 4; the riskfree return lies on the intersection of the 45^o line (same payoff in both states) and the price = 1 plane (the set of all returns).

Geometry with m in place of pc.

The geometric interpretation of Figure 4 goes through with the discount factor m in the place of pc. We can define an inner product between the random variables x and y by

$$x \cdot y \equiv E(xy),$$

and retain all the properties of an inner product. For this reason, random variables for which $E(xy) = 0$ are often called "orthogonal."

When the inner product is defined by a second moment, the operation "project y onto x" is a *regression.* (If x does not include a constant, you don't add one.) To see this fact, the idea of projection is to define

$$y = proj(y|x) + \varepsilon$$

in such a way that the residual ε is orthogonal to the projection,

$$\varepsilon \cdot proj(y|x) = E\left[\varepsilon \times proj(y|x)\right] = 0$$

This property is achieved by the construction

$$proj(y|x) = (x \cdot x)^{-1}(x \cdot y)\; x = E(x^2)^{-1}E(xy)\; x$$

which is the formula for OLS regression. For this reason, econometrics books often graph OLS regression as a projection of a point y on to a plane spanned by x with a residual ε that is at right angles to the plane x. We use the same geometry.

The geometric interpretation of Figure 4 also is valid if we generalize the setup to an infinite dimensional state space. Instead of vectors, which are functions from $\mathcal{R}^S$ to $\mathcal{R}$, random variables are (measurable) functions from Ω to $\mathcal{R}$. Nonetheless, we can still think of them as vectors. The equivalent of $\mathcal{R}^s$ is now a *Hilbert Space* L^2, which denotes spaces generated by linear combinations of square integrable *functions* from Ω to the real line, or the space of random variables with finite second moments. We can still define an "inner product" between two such elements by $x \cdot y = E(xy)$. In particular, $p(x) = E(mx)$ can still be interpreted as "m is orthogonal to (hyper)planes of constant price." *Proving* theorems in this context is much harder, and you are referred to the references (Especially Hansen and Richard (1987)) for such proofs.

4.5 Law of one price and existence of a discount factor

Definition of law of one price.

$p = E(mx)$ implies law of one price.

The law of one price implies that a discount factor exists: There exists a unique x^* in $\underline{X}$ such that $p = E(x^* x)$ for all $x \in \underline{X}$ = space of all available payoffs.

Furthermore, for any valid discount factor m,

$$x^* = \text{proj}(m \mid \underline{X}).$$

So far we have derived the basic pricing relation $p = E(mx)$ from environments with a lot of structure: either the consumption-based model or complete markets. We deduced that in any sensible model with consumers, the discount factor should be positive, and we deduced that price is a linear function of payoff in a contingent claim market.

Does thinking about asset pricing in this way require all this structure? Suppose we observe a set of prices p and payoffs x, and that markets — either the markets faced by investors or the markets under study in a particular application — are *incomplete*, meaning they do not span the entire set of contingencies. In what minimal set of circumstances does some discount factor exists which represents the observed prices by $p = E(mx)$? This section and the following answer this important question. This treatment is a simplified version of Hansen and Richard (1987), which contains rigorous proofs and some technical assumptions.

4.5.1 The theorem

Payoff space

The *payoff space* $\underline{X}$ is the set (or a subset) of all the payoffs that investors can purchase, or it is a subset of the tradeable payoffs that is used in a particular study. For example, if there are complete contingent claims to S states of nature, then $\underline{X} = \mathcal{R}^S$. But the whole point is that markets are (as in real life) *incomplete*, so we will generally think of $\underline{X}$ as a proper subset of complete markets $\mathcal{R}^S$.

The payoff space will include some set of primitive assets, but investors can also form new payoffs by forming portfolios. I assume that investors can form any portfolio of traded assets:

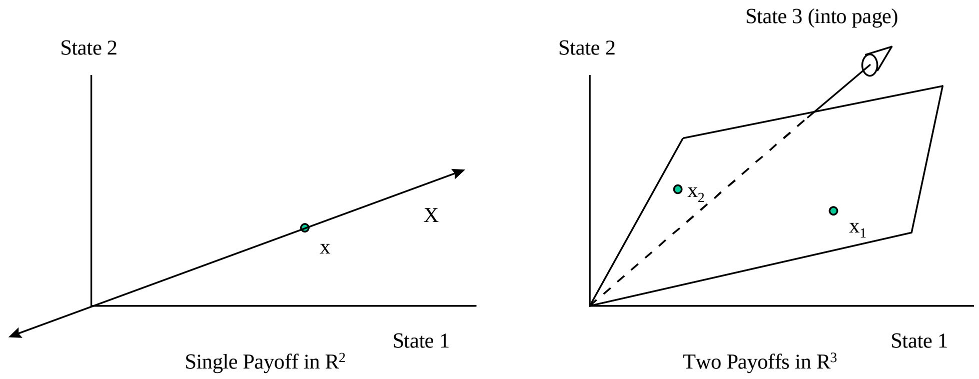
A1: (Portfolio formation) $x_1, x_2 \in \underline{X} \Rightarrow ax_1 + bx_2 \in \underline{X}$ for any real a, b.

Of course, $\underline{X} = \mathcal{R}^S$ for complete markets satisfies the portfolio formation assumption. If there is a single basic payoff x, then the payoff space must be at least the ray from the origin through x. If there are two basic payoffs in $\mathcal{R}^3$, then the payoff space $\underline{X}$ must include the plane defined by these two payoffs and the origin. Figure 5 illustrates these possibilities.

The payoff space is *not* the space of returns. The return space is a subset of the payoff space; if a return R is in the payoff space, then you can pay a price \$2 to get a payoff $2R$, so the payoff $2R$ with price 2 is also in the payoff space. Also, $-R$ is in the payoff space.

Free portfolio formation is in fact an important and restrictive simplifying assumption. It rules out short sales constraints, bid/ask spreads, leverage limitations and so on. The theory can be modified to incorporate these frictions, and I treat this modification later.

If investors can form portfolios of a vector of basic payoffs $\mathbf{x}$ (say, the returns on the NYSE stocks), then the payoff space consists of all portfolios or linear combinations of these original payoffs $\underline{X} = \{c'\mathbf{x}\}$. We also can allow truly infinite-dimensional payoff spaces. For example, consumers might be able to trade *nonlinear* functions of a basis payoff $\mathbf{x}$, such as



Single Payoff in R^2

Two Payoffs in R^3

Figure 5. Payoff spaces X generated by one (left) and two (right) basis payoffs.

call options on x with strike price K, which have payoff $\max\left[x(s) - K, 0\right]$.

The law of one price.

A2: (Law of one price, linearity) $p(ax_1 + bx_2) = ap(x_1) + bp(x_2)$

It doesn't matter how one forms the payoff x. The price of a burger, shake and fries must be the same as the price of a happy meal. Keep in mind that we are describing a market that has already reached equilibrium. The point is that if there are any violations of the law of one price, traders will quickly eliminate them so they can't survive in equilibrium. Graphically, if the iso-price curves were not planes, then one could buy two payoffs on the same iso-price curve, form a portfolio, which is on the straight line connecting the two original payoffs, and sell it for a higher price than the cost of the portfolio. Thus, law of one price basically says that investors can't make profits by repackaging portfolios. It is a (weak) characterization of preferences.

A1 and A2 also mean that the 0 payoff must be available, and must have price 0.

The Theorem

The existence of a discount factor implies the law of one price. As we have already seen, $p(x) = E(mx)$ implies linearity, and linearity implies the law of one price. This is obvious to the point of triviality: if $x = y + z$ then $E(mx) = E[m(y + z)]$. More directly, if the law of one price were violated, investors would take infinite positions and make infinite profits. Hence, the law of one price is a weak *implication* of the utility-based framework.

Our basic theorem in this section reverses this logic. We show that the *law of one price* implies the *existence of a discount factor*. Even if all we know about investors is that they can see past packaging and will take sure profits available from packaging, that is enough to guarantee the existence of a discount factor.

A1 and A2 imply that the price *function* on $\underline{X}$ looks like Figure4: parallel hyperplanes marching out from the origin. The only difference is that $\underline{X}$ may be a subspace of the original state space, as shown in Figure 5. We are ready to prove that a discount factor exists.

Theorem: Given free portfolio formation A1, the law of one price A2, there exists a unique payoff $x^* \in \underline{X}$ such that $p(x) = E(x^*x)$ for all $x \in \underline{X}$.

x^* is a discount factor. I offer an algebraic and a geometric proof.

Proof 1: (Algebraic.) We can prove the theorem by construction when the payoff space $\underline{X}$ is generated by portfolios of a N basis payoffs (for example, N stocks). Organize the basis payoffs into a vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & ... & x_N \end{bmatrix}'$ and similarly their prices $\mathbf{p}$. The payoff space is then $\underline{X} = \{\mathbf{c}'\mathbf{x}\}$. We want a discount factor that is in the payoff space, as the theorem requires. Thus, it must be of the form $x^* = \mathbf{b}'\mathbf{x}$. Find $\mathbf{b}$ so that x^* prices the basis assets. We want $\mathbf{p} = E(x^*\mathbf{x}) = E(\mathbf{x}\mathbf{x}'\mathbf{b})$. Thus we need $\mathbf{b} = E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{p}$. If $E(\mathbf{x}\mathbf{x}')$ is nonsingular, this $\mathbf{b}$ exists and is unique. A2 implies that $E(\mathbf{x}\mathbf{x}')$ is nonsingular. Thus, $x^* = \mathbf{p}'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}$ is our discount factor. It is a linear combination of $\mathbf{x}$ so it is in $\underline{X}$. It prices the basis assets by construction. It prices every $x \in \underline{X} : E[x^*(\mathbf{x}'\mathbf{c})] = E[\mathbf{p}'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}\mathbf{x}'\mathbf{c}] = \mathbf{p}'\mathbf{c}$. By linearity, $p(\mathbf{c}'\mathbf{x}) = \mathbf{c}'\mathbf{p}$.

Proof 2: (Geometric.) We have established that the price is a linear function as shown in Figure 6. (Figure 6 can be interpreted as the plane $\underline{X}$ of a larger dimensional space as in the right hand panel of Figure 5, laid flat on the page for clarity.) Now, to every plane we can draw a line from the origin at right angles to the plane. Choose a vector x^*on this line. The inner product between any payoff x on the price=1 plane $x^*$is $x \cdot x^* = |proj(x|x^*)| \times |x^*|$ Thus, every payoff on the price = 1 plane has the *same* inner product with x^*. All we have to do is pick x^* to have the right length, and we obtain $p(x) = 1 = x^* \cdot x = E(x^*x)$ for every x on the price = 1 plane. Then, of course we have $p(x) = x^* \cdot x = E(x^*x)$ for payoffs x on the other planes as well. Thus, the *linear* pricing function implied by the Law of One Price can be *represented* by inner products with x^*. $\square$.

You can see that the basic mathematics here is just that any linear function can be represented by an inner product. This theorem extends to infinite-dimensional spaces too. In this case, the *Riesz representation theorem* says that there is always a "line" orthogonal to any "plane", so one can always represent a linear function $p(x)$ by an inner product $p(x) = E(x^*x)$. See Hansen and Richard (1987) for the details.

What the theorem does and does not say

The theorem says there is a unique x^* *in* $\underline{X}$. There may be many other discount factors m *not* in $\underline{X}$. In fact, unless markets are complete, there are an *infinite* number of random variables that satisfy $p = E(mx)$. If $p = E(mx)$ then $p = E[(m + \varepsilon)x]$ for any ε orthogonal to x, $E(\varepsilon x) = 0$.

Not only does this construction generate some additional discount factors, it generates

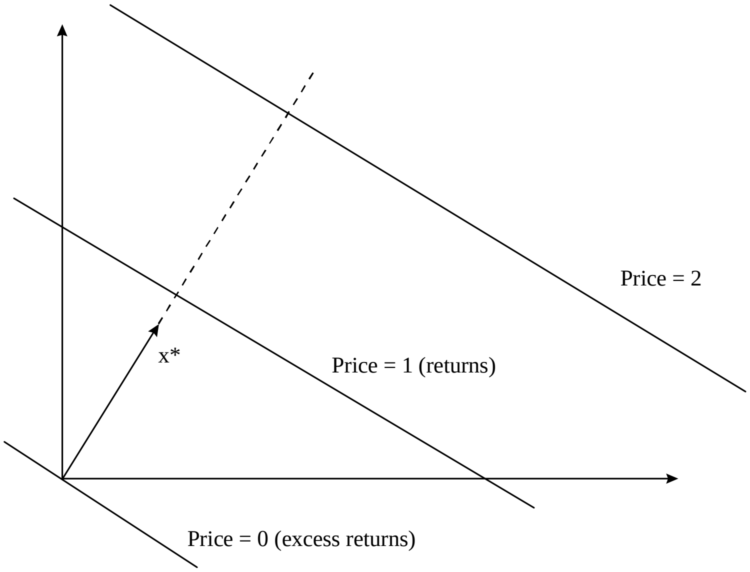


Figure 6. Existence of a discount factor x^*.

all of them: *Any discount factor* m (any random variable that satisfies $p = E(mx)$) *can be represented as* $m = x^* + \varepsilon$ with $E(\varepsilon x) = 0$. Figure 7 gives an example of a one-dimensional $\underline{X}$ in a two-dimensional state space, in which case there is a whole line of possible discount factors m. If markets are complete, there is nowhere to go orthogonal to the payoff space $\underline{X}$, so x^* is the only possible discount factor.

Reversing the argument, x^* *is the projection of any stochastic discount factor* m *on the space* $\underline{X}$ *of payoffs.* This is a very important fact: *the pricing implications of any model of* m *for a set of payoffs* $\underline{X}$ *are the same as those of the projection of* m *on* $\underline{X}$, or of the *mimicking portfolios* of m. Algebraically,

$$p = E(mx) = E\left[(proj(m|\underline{X}) + \varepsilon)x\right] = E\left[proj(m|\underline{X})\ x\right]$$

Let me repeat and emphasize the logic. Above, we started with investors or a contingent claim market, and derived a discount factor. $p = E(mx)$ *implies* the linearity of the pricing function and hence the law of one price, a pretty obvious statement in those contexts. Here we work backwards. Markets are *in*complete in that contingent claims to lots of states of nature are not available. We *do* allow arbitrary portfolio formation, and that sort of “completeness” is important to the result. If consumers cannot form a portfolio $ax + by$, they cannot force the price of this portfolio to equal the price of its constituents.. We found that the law of one price implies a linear pricing function, and a linear pricing function implies that there exists at least one and usually many discount factors. The law of one price is not innocuous; it is an

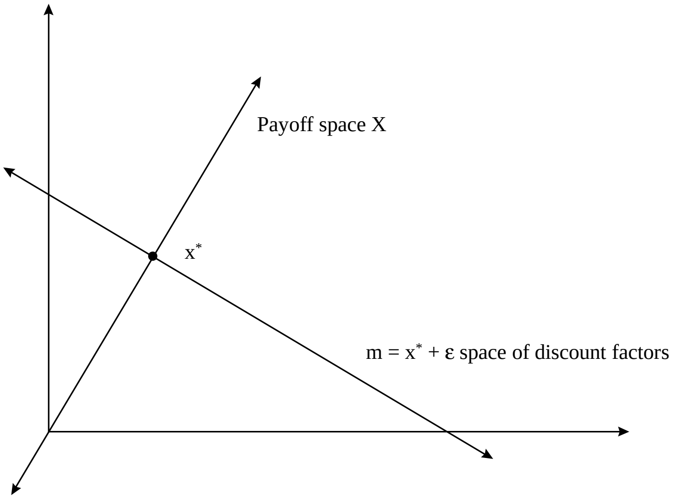


Figure 7. Many discount facotors m can price a given set of assets in incomplete markets.

assumption about preferences albeit a weak one. The point of the theorem is that this is *just* enough information about preferences to deduce a discount factor.

4.6 No-Arbitrage and positive discount factors

The definition of arbitrage.

There is a *strictly positive* discount factor m such that $p = E(mx)$ if and only if there are *no arbitrage opportunities*.

Next, another *implication* of marginal utility that holds for a wide class of preferences, that can be reversed to deduce properties of discount factors. Start with the following:

Definition (Absence of arbitrage). A payoff space $\underline{X}$ and pricing function $p(x)$ leave no *arbitrage opportunities* if every payoff x that is always non-negative, $x \geq 0$ (almost surely), and sometimes strictly positive, $x > 0$ with some positive probability, has positive price, $p(x) > 0$.

No-arbitrage says that you can't get for free a portfolio that *might* pay off positively, but will certainly never cost you anything. This definition is different from the colloquial use of the word "arbitrage." Most people use "arbitrage" to mean a violation of the law of one price – a riskless way of buying something cheap and selling it for a higher price. "Arbitrages" here might pay off, but then again they might not. The word "arbitrage" is also widely abused. "Risk arbitrage" is an oxymoron that means taking bets.

An equivalent statement is that if one payoff *dominates* another – if $x \geq y$ – then $p(x) \geq p(y)$ (Or, a bit more carefully but more long-windedly, if $x \geq y$ almost surely and $x > y$ with positive probability, then $p(x) > p(y)$.)

No-arbitrage as a consequence of utility maximization

The absence of arbitrage opportunities is clearly a *consequence* of utility maximization. Recall,

$$m(s) = \delta \frac{u'[c(s)]}{u'(c)} > 0.$$

It is a sensible characterization of a wide class of preferences that marginal utility is always positive. Few people are so satiated that they will throw away money. Therefore, *the marginal rate of substitution is positive.* Watch out: the marginal rate of substitution is a random variable, so "positive" means "positive in every state of nature" or "in every possible realization."

Now, since each contingent claim price is positive, a bundle of positive amounts of contingent claims must also have a positive price, even in incomplete markets. A little more formally,

Theorem: $p = E(mx)$ and $m(s) > 0$ imply no-arbitrage.
Proof: $p(x) = \sum_s \pi(s)m(s)x(s)$. If there is a payoff with $x(s) \geq 0$, and $x(s) > 0$ with positive probability, then the right hand side is positive. □
Similarly, if $m > 0$ then $x \geq y$ implies $p(x) = E(mx) \geq p(y) = E(my)$.

The theorem

Now we turn the observation around. As the LOOP property guaranteed the existence of a discount factor m, no-arbitrage guarantees the existence of a positive m.

The basic idea is pretty simple. No-arbitrage means that the prices of any payoff in the positive orthant (except zero, but including the axes) must be strictly positive. Thus the iso-price lines must march up and to the right, and the discount factor m, perpendicular to the iso-price lines, must point up and to the right. Figure 8 gives an illustration of the case that is ruled out: the payoff x is strictly positive, but has a negative price. As a result, the (unique, since this market is complete) m is negative in the y-axis state.

To prove the theorem a little more formally, start in complete markets. There is only one $m,\ x^*$. If it isn't positive in some state, then the contingent claim in that state has a positive payoff and a negative price, which violates no arbitrage. More formally,

Theorem: In complete markets, no-arbitrage implies that there exists a unique $m > 0$ such that $p = E(mx)$.
Proof: No-arbitrage implies the law of one price, so there is a unique x^* such that $p = E(x^*x)$. Suppose that $x^* \leq 0$ for some states. Then, form a payoff x that is 1 in those states, and zero elsewhere. This payoff is strictly positive, but its price, $\sum_{s:x^*(s)<0} \pi(s)x^*(s)$ is negative, negating the assumption of no-arbitrage. □

Next, what if markets are incomplete? There are now many m's that price assets. Any m of the form $m = x^* + \epsilon$, with $E(\epsilon x) = 0$ will do. We want to show that at least *one* of these is positive. But that one may not be x^*. Since the other m's are not in the payoff space $\underline{X}$, the construction given above may yield a payoff that is not in $\underline{X}$, and hence to which we can't assign a price. To handle this case, I adopt a different strategy of proof. My proximate source is Duffie (1992), the original proof is due to Ross (1978). The proof is not particularly intuitive.

Theorem: No arbitrage implies the existence of an $m > 0$ s. t. $p = E(mx)$.
Proof: Join $(-p(x), x)$ together to form vectors in $\mathcal{R}^{S+1}$. Call M the set of all $(-p(x),\ x)$ pairs,

$$M = \{(-p(x), x);\ x \in \underline{X}\}$$

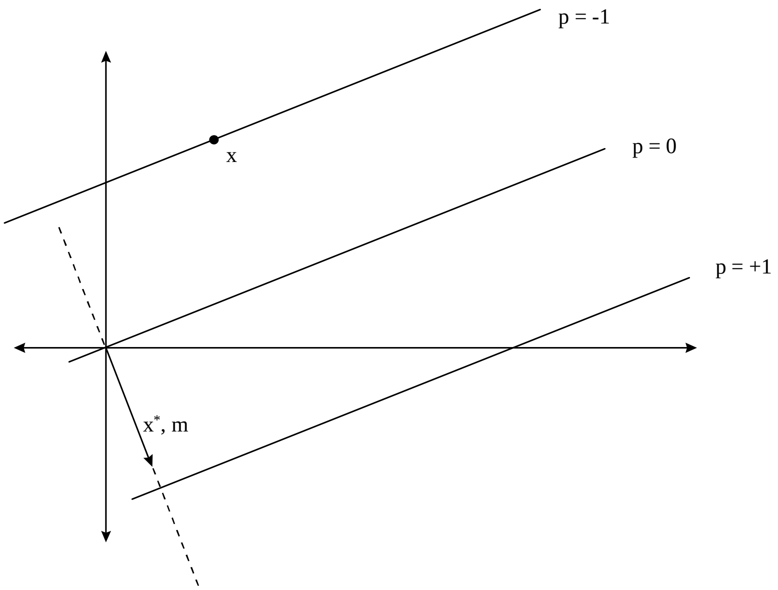


Figure 8. Counter-example for no-arbitrage. The payoff x is positive, but has negative price. The discount factor is not strictly positive

M is still a linear space: $m_1 \in M,\ m_2 \in M \Rightarrow am_1 + bm_2 \in M$. No-arbitrage means that elements of m can't have all positive elements. If x is positive, $-p(x)$ must be negative. Thus, M only intersects the positive orthant $\mathcal{R}_+^{S+1}$ at the point 0.
M and $\mathcal{R}_+^{S+1}$ are thus convex sets that intersect at one point, 0. By the *separating hyperplane* theorem, there is a linear function that separates the two convex sets; there is an $F : \mathcal{R}^{S+1} \Rightarrow \mathcal{R}$, such that $F(-p, x) = 0$ for $(-p, x) \in M$, and $F(-p, x) > 0$ for $(-p, x) \in \mathcal{R}_+^{S+1}$ except the origin. By the *Riesz representation theorem* we can represent F as an inner product with some vector m, by $F(-p, x) = -p + m \cdot x$, or $-p + E(mx)$ using the second moment inner product. Finally, since $F(-p, x)$ must be positive for $(-p, x) > 0$, m must be positive. □

What the theorem says and does not say

The theorem says that a discount factor $m > 0$ exists, but it does *not* say that $m > 0$ is *unique*. The left hand panel of Figure 9 illustrates. Any m on the line through x^* orthogonal to $\underline{X}$ also prices assets. Again, $p = E[(m+\varepsilon)x]$ if $E(\varepsilon x) = 0$. Any of these discount factors in the positive orthant are positive, and thus satisfy the theorem. There are lots of them!

The theorem says that a positive m exists, but it also does *not* say that *every* discount factor m must be positive. The discount factors in the left hand panel of Figure 9 *outside* the positive orthant are perfectly valid – they satisfy $p = E(mx)$, and the prices they generate on $\underline{X}$ are arbitrage free, but they are not positive in every state of nature. In particular, the discount factor x^* in the payoff space is still perfectly valid — $p(x) = E(x^*x)$ — but it need not be positive, again as illustrated in the left hand panel of Figure 9.

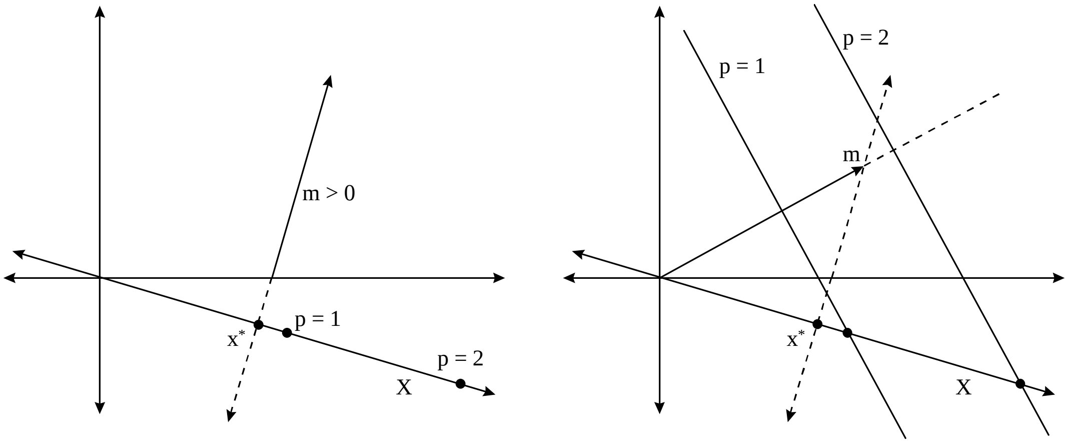


Figure 9. Existence of a discount factor and extensions. The left graph shows that the postive discount factor is not unique, and that discount factors may also exist that are not strictly positive. In particular, x^* need not be positive. The right hand graph shows that each particular choice of $m > 0$ induces an *arbitrage free extension* of the prices on X to all contingent claims.

Another interpretation: This theorem shows that we can *extend* the pricing function defined on $\underline{X}$ to all possible payoffs $\mathcal{R}^S$, and not imply any arbitrage opportunities on that larger space of payoffs. It says that there is a pricing function $p(x)$ defined over *all* of $\mathcal{R}^S$, that assigns the same (correct, or observed) prices on $\underline{X}$ and that displays no arbitrage on all of $\mathcal{R}^S$. Graphically, it just says we can draw parallel planes to represent prices on all of $\mathcal{R}^S$ in such a way that the planes intersect $\underline{X}$ in the right places and march up and to the right so the positive orthant always has positive prices. In fact, there are many ways to do this. Any positive m generates such a no-arbitrage extension, as illustrated in the right hand panel of Figure 9. As $m > 0$ exists but is not unique, so the extension it generates is not unique.

We can think of strictly positive discount factors as possible contingent claims prices. We can think of the theorem as answering the question: is it possible that an observed and incomplete set of prices and payoffs is generated by some complete markets, contingent claim economy? The answer is, yes, if there is no arbitrage on the observed prices and payoffs. In fact, since there are typically many positive m's consistent with an $\{\underline{X},\ p(x)\}$, there exist many contingent claims economies consistent with our observations.

Finally, the absence of arbitrage is another very weak characterization of preferences. The theorem tells us that this is enough to allow us to use the $p = E(mx)$ formalism with $m > 0$.

As usual, this theorem and proof do not require that the state space is $\mathcal{R}^S$. State spaces generated by continuous random variables work just as well.

4.7 Existence in continuous time

Just like x^* in discrete time,

$$\frac{d\Lambda^*}{\Lambda^*} = -r_t dt - \left(\boldsymbol{\mu}_t + \frac{\mathbf{D}}{\mathbf{p}} - r_t\right)' \Sigma_t^{-1} d\mathbf{z}.$$

prices assets by construction in continuous time.

The law of one price implies the existence of a discount factor process, and absence of arbitrage a positive discount factor process in continuous time as well as discrete time.

At one level, this statement requires no new mathematics. If we reinvest dividends for simplicity, then a discount factor must satisfy

$$p_t \Lambda_t = E_t \Lambda_{t+s} p_{t+s}.$$

Calling $p_{t+s} = x_{t+s}$, this *is* precisely the discrete time $p = E(mx)$ that we have studied all along. Thus, the law of one price and absence of arbitrage are equivalent to the existence of a or a positive Λ_{t+s}; the same conditions at all horizons s are thus equivalent to the existence of a or a positive discount factor process Λ_t for all time t.

For calculations it is useful to find explicit formulas for a discount factors, the analogue to the discrete time discount factor $x^* = \mathbf{p}' E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}$. Suppose a set of securities pays dividends

$$\mathbf{D}_t dt$$

and their prices follow

$$\frac{d\mathbf{p}_t}{\mathbf{p}_t} = \boldsymbol{\mu}_t dt + \Sigma_t d\mathbf{z}$$

where $\mathbf{p}$ and $\mathbf{z}$ are $N \times 1$ vectors, $\boldsymbol{\mu}$ and Σ may vary, $\boldsymbol{\mu}(\mathbf{p}_t, t,$other variables$)$, $\Sigma(\mathbf{p_t}, t,$other variables$)$ is full rank, $E\left(d\mathbf{z}d\mathbf{z}'\right) = I$ and the division on the left hand side is element-by element.

We can form a discount factor that prices these assets from a linear combination of the shocks that drive the original assets,

$$\frac{d\Lambda^*}{\Lambda^*} = -r_t^f dt - \left(\boldsymbol{\mu}_t + \frac{\mathbf{D}}{\mathbf{p}} - r_t^f\right)' \Sigma_t^{-1} d\mathbf{z}. \tag{34}$$

If there is a risk free rate r_t^f (also potentially time-varying), then that rate determines r_t^f in the above equation. If there is no risk free rate, the above discount factor will price the risky assets for any arbitrary (or convenient) choice of r_t^f. As usual, this discount factor is not unique; Λ^* plus orthogonal noise will also act as a discount factor:

$$\frac{d\Lambda}{\Lambda} = \frac{d\Lambda^*}{\Lambda^*} + dw;\ E(dw) = 0;\ \ E(d\mathbf{z}dw) = 0$$

We can easily check that (34) does in fact price the basis assets. Writing the basic pricing equation (20) in continuous time,

$$\begin{aligned} & \frac{\mathbf{D}}{\mathbf{p}}dt + E_t\left(\frac{d\Lambda^*}{\Lambda^*} + \frac{d\mathbf{p}}{\mathbf{p}} + \frac{d\Lambda^*}{\Lambda^*}\frac{d\mathbf{p}}{\mathbf{p}}\right) \\ = \quad & \frac{\mathbf{D}}{\mathbf{p}}dt - r_t^f dt + \boldsymbol{\mu}_t dt - \left(\boldsymbol{\mu}_t + \frac{\mathbf{D}}{\mathbf{p}} - r_t^f\right)' \Sigma_t^{-1}\Sigma_t dt = 0. \end{aligned}$$

In the continuous time case it is easier to write the discount factor in terms of the *covariance* matrix of the original securities, where we wrote it in terms of the second moment matrix in discrete time. There is nothing essential in this difference. We could have written the discrete-time x^* in terms of the covariance matrix of a set of returns just as well: you can check that

$$x^* = \frac{1}{R^f} - \frac{E(\mathbf{R}) - R^f}{R^f}\Sigma^{-1}\left[\mathbf{R} - E(R)\right];\ \ \Sigma \equiv cov(\mathbf{RR}')$$

satisfies $\mathbf{1} = E(x^*\mathbf{R})$ by construction, and this formula is obviously closely analogous to the continuous time formula.

Chapter 5. Mean-variance frontier and beta representations

Much empirical work in asset pricing is couched in terms of expected return - beta representations and mean-variance frontiers. In this chapter, I draw the connection between the discount factor view and these more traditional views of asset pricing. In the process, I introduce a number of useful tools and representations. In the first chapter, I showed how mean-variance and a beta representation follow from $p = E(mx)$ and (in the mean-variance case) complete markets. Here, I take a closer look at the representations and I draw the connections in incomplete markets. I start by defining the terminology of beta pricing models and mean variance frontiers.

5.1 Expected return - Beta representations

An expected return-beta model is,

$$E(R^i) = \alpha + \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \dots$$

α equals the risk free or zero beta rate.

When the factors are returns, $f^a = R^a$ then $\lambda_a = E(R^a) - \alpha$, and factor pricing is equivalent to a restriction on intercepts in time-series regressions.

When the factors are not returns, we can reexpress the beta pricing model in terms of factor mimicking portfolios that are returns.

Much empirical work in finance is cast in terms of beta representations for expected returns. A beta model is a characterization of expected returns across assets of the form

$$E(R^i) = \alpha + \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \dots,\ i = 1, 2, ...N. \tag{35}$$

α, λ are constant for all assets and $\beta_{i,a}$ is the multiple regression coefficient of return i on factor a. This qualification is important: if the betas are arbitrary numbers or characteristics of the securities there is no content to the equation! $\beta_{i,a}$ is interpreted as the amount of exposure of asset i to factor a risks, and λ_a is interpreted as the price of such risk-exposure. Read the beta pricing model to say: "for each unit of exposure β to risk factor a, you must provide investors with an expected return premium λ_a."

If there is a risk free rate, its betas are all zero[1], so the intercept is equal to the risk free

[1] The betas are zero because the risk free rate is known ahead of time. When we consider the effects of conditioning information, i.e. that the interest rate could vary over time, we have to interpret the means and betas

rate,

$$R^f = \alpha$$

If not, then α is the expected rate of return on a portfolio of risky assets with zero betas on all factors. α is called the (expected) *zero-beta rate* in this circumstance.

Beta pricing models are constructed to explain the variation in average returns across assets. I write $i = 1, 2, ...N$ to emphasize this fact. For example, equation (35) says that if we plot expected returns versus betas in a one-factor model, we should expect all $(E(R^i),\ \beta_{i,a})$ pairs to line up on a straight line with slope λ_a. A low price is equivalent to a high expected return, so this is "asset pricing" by any other name. One way to test (35) is to run a *cross sectional regression* of average returns on betas,

$$E(R^i) = \alpha + \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \ldots + \alpha_i,\ i = 1, 2, ...N.$$

This regression is tricky because of the nonstandard notation: the β_i are the right hand variables, the λ are the slope coefficients, and the α_i are *pricing errors*. The model predicts $\alpha_i = 0$, and they should be statistically insignificant in a test.

The "factors" are proxies for marginal utility growth. I discuss the stories used to select factors at some length below. For the moment keep in mind the canonical examples of risk factors, $f =$ consumption growth or the return on the market portfolio of all assets.

We can write the multiple regressions that define the betas as

$$R^i_t = a_i + \beta_{i,a} f^a_t + \beta_{i,b} f^b_t + \ldots + \varepsilon^i_t.\ t = 1, 2, ...T \tag{36}$$

This is often called a *time-series regression*, to distinguish it from the cross-sectional regression of average returns on betas. Notice that we run returns R^i_t on contemporaneous factors f_t. This regression is not about predicting returns from variables seen ahead of time. Its objective is to measure contemporaneous relations; risk exposure; whether returns are typically high in "good times" or "bad times" and thus whether the asset is useful to smooth risks.

Rather than estimate a zero-beta rate, one often examines a factor pricing model using excess returns. Differencing (35) between any two returns $R^{ei} = R^i - R^j$ (R^j may but does not have to be risk free), we obtain

$$E(R^{ei}) = \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \ldots,\ i = 1, 2, ...N. \tag{37}$$

Here, β_{ia} represents the regression coefficient of the excess return R^{ei} on the factors. This subtraction is more than a convenience: it allows us to focus on the central task of understanding risk premia and risk corrections.

as *conditional* moments. Thus, if you are worried about time-varying risk free rates, betas, and so forth, either assume all variables are i.i.d. (and thus the risk free rate is constant), or interpret all moments as conditional on time *t* information. We incorporate conditioning information explicitly in the next chapter.

Finally, we often express factor risk premia λ in terms of portfolio returns. If the factors already are excess returns, $f^a = R^{ea}$, this is easy. Otherwise, find an excess return R^{ea} with beta of 1 on one factor and beta zero on the rest. In either case, the factor model (37) applied to the excess return R^{ea} implies $\lambda_a = E(R^{ea})$ so we can write the factor model as

$$E(R^{ei}) = \beta_{i,a}E(R^{ea}) + \beta_{i,b}E(R^{eb}) + \dots\, ,\ i = 1, 2, ...N.$$

The beta pricing model (35)-(37) and the regression definition of the betas in (36) look very similar. It seems like one can take expectations of the time-series regression (36) and arrive at the beta model (35), in which case the latter would be vacuous since one can always run a regression of anything on anything. Yet the beta model and regression are distinct equations and capture very different ideas. The difference is subtle but crucial: the time-series regressions (36) we will in general have a different intercept a_i for each return i, while the intercept α is the same for all assets in the beta pricing equation (35). The beta pricing equation is a restriction on expected returns, and thus imposes a restriction on intercepts in the time-series regression. In the special case that the factors are themselves excess returns, the restriction is particularly simple: the intercepts should all be zero.

5.2 Mean-variance frontiers

The *mean-variance frontier* of a given set of assets is the boundary of the set of means and variances of the returns on all portfolios of the given assets. One can find or define this boundary by minimizing return variance for a given mean return. Many asset pricing propositions and test statistics have interpretations in terms of the mean-variance frontier.

Figure 10 displays a typical mean-variance frontier. As displayed in Figure 10, it is common to distinguish the mean-variance frontier of all risky assets, graphed as the hyperbolic region, and the mean-variance frontier of all assets, i.e. including a risk free rate if there is one, which is the larger wedge-shaped region. Some authors reserve the terminology "mean-variance frontier" for the upper portion, calling the whole thing the *minimum variance frontier*. The risky asset frontier is a hyperbola, which means it lies between two asymptotes, shown as dotted lines. The risk free rate is typically drawn below the intersection of the asymptotes and the vertical axis, or the point of minimum variance on the risky frontier. If it were above this point, investors with a mean-variance objective would try to short the risky assets.

In Chapter 1, we derived a similar wedge-shaped region as the set of means and variances of all assets that are priced by a given discount factor. This chapter is about incomplete markets, so we think of a mean-variance frontier generated by a given set of assets, typically less than complete.

When does the mean-variance frontier exist? I.e., when is the set of portfolio means and variances less than the whole $\{E, \sigma\}$ space? We basically have to rule out a special case: two returns are perfectly correlated but yield different means. In that case one could short one, long the other, and achieve infinite expected returns with no risk. More formally, eliminate purely redundant securities from consideration, then

Theorem: So long as the variance-covariance matrix of returns is non-singular, there is a mean-variance frontier.

To prove this theorem, just follow the construction below. This theorem should sound very familiar: Two perfectly correlated returns with different mean are a violation of the law of one price. Thus, the law of one price implies that there is a mean variance frontier as well as a discount factor.

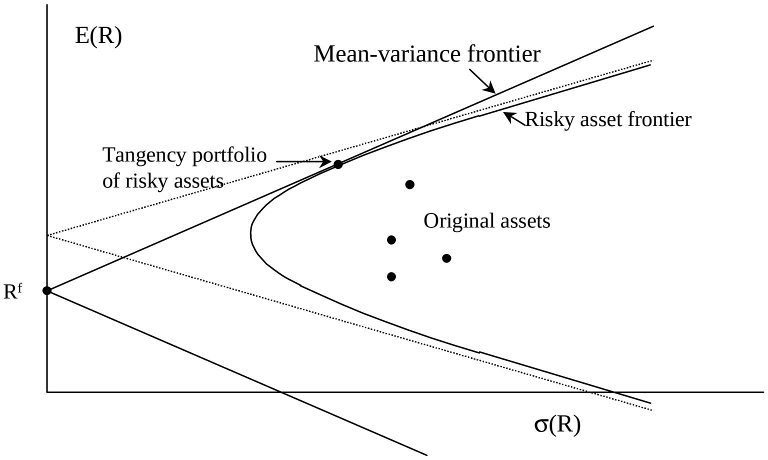


Figure 10. Mean-variance frontier

5.2.1 Lagrangian approach to mean-variance frontier

The standard definition and the computation of the mean-variance frontier follows a brute force approach.

Problem: Start with a vector of asset returns $\mathbf{R}$. Denote by $\mathbf{E}$ the vector of mean returns, $\mathbf{E} \equiv E(\mathbf{R})$, and denote by Σ the variance-covariance matrix $\Sigma = [E(\mathbf{R} - \mathbf{E})(\mathbf{R} - \mathbf{E})']$. A portfolio is defined by its weights $\mathbf{w}$ on the initial securities. The portfolio return is $\mathbf{w}'\mathbf{R}$ where the weights sum to one, $\mathbf{w}'\mathbf{1} = 1$.The problem "choose a portfolio to minimize variance for a given mean" is then

$$min_{\{\mathbf{w}\}} \ \mathbf{w}'\Sigma\mathbf{w} \text{ s.t. } \mathbf{w}'\mathbf{E} = \mu; \ \mathbf{w}'\mathbf{1} = 1. \tag{38}$$

Solution: Let

$$A = \mathbf{E}'\Sigma^{-1}\mathbf{E};\;\; B = \mathbf{E}'\Sigma^{-1}\mathbf{1};\;\; C = \mathbf{1}'\Sigma^{-1}\mathbf{1}.$$

Then, for a given mean portfolio return μ, the minimum variance portfolio has variance

$$var\left(R^p\right) = \frac{C\mu^2 - 2B\mu + A}{AC - B^2} \tag{39}$$

and is formed by portfolio weights

$$\mathbf{w} = \Sigma^{-1}\frac{\mathbf{E}\left(C\mu - B\right) + \mathbf{1}\left(A - B\mu\right)}{\left(AC - B^2\right)}.$$

Equation (39) shows that the variance is a quadratic function of the mean. The square root of a parabola is a hyperbola, which is why we draw hyperbolic regions in mean-standard deviation space.

The *minimum-variance portfolio* is interesting in its own right and appears as a special case in many theorems and it appears in several test statistics. We can find it by minimizing (39) over μ, giving $\mu^{\text{min var}} = B/C$. The weights of the minimum variance portfolio are thus

$$\mathbf{w} = \Sigma^{-1}\mathbf{1}/(\mathbf{1}'\Sigma^{-1}\mathbf{1}).$$

We can get to any point on the mean-variance frontier by starting with two returns on the frontier and forming portfolios. The frontier is *spanned* by any two frontier returns. To see this fact, notice that $\mathbf{w}$ is a linear function of μ. Thus, if you take the portfolios corresponding to any two distinct mean returns μ_1 and μ_2, the weights on a third portfolio with mean $\mu_3 = \lambda\mu_1 + (1-\lambda)\mu_2$ are given by $\mathbf{w}_3 = \lambda\mathbf{w}_1 + (1-\lambda)\mathbf{w}_2$.

Derivation: To derive the solution, introduce Lagrange multipliers 2λ and 2δ on the constraints. The first order conditions to (38) are then

$$\Sigma\mathbf{w} - \lambda\mathbf{E} - \boldsymbol{\delta}\mathbf{1} = 0$$

$$\mathbf{w} = \Sigma^{-1}(\lambda\mathbf{E} + \delta\mathbf{1}). \tag{40}$$

We find the Lagrange multipliers from the constraints,

$$\mathbf{E}'\mathbf{w} = \mathbf{E}'\Sigma^{-1}(\lambda\mathbf{E} + \delta\mathbf{1}) = \mu$$

$$\mathbf{1}'\mathbf{w} = \mathbf{1}'\Sigma^{-1}(\lambda\mathbf{E} + \delta\mathbf{1}) = 1$$

or

$$\left[\begin{array}{cc} \mathbf{E}'\Sigma^{-1}\mathbf{E} & \mathbf{E}'\Sigma^{-1}\mathbf{1} \\ \mathbf{1}'\Sigma^{-1}\mathbf{E} & \mathbf{1}'\Sigma^{-1}\mathbf{1} \end{array}\right]\left[\begin{array}{c} \lambda \\ \delta \end{array}\right] = \left[\begin{array}{c} \mu \\ 1 \end{array}\right]$$

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \lambda \\ \delta \end{bmatrix} = \begin{bmatrix} \mu \\ 1 \end{bmatrix}$$

Hence,

$$\lambda = \frac{C\mu - B}{AC - B^2}$$

$$\delta = \frac{A - B\mu}{AC - B^2}$$

Plugging in to (40), we get the portfolio weights and variance.

5.2.2 Characterizing the mean-variance frontier

Every return can be expressed as $R^i = R^* + w^i R^{e*} + n^i$.

The mean-variance frontier is $R^{mv} = R^* + wR^{e*}$

R^{e*} is defined as $R^{e*} = proj(1|\underline{R^e})$. It represents mean excess returns, $E(R^e) = E(R^{e*}R^e)$ $\forall R^e \in \underline{R^e}$

The Lagrangian approach to the mean-variance frontier is straightforward but cumbersome. Our further manipulations will be easier if we follow an alternative approach due to Hansen and Richard (1987). Technically, Hansen and Richard's approach is also valid in infinite-dimensional payoff spaces, which we will not be able to avoid when we include conditioning information. Also, it is the natural geometric way to think about the mean-variance frontier given that we have started to think of payoffs, discount factors and other random variables as vectors with a second moment norm.

Definitions of R^*, R^{e*}.

I start by defining two special returns. R^* is the return corresponding to the payoff x^* that can act as the discount factor. The price of x^*, is, like any other price, $p(x^*) = E(x^*x^*)$. Thus,

The definition of R^* is

$$R^* \equiv \frac{x^*}{E(x^{*2})} \tag{41}$$

The definition of $R^{e*}$is

$$R^{e*} \equiv proj(1 \mid \underline{R}^e) \tag{42}$$

$$\underline{R^e} \equiv \text{space of excess returns} = \{x \in \underline{X} \; s.t. \; p(x) = 0\}$$

R^* and R^{e*} have many interesting properties, which I discuss below. Now we can state a beautiful orthogonal decomposition.

Theorem: Every return R^i can be expressed as

$$R^i = R^* + w^i R^{e*} + n^i$$

where w^i is a number, and n^i is an excess return with the property

$$E(n^i) = 0.$$

The three components are orthogonal,

$$E(R^* R^{e*}) = E(R^* n^i) = E(R^{e*} n^i) = 0.$$

This theorem quickly implies

Theorem: R^{mv} is on the mean-variance frontier if and only if

$$R^{mv} = R^* + wR^{e*}$$

for some real number w.

As usual, first I'll argue why the theorems are sensible, then I'll offer a fairly loose algebraic proof. Hansen and Richard (1987) give a much more careful proof.

Graphical construction

Figure 11 illustrates the decomposition. Start at the origin (0). Recall that the x^* vector is orthogonal to planes of constant price; thus the R^* vector lies at right angles to the plane of returns as shown. Go to R^*. R^{e*} is the excess return that is closest to the vector 1; it is orthogonal to planes of constant *mean* return, shown in the $E = 1, E = 2$ lines, just as the return R^* is orthogonal to all excess returns. Proceed an amount w^i in the direction of R^{e*}, getting as close to R^i as possible. Now move, again in an orthogonal direction, by an amount n^i to get to the return R^i. We have thus expressed $R^i = R^* + w^i R^{e*} + n^i$ in a way that all three components are orthogonal.

Returns with $n = 0$, $R^* + wR^{e*}$, are the mean-variance frontier. Here's why. Since $E(R^2) = \sigma^2(R) + E(R)^2$, we can define the mean-variance frontier by minimizing second moment for a given mean. The length of each vector in Figure 11 is its second moment, so we want the shortest vector that is on the return plane for a given mean. The shortest vectors in the return plane with given mean or value of E in the picture are on the $R^* + wR^{e*}$ line.

R^{e*} is an excess return in $\underline{R^e}$ that represents *means* on $\underline{R^e}$ with an inner product in just

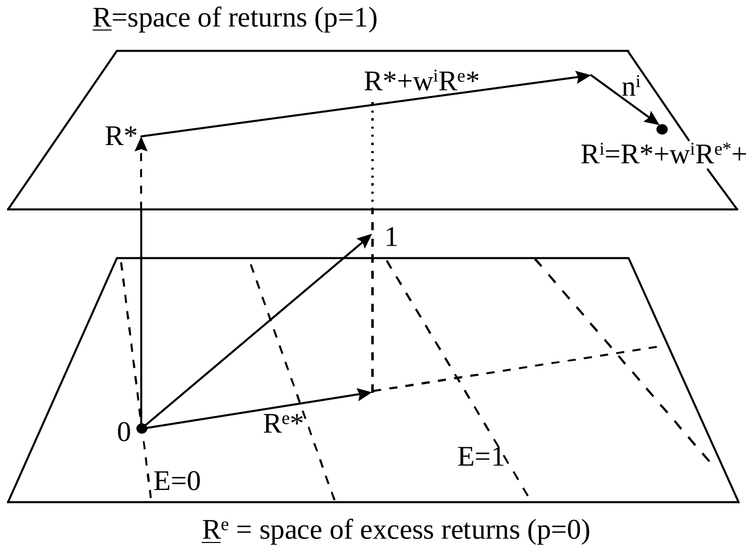


Figure 11. Orthogonal decomposition and mean-variance frontier.

the same way that x^* is a portfolio in $\underline{X}$ that represents prices on $\underline{X}$.

$$E(R^e) = E(R^{e*}R^e) \;\; \forall R^e \in \underline{R^e}$$

To see this fact algebraically,

$$E(R^e) = E(1 \times R^e) = E\left[proj(1 \mid R^e) \times R^e\right] = E(R^{e*}R^e).$$

Here's the idea intuitively. Expectation is the inner product with 1. Planes of constant expected value in Figure 11 are orthogonal to the 1 vector, just as planes of constant price are orthogonal to the x^* or R^* vectors. I don't show these planes for clarity; I do show lines of constant expected return in $\underline{R^e}$, which are the intersection of planes of constant expected payoff with the $\underline{R^e}$ plane. Therefore, just as we found an x^* *in* $\underline{X}$ to represent prices in $\underline{X}$ by projecting m onto $\underline{X}$, we find R^{e*} in $\underline{R^e}$ by projecting of 1 onto $\underline{R^e}$. Yes, a regression with one on the left hand side. Planes perpendicular to R^{e*} in $\underline{R^e}$ (and in $\underline{R}$) are payoffs with constant *mean*, just as planes perpendicular to x^* in $\underline{X}$ generate payoffs with the same *price*.

Algebraic argument

Now, on to an algebraic proof of the decomposition and characterization of mean variance frontier.

Proof: Straight from their definitions, (41) and (42) we know that R^e is an excess return (price zero), and that R^* and R^e are orthogonal,

$$E(R^*R^e) = \frac{E(x^*R^e)}{E(x^{*2})} = 0.$$

We define n^i so that the decomposition adds up to R^i as claimed – n^i is what is left over – and we define w^i to make sure that n^i is orthogonal to the other two components. Then we prove that $E(n^i) = 0$ for the mean-variance frontier. Define

$$n^i \equiv R^i - R^* - w^i R^{e*}.$$

For any w^i, n^i is an excess return so already orthogonal to R^*,

$$E(R^* n^i) = 0.$$

To show $E(n^i) = 0$ and n^i orthogonal to R^{e*}, we exploit the fact that since n^i is an excess return,

$$E(n^i) = E(R^{e*} n^i).$$

Therefore, R^{e*} is orthogonal to n^i if and only if we pick w^i so that $E(n^i) = 0$. We don't have to explicitly calculate w^i for the proof[2].
Once we have constructed the decomposition, the frontier drops out. Since $E(n^i) = 0$ and the three components are orthogonal,

$$E(R^i) = E(R^*) + w^i E(R^{e*})$$

$$\sigma^2(R^i) = \sigma^2(R^* + w^i R^{e*}) + \sigma^2(n^i).$$

Thus, for each desired value of the mean return, there is a unique w^i. Returns with $n^i = 0$ minimize variance for each mean. ■

Decomposition in mean-variance space

Figure 12 illustrates the decomposition in mean-variance space rather than in state-space.

First, let's locate R^*. R^* is the minimum second moment return. One can see this fact from the geometry of Figure 11: R^* is the return closest to the origin, and thus the return with the smallest "length" which is second moment. As with OLS regression, minimizing the length of R^* and creating an R^* orthogonal to all excess returns is the same thing. One can also verify this property algebraically. Since any return can be expressed as $R = R^* +$

[2] Its value

$$w^i = \frac{E(R^i) - E(R^*)}{E(R^{e*})}$$

is not particularly enlightening.

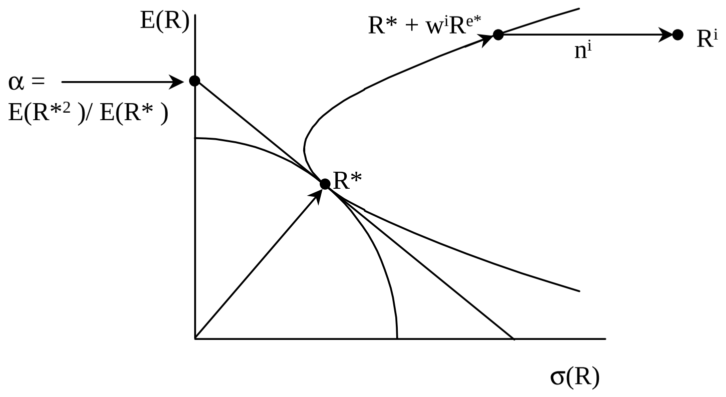


Figure 12. Orthogonal decomposition of a return R^i in mean-standard deviation space.

$wR^{e*} + n$, $E(R^2) = E(R^{*2}) + w^2 E(R^{e*2}) + E(n^2)$. $n = 0$ and $w = 0$ thus give the minimum second moment return.

In mean-standard deviation space, lines of constant second moment are circles. Thus, the minimum second-moment return R^* is on the smallest circle that intersects the set of all assets, or the mean-variance frontier as in the right hand panel of Figure 14. Notice that R^* is on the lower, or "inefficient" segment of the mean-variance frontier. It is initially surprising that this is the location of the most interesting return on the frontier! R^* is *not* the "market portfolio" or "wealth portfolio."

R^{e*} moves one along the frontier. Adding n does not change mean but does change variance, so it is an *idiosyncratic* return that just moves an asset off the frontier as graphed. I'll return to the vertical intercept α below.

A compilation of properties of R^*, R^{e*} and x^*

There are lots of interesting and useful properties of the special returns that generate the mean variance frontier. I list a few here. Some I derived above, some I will derive and discuss below in more detail, and some will be useful tricks later on. In every case, I urge you to draw a little picture to go along with the algebraic discussion.

1)

$$E(R^{*2}) = \frac{1}{E(x^{*2})}. \tag{43}$$

To derive this fact, multiply both sides of (41) by R^*, take expectations, and remember R^* is a return so $1 = E(x^* R^*)$.

2) We can reverse the definition and recover x^* from R^* via

$$x^* = \frac{R^*}{E(R^{*2})}. \tag{44}$$

To derive this formula, start with the definition $R^* = x^*/E(x^{*2})$ and substitute from (43) for $E(x^{*2})$3) R^* can be used to represent prices just like x^*. This is not surprising, since they both point in the same direction, orthogonal to planes of constant price. Algebraically,

$$E(R^{*2}) = E(R^* R) \; \forall R \in \underline{R}. \tag{45}$$

$$E(R^{*2})p(x) = E(R^* x) \; \forall x \in \underline{X}.$$

This fact can also serve as an alternative defining property of R^*. To derive (45), use $1 = E(x^* R)$ and (44).

3) R^* is the minimum second moment return.

4) R^{e*} represents means on $\underline{R^e}$ via an inner product in the same way that x^* represents prices on $\underline{X}$ via an inner product. R^{e*} is orthogonal to planes of constant mean in $\underline{R^e}$ as x^* is orthogonal to planes of constant price. Algebraically, in analogy to $p(x) = E(x^* x)$ we have

$$E(R^e) = E(R^{e*} R^e) \; \forall R^e \in \underline{R^e}. \tag{46}$$

This fact can serve as an alternative defining property of R^{e*}.

To see this fact, recall that R^{e*} is defined by

$$R^{e*} \equiv proj(1|\underline{R^e})$$

which is analogous to $x^* = proj(m|\underline{X})$. Therefore,

$$E(R^e) = E(1 \times R^e) = E\left(proj(1|\underline{R^e}) \times R^e\right) = E(R^{e*} R^e).$$

5) R^{e*} and R^* are orthogonal,

$$E(R^* R^{e*}) = 0.$$

6) The mean variance frontier is given by

$$R^{mv} = R^* + w R^{e*}.$$

7) Since R^* and R^{e*} are orthogonal, the last fact implies that

$$R^* = \text{ minimum second moment return.}$$

Graphically, R^* is the return closest to the origin. I discuss this property at some length in section 6 below.

The remaining properties are minor; I use them once or twice below and they make great test questions, but are not that deep.

8) Applying fact (46) to R^{e*} itself, R^{e*} has the same first and second moment,

$$E(R^{e*}) = E(R^{e*2})$$

and therefore

$$var(R^{e*}) = E(R^{e*2}) - E(R^{e*})^2 = E(R^{e*})\left[1 - E(R^{e*})\right].$$

9) If there is a riskfree rate, then R^{e*} can also be defined as the residual in the projection of 1 on R^* :

$$R^{e*} = 1 - proj(1|R^*) = 1 - \frac{E(R^*)}{E(R^{*2})}R^* = 1 - \frac{1}{R^f}R^* \tag{47}$$

See Figure 11! To verify this statement analytically, check that R^{e*} so defined is an excess return in $\underline{X}$, and $E(R^{e*}R^e) = E(R^e);\ E(R^*R^{e*}) = 0.$

Riskfree return, zero beta return, and minimum variance returns

The riskfree rate is an obviously interesting point on the mean variance frontier, and it should be no surprise that it will show up often in asset pricing formulas. Thus, it's interesting to characterize it by finding the appropriate w in $R^* + wR^{e*}$. When no risk-free rate is traded, three generalizations of the riskfree rate are interesting and can take its place in asset pricing formulas. These are the *zero-beta* rate, the *minimum-variance return* and the *constant-mimicking portfolio* return. We will also characterize these quantities by finding the appropriate w in $R^* + wR^{e*}$.

Risk free rate

If there is a risk free rate—if the payoff space $\underline{X}$ includes a unit payoff—then $E(R^{*2}) = E(R^*R^f) = E(R^*)R^f$, and we can recover the value of the risk free rate from R^* itself, or x^*,

$$R^f = \frac{E(R^{*2})}{E(R^*)} = \frac{1}{E(x^*)}. \tag{48}$$

Since we have decomposed every frontier return as $R^* + wR^{e*}$, it is interesting to express the risk free rate in this way as well. There are a number of equivalent representations,

$$R^f = R^* + R^f R^{e*} \tag{49}$$

$$R^f = R^* + \frac{E(R^{*2})}{E(R^*)} R^{e*} \tag{50}$$

$$R^f = R^* + \frac{E(R^*)}{1 - E(R^{e*})} R^{e*} \tag{51}$$

$$R^f = R^* + \frac{var(R^*)}{E(R^*)E(R^{e*})} R^{e*}. \tag{52}$$

To derive (49) and (50), start from (47) and multiply through by R^f. To establish (51) and (52), we need to show that if there is a risk free rate, then

$$R^f = \frac{E(R^{*2})}{E(R^*)} = \frac{E(R^*)}{1 - E(R^{e*})} = \frac{var(R^*)}{E(R^*)E(R^{e*})} \tag{53}$$

The first equality is given by (48). To derive the second equality, take expectations of (47), take expectations of (49), or note that $||R^{e*}||^2 + ||proj(1|R^*)||^2 = 1$, since the three quantities form a right triangle. To check the third equality, take expectations of (52),

$$R^f = E(R^*) + \frac{E(R^{*2}) - E(R^*)^2}{E(R^*)} = \frac{E(R^{*2})}{E(R^*)}.$$

The equalities in (53) all depend on the presence of a riskfree rate. When there is no riskfree rate, these three different expressions generate three different and interesting returns on the frontier, and each takes the place of the riskfree return in some asset pricing formulas.

The remaining cases assume there is no riskfree rate – the unit payoff is not in $\underline{X}$.

Zero-beta rate for R^*

The riskfree rate R^f is of course uncorrelated with R^*. Risky returns uncorrelated with R^* earn the same average return as the risk free rate if there is one, so they might take the place of R^f when the latter does not exist. For any return R^α that is uncorrelated with R^* we have $E(R^* R^\alpha) = E(R^*)E(R^\alpha)$, so

$$\alpha = E(R^\alpha) = \frac{E(R^{*2})}{E(R^*)} = \frac{1}{E(x^*)}.$$

The first equality introduces a popular notation α for this rate.

The zero-beta rate is the inverse of the price that R^* and x^* assign to the unit payoff, which is another natural generalization of the riskfree rate. It is called the zero *beta* rate because such that $cov(R^*, R^\alpha) = 0$ implies that the regression beta of R^α on R^* is zero, and everything used to be written in terms of such regression coefficients rather than covariances or second moments. More precisely, one might call it the zero beta rate *on* R^*, since one can calculate zero-beta rates for returns other than R^* and they are different. In particular, the zero-beta rate on the "market portfolio" will generally be different from the zero beta rate on R^*

I drew α in 12 as the intersection of the tangency and the vertical axis. This is a property of any return on the mean variance frontier: The expected return on an asset uncorrelated with the mean-variance efficient asset (a *zero-beta* asset) lies at the point so constructed. To check this geometry, use similar triangles: The length of R^* in 12 is $\sqrt{E(R^{*2})}$, and its vertical extent is $E(R^*)$. Therefore, $\alpha/\sqrt{E(R^{*2})} = \sqrt{E(R^{*2})}/E(R^*)$, or $\alpha = E(R^{*2})/E(R^*)$. Since R^* is on the lower portion of the mean-variance frontier, this zero beta rate α is above the minimum variance return.

Note that in general $\alpha \neq 1/E(m)$. Projecting m on $\underline{X}$ preserves asset pricing implications on $\underline{X}$ but not for payoffs not in $\underline{X}$. Thus if a risk free rate is not traded, x^* and m may differ in their predictions for the riskfree rate as for other nontraded assets.

We want to see the zero beta return that is also on the mean variance frontier in $R^* + wR^{e*}$ form. Expression (52) becomes the zero-beta return when there is no risk free rate,

$$R^\alpha = R^* + \frac{var(R^*)}{E(R^*)E(R^{e*})}R^{e*}.$$

To check, verify that it gives the correct mean,

$$E(R^\alpha) = E(R^*) + \frac{var(R^*)}{E(R^*)E(R^{e*})}E(R^{e*}) = \frac{E(R^*)^2 + var(R^*)}{E(R^*)} = \frac{E(R^{*2})}{E(R^*)}.$$

Minimum variance return.

The riskfree rate obviously is the minimum variance return when it exists. When there is no risk free rate, expression (51) becomes the minimum variance return

$$R^{\text{min. var.}} = R^* + \frac{E(R^*)}{1 - E(R^{e*})}R^{e*}.$$

Taking expectations,

$$E(R^{\text{min. var.}}) = E(R^*) + \frac{E(R^*)}{1 - E(R^{e*})}E(R^{e*}) = \frac{E(R^*)}{1 - E(R^{e*})}.$$

The minimum variance return retains the property (49) of the risk free rate above, that its weight on R^{e*} is the same as its mean

$$R^{\text{min. var.}} = R^* + E(R^{\text{min. var.}})R^{e*}.$$

When there is no risk free rate, the zero-beta and minimum variance returns are *not* the same.

We can derive this expression for the minimum variance return by brute force: choose w in $R^* + wR^{e*}$ to minimize variance.

$$\min_w \; var(R^* + wR^{e*}) = E[(R^* + wR^{e*})^2] - E(R^* + wR^{e*})^2 =$$

$$= E(R^{*2}) + w^2E(R^{e*}) - E(R^*)^2 - 2wE(R^*)E(R^{e*}) - w^2E(R^{e*})^2.$$

The first order condition is

$$0 = wE(R^{e*})[1 - E(R^{e*})] - E(R^*)E(R^{e*})$$

$$w = \frac{E(R^*)}{1 - E(R^{e*})}.$$

Variance is the size or second moment of the residual in a projection (regression) on 1.

$$var(x) = E\left[(x - E(x))^2\right] = E\left[(x - proj(x|1))^2\right] = ||x - proj(x|1)||^2$$

Thus, the minimum variance return is the return closest to extensions of the unit vector.

Return on constant-mimicking portfolio.

The riskfree rate is of course the return on the unit payoff. When there is no riskfree rate, expression (50),

$$\hat{R} \equiv R^* + \alpha R^{e*} = R^* + \frac{E(R^{*2})}{E(R^*)}R^{e*}$$

is the return on the traded payoff that is closest to the (nontraded) unit payoff. Precisely, this return has the property[3]

$$\hat{R} = \frac{proj(1|\underline{X})}{p\left[proj(1|\underline{X})\right]},$$

which is a natural generalization of the risk free rate property $R^f = 1/p(1)$. Since we form *mimicking portfolios* by projecting nontraded random variables on the space of payoffs, a natural name for this construct is the *constant-mimicking portfolio return.*

Note the subtle difference: the minimum variance return is the *return* closest to an extension of the unit vector. The constant-mimicking portfolio return is the return on the *payoff* closest to 1. They are not the same object when there is no risk free rate.

5.3 Relation between $p = E(mx)$, beta, and mean-variance frontiers

$p = E(mx),\ \beta$ representations and mean-variance frontiers look unrelated, but in fact they all express the same thing. This section is devoted to linking these three asset pricing representations. An overview of the ideas:

1. $p = E(mx) \Rightarrow \beta$: Given m such that $p = E(mx)$, we can derive an expected return $-\ \beta$ relationship. $m,\ x^*$or R^* all can serve as reference variables for betas. If $m = \mathbf{b}'\mathbf{f}$, then

[3] I think this is a novel result, so here's the algebra. $\underline{X}$ is spanned by R^*, R^{e*} and n so we can find $proj(1|\underline{X}) = aR^* + bR^{e*} + n$ by making sure the residual is orthogonal to R^*, R^{e*} and n :

$$0 = E\left[R^*(1 - aR^* - bR^{e*} - n)\right] = E(R^*) - aE(R^{*2}) \Rightarrow a = \frac{E(R^*)}{E(R^{*2})}$$

$$0 = E\left[R^{e*}(1 - aR^* - bR^{e*} - n)\right] = E(R^{e*}) - bE(R^{e*}) = 0 \Rightarrow b = 1$$

$$0 = E\left[n(1 - aR^* - bR^{e*} - n)\right] = E(n^2) = 0 \Rightarrow n = 0$$

Thus,

$$proj(1|X) = \frac{E(R^*)}{E(R^{*2})}R^* + R^{e*}$$

Its price is

$$p\left[proj(1|X)\right] = \frac{E(R^*\left[proj(1|X)\right])}{E(R^{*2})} = \frac{1}{E\left(R^{*2}\right)}\left[\frac{E(R^*)}{E(R^{*2})}E(R^{*2})\right] = \frac{E(R^*)}{E(R^{*2})}$$

and finally

$$\frac{proj(1|\underline{X})}{p\left[proj(1|\underline{X})\right]} = \frac{\frac{E(R^*)}{E(R^{*2})}R^* + R^{e*}}{\frac{E(R^*)}{E(R^{*2})}} = R^* + \frac{E(R^{*2})}{E(R^*)}R^{e*}.$$

$\mathbf{f}$, $proj(\mathbf{f} \mid \underline{X})$ can serve as multiple factors in a multiple beta model.

2. $p = E(mx) \Rightarrow$ mean-variance frontier. R^* is on the mean-variance frontier.
3. $\beta \Rightarrow p = E(mx)$. If we have an expected return/beta model, then $m = \mathbf{b}'\mathbf{f}$ linear in the factors satisfies $p = E(mx)$.
4. Mean-variance frontier $\Rightarrow p = E(mx)$. If a return R^{mv} is on the mean-variance frontier, then $m = a + bR^{mv}$ linear in that return is a discount factor; it satisfies $p = E(mx)$.
5. If a return is on the mean-variance frontier, then there is an expected return/beta model with that return as reference variable.

Figure 13 summarizes the equivalence of the three asset pricing views.

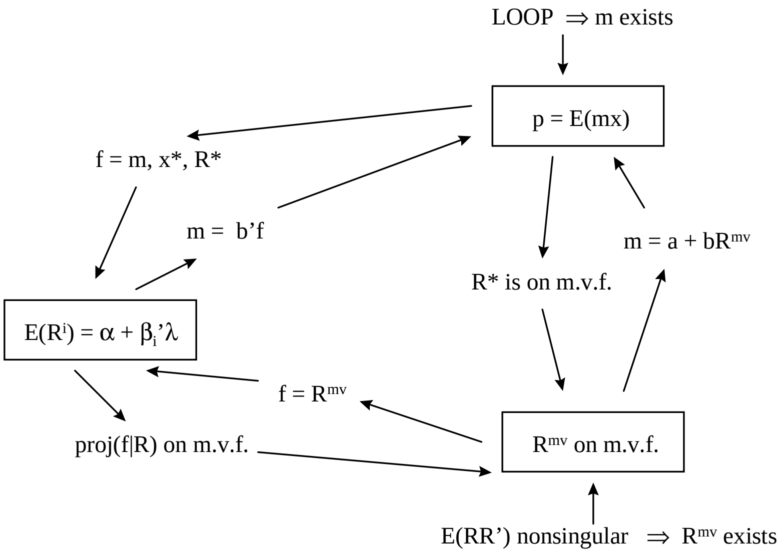


Figure 13. Relation between three views of asset pricing.

The following subsections discuss the mechanics of going from one representation to the other in detail. The next chapter discusses the implications of the existence and equivalence theorems.

5.3.1 From $p = E(mx)$ to a single beta representations

Single β representation using m

$p = E(mx)$ implies $E(R^i) = \alpha + \beta_{i,m}\lambda_m$

Start with

$$1 = E(mR^i) = E(m)E(R^i) + cov(m, R^i).$$

Thus,

$$E(R^i) = \frac{1}{E(m)} - \frac{cov(m, R^i)}{E(m)}$$

Multiply and divide by $var(m)$, define $\alpha \equiv 1/E(m)$ to get

$$E(R^i) = \alpha + \left(\frac{cov(m, R^i)}{var(m)}\right)\left(-\frac{var(m)}{E(m)}\right) = \alpha + \beta_{i,m}\lambda_m.$$

As advertised, $1 = E(mR)$ implies a single beta representation.

For example, we can equivalently state the consumption-based model as: mean asset returns should be linear in the regression betas of asset returns on $(c_{t+1}/c_t)^{-\gamma}$. Furthermore, the slope of this cross-sectional relationship λ_m is not a free parameter, though it is usually treated as such in empirical evaluation of factor pricing models. λ_m should equal the ratio of variance to mean of $(c_{t+1}/c_t)^{-\gamma}$.

The factor risk premium λ_m for marginal utility growth is negative. Positive expected returns are associated with positive correlation with consumption growth, and hence negative correlation with marginal utility growth and m. Thus, we expect $\lambda_m < 0$.

Single β representation using x^* and R^*

$1 = E(mR^i)$ implies a beta model with $x^* = proj(m|\underline{X})$ or $R^* \equiv x^*/E(x^{*2})$ as factors, e.g. $E(R^i) = \alpha + \beta_{i,R^*}[E(R^*) - \alpha]$.

It is traditional and sometimes desirable to express a pricing model in terms of returns on some portfolio rather than in terms of some real factor, such as consumption growth. Even if we found the perfect measure of utility and consumption, the fact that asset return data are measured much better and more frequently would lead us to use an equivalent return formulation for many practical purposes.

We have already seen the idea of “factor mimicking portfolios” formed by projection. We can use the same idea here: project m on to $\underline{X}$, and the result also serves as a pricing factor.

Single beta representation with x^*.

Recall that $p = E(mx)$ implies $p = E\left[proj(m \mid \underline{X})\ x\right]$, or $p = E(x^*x)$. Then we know

$$1 = E(mR^i) = E(x^*R^i) = E(x^*)E(R^i) + cov(x^*, R^i).$$

Solving for the expected return,

$$E(R^i) = \frac{1}{E(x^*)} - \frac{cov(x^*, R^i)}{E(x^*)} = \frac{1}{E(x^*)} - \frac{cov(x^*, R^i)}{var(x^*)}\frac{var(x^*)}{E(x^*)} \tag{54}$$

which we can write as the desired single-beta model,

$$E(R^i) = \alpha + \beta_{i,x^*}\lambda_{x^*}.$$

Notice that the zero-beta rate $1/E(x^*)$ appears when there is no riskfree rate.

Single beta representation with R^*.

Recall the definition,

$$R^* = \frac{x^*}{E(x^{*2})}$$

Substituting R^* for x^*, equation (54) implies that we can in fact construct a return R^* from m that acts as the single factor in a beta model,

$$E(R^i) = \frac{E(R^{*2})}{E(R^*)} - \frac{cov(R^*, R^i)}{E(R^*)} = \frac{E(R^{*2})}{E(R^*)} + \left(\frac{cov(R^*, R^i)}{var(R^*)}\right)\left(-\frac{var(R^*)}{E(R^*)}\right)$$

or, defining Greek letters in the obvious way,

$$E(R^i) = \alpha + \beta_{R^i,R^*}\lambda_{R^*} \tag{55}$$

Since the factor R^* is also a return, its expected excess return over the zero beta rate gives the factor risk premium λ_{R^*}. Applying equation (55) to R^* itself,

$$E(R^*) = \alpha - \frac{var(R^*)}{E(R^*)}. \tag{56}$$

So we can write the beta model in an even more traditional form

$$E(R^i) = \alpha + \beta_{R^i,R^*}[E(R^*) - \alpha]. \tag{57}$$

Recall that R^* is the minimum second moment frontier, on the lower portion of the mean-variance frontier. This is why R^* has an unusual negative expected excess return or factor risk premium, $\lambda_{R^*} = -var(R^*)/E(R^*) < 0$. Note that α is the zero-beta rate on R^* that I defined and discussed above, and is shown in Figure12.

Special cases

The one thing that can go wrong in these constructions is that $E(m)$, $E(x^*)$, or $E(R^*)$ might be zero, so you can't divide by them. $E(m)$ cannot be zero since, by absence of arbitrage, $m > 0$. If there is a riskfree rate, then $1/R^f = E(m) = E(x^*) = E(R^*)/E(R^{*2})$, so the presence of a finite riskfree rate also eliminates the potential problem. However, if the payoff space $\underline{X}$ under study does not include a riskfree rate, then some discount factors, including x^* and R^* may have mean zero – they may imply an infinite price for the nontraded unit payoff. We simply have to rule this out as a special case: amend the theorems to read "there is an expected return - beta representation *if* $E(x^*) \neq 0$, $E(R^*) \neq 0$". This is a technical special case, of little importance for practice. One can easily find alternative discount factors $x^* + \varepsilon$, with nonzero mean, and use them for a single beta representation. All alternative discount factors agree on the expected returns of the traded assets, though they disagree on α. Moral: Don't use mean zero discount factors for single beta representations.

$p = E(mx)$ to mean - variance frontier

R^* is the minimum second moment return, and hence R^* is on the mean-variance frontier.

R^* is the minimum second moment return. Since $E(R^2) = E(R)^2 + \sigma^2(R)$, the minimum second moment return is of course on the mean-variance frontier.

5.3.2 Mean-variance frontier to β and m

R^{mv} is on mean-variance frontier $\Leftrightarrow m = a + bR^{mv};\; E(R^i) - \alpha = \beta_i \left[E(R^{mv}) - \alpha\right]$

We rule out special cases.

We have seen that $p = E(mx)$ implies a single$-\beta$ model with *a* mean-variance efficient reference return, namely R^*. The converse is also true: for (almost) any return on the mean-variance frontier, we can define a discount factor m that prices assets as a linear function of the mean-variance efficient return, and expected returns mechanically follow a single$-\beta$ representation using the mean-variance efficient return as reference.

Mean-variance frontier to m

Theorem: There is a discount factor of the form $m = a + bR^{mv}$ if and only if R^{mv} is on the mean-variance frontier, and $R^{mv}$is not the constant-mimicking portfolio return.

Graphical argument.

Geometrically, this is a straightforward theorem. The space of discount factors is x^* plus any random variables orthogonal to the space $\underline{X}$ .We want to know when the space spanned by the unit vector 1 and a return R includes one of these discount factors.

To think about this question, look at Figure 14. In this case $\underline{X}$ is the whole space, including a unit payoff or risk free rate. We want to know when the space spanned by a return and the unit payoff includes the unique discount factor x^*. Pick a vector R^{mv} on the mean-variance frontier as shown. Then stretch it (bR^{mv}) and then subtract some of the 1 vector (a). If we pick the right a and $b,$ we can recover the discount factor $x^*.$

If the original return vector were not on the mean-variance frontier, then $a + bR^{mv}$ would point in some of the n direction orthogonal to the mean-variance frontier, for any $b \neq 0$. If $b = 0$, though, just stretching up and down the 1 vector will not get us to x^*. Thus, we can only get a discount factor of the form $a + bR^{mv}$ if R^{mv} is on the frontier.

Special cases

If the mean-variance efficient return R^{mv} that we start with happens to lie right on the intersection of the stretched unit vector and the frontier, then stretching the R^{mv} vector and adding some unit vector are the same thing, so we again can't get back to x^* by stretching and adding some unit vector. The stretched unit payoff is the riskfree rate, which is the same as the constant-mimicking portfolio return when there is a riskfree rate.

Now think about the case that the unit payoff does not intersect the space of returns. Figure 15 shows the geometry of this case. To use no more than three dimensions I had to reduce the return and excess return spaces to lines. The payoff space $\underline{X}$ is the plane joining the return and excess return sets as shown. The set of all discount factors is $m = x^* + \varepsilon$, $E(\varepsilon x) = 0$, the line through x^* orthogonal to the payoff space $\underline{X}$ in the figure. I draw the unit payoff (the dot marked "1" in Figure 15) closer to the viewer than the plane $\underline{R}$, and I draw a vector through the unit payoff coming out of the page.

Take a payoff on the mean-variance frontier, R^{mv}. (Since the return space only has two dimensions, all returns are on the frontier.) For a given R^{mv}, the space $a + bR^{mv}$ is the plane spanned by R^{mv} and 1. This plane lies sideways in the figure. As the figure shows, there is a vector $a + bR^{mv}$ in this plane that lies on the line of discount factors.

Next, the special case. This construction would go awry if the plane spanning 1 and the return R^{mv} were parallel to the plane containing the discount factor. Thus, the construction would not work for the return marked $\hat{R}$ in the Figure. The special case return is an extension of the projection of the unit vector on $\underline{X}$, which was the defining property of the constant-mimicking portfolio return $\hat{R}$.

Algebraic proof

Now, an algebraic proof that captures the same ideas. For an arbitrary $R,$ try the discount

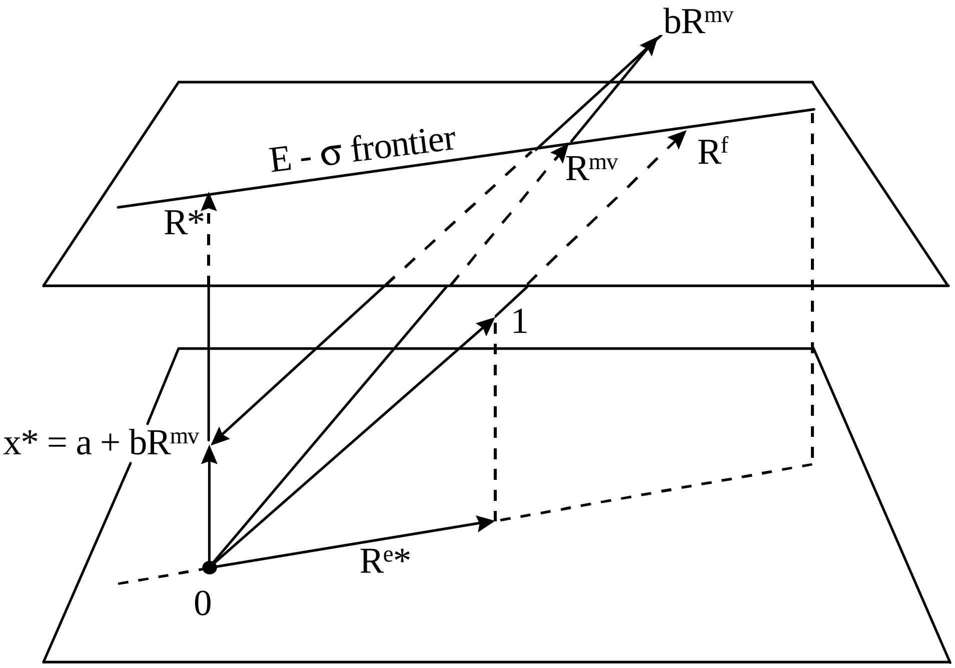


Figure 14. There is a discount factor $m = a + bR^{mv}$ if and only if R^{mv} is on the mean-variance frontier and not minimum variance.

factor model

$$m = a + bR = a + b(R^* + wR^{e*} + n). \tag{58}$$

I show that this model prices an arbitrary payoff if and only if $n = 0$ and R is not the constant-mimicking portfolio return.

We can determine a and b by forcing m to price any two assets. I find a and b to make the model price R^* and R^{e*}.

$$\begin{aligned} 1 &= E(mR^*) = aE(R^*) + bE(R^{*2}) \\ 0 &= E(mR^{e*}) = aE(R^{e*}) + bwE(R^{e*2}) = (a + bw)\, E(R^{e*}). \end{aligned}$$

Solving for a and b,

$$\begin{aligned} a &= \frac{w}{wE(R^*) - E(R^{*2})} \\ b &= -\frac{1}{wE(R^*) - E(R^{*2})}. \end{aligned}$$

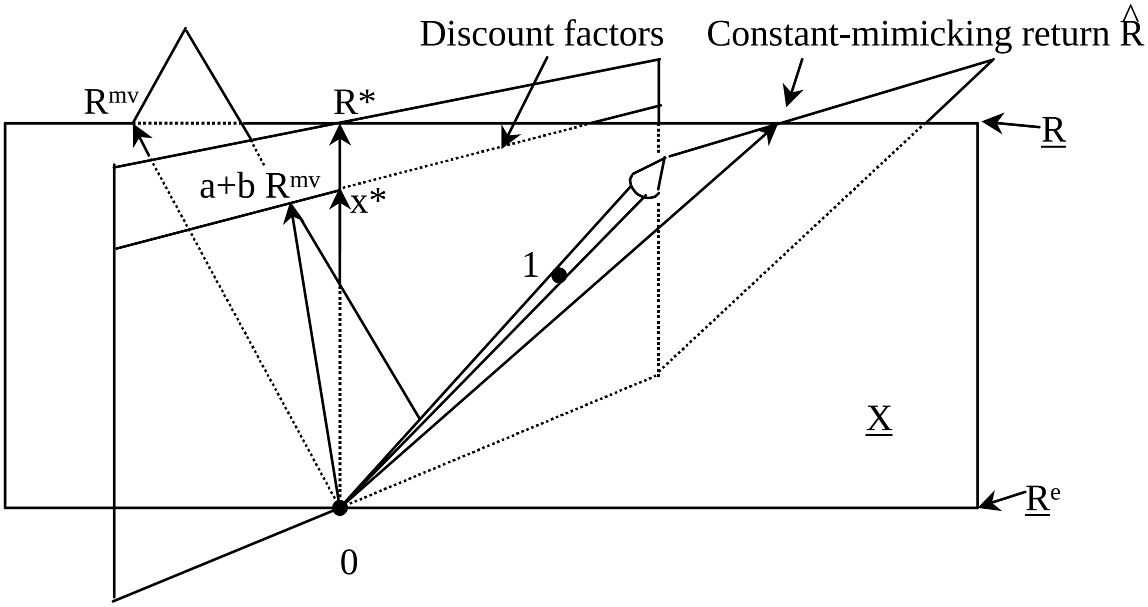


Figure 15. One can construct a discount factor $m = a + bR^{mv}$ from any mean-variance-efficient return except the constant-mimicking return $\hat{R}$.

Thus, if it is to price R^* and R^{e*}, the discount factor must be

$$m = \frac{w - (R^* + wR^{e*} + n)}{wE(R^*) - E(R^{*2})}.$$

Obviously, this construction can't work if the denominator is zero, i.e. if $w = E(R^{*2})/E(R^*)$. We saw above that the constant-mimicking portfolio return $\hat{R} = R^* + E(R^{*2})/E(R^*)R^{e*}$, so that is the case we are ruling out.

Now, let's see if m prices an arbitrary payoff x^i. Any $x^i \in \underline{X}$ can also be decomposed as

$$x^i = y^i R^* + w^i R^{e*} + n^i.$$

(See Figure 11 if this isn't obvious.) The price of x^i is y^i, since both R^{e*} and n^i are zero-price (excess return) payoffs. Therefore, we want $E(mx^i) = y^i$. Does it?

$$E(mx^i) = E\left(\frac{(w - R^* - wR^{e*} - n)(y^i R^* + w^i R^{e*} + n^i)}{wE(R^*) - E(R^{*2})}\right)$$

Using the orthogonality of R^*, R^{e*} n; $E(n) = 0$ and $E(R^{e*2}) = E(R^{e*})$ to simplify the product,

$$E(mx^i) = \frac{wy^i E(R^*) - y^i E(R^{*2}) - E(nn^i)}{wE(R^*) - E(R^{*2})} = y^i - \frac{E(nn^i)}{wE(R^*) - E(R^{*2})}.$$

To get $p(x^i) = y^i = E(mx^i)$, we need $E(nn^i) = 0$. The only way to guarantee this condition for *every* payoff $x^i \in \underline{X}$ is to insist that $n = 0$. ■

We can generalize the theorem somewhat. Nothing is special about returns; any payoff of the form $yR^* + wR^{e*}$ or $yx^* + wR^{e*}$ can be used to price assets; such payoffs have minimum variance among all payoffs with given mean and price. Of course, we proved existence not uniqueness: $m = a + bR^{mv} + \epsilon,\ E(\epsilon x) = 0$ also price assets as always.

Mean-variance frontier to β

Now, let's think about the tie between mean-variance efficiency and single beta representations. We already know mean variance frontiers ⇔ discount factor and discount factor ⇔ single beta representation, so at a superficial level we can string the two theorems together. However it is more elegant to go directly, and the special cases are also a bit simpler this way.

Theorem: There is a single beta representation with a return R^{mv} as factor,

$$E(R^i) = \alpha_{R^{mv}} + \beta_{i,R^{mv}} \left[E(R^{mv}) - \alpha\right],$$

if and only if R^{mv} is mean-variance efficient and not the minimum variance return.

This famous theorem is given by Roll (1976) and Hansen and Richard (1987). We rule out minimum variance to rule out the special case $E(m) = 0$. Graphically, the zero-beta rate is formed from the tangency to the mean-variance frontier as in figure 12. If we started at the minimum-variance return, that would lead to an infinite zero-beta rate.

Proof The mean-variance frontier is $R^{mv} = R^* + wR^{e*}$. Any return is $R^i = R^* + w^i R^{e*} + n$. Thus,

$$E(R^i) = E(R^*) + w^i E(R^{e*})$$

Now,

$$\begin{aligned} cov(R^i, R^{mv}) &= cov\left[(R^* + wR^{e*}), (R^* + w^i R^{e*})\right] \\ &= var(R^*) + ww^i var(R^{e*}) - (w + w^i)E(R^*)E(R^{e*}) \\ &= var(R^*) - wE(R^*)E(R^{e*}) + w^i\left[w\ var(R^{e*}) - E(R^*)E(R^{e*})\right] \end{aligned}$$

Thus, $cov(R^i, R^{mv})$ and $E(R^i)$ are both linear functions of w^i. We can solve $cov(R^i, R^{mv})$ for w^i, plug into the expression for $E(R^i)$ and we're done. To do this, of course, we must be able to solve $cov(R^i, R^{mv})$ for w^i. This requires

$$w \neq \frac{E(R^*)E(R^{e*})}{var(R^{e*})} = \frac{E(R^*)E(R^{e*})}{E(R^{e*2}) - E(R^{e*})^2} = \frac{E(R^*)}{1 - E(R^{e*})}$$

which is the condition for the minimum variance return. ■

5.3.3 Beta pricing ⇔linear discount factor models

Beta-pricing models are equivalent to linear models for the discount factor m.

$$m = a + \mathbf{b}'\mathbf{f} \Leftrightarrow E(R^i) = \alpha + \boldsymbol{\lambda}'\boldsymbol{\beta}_i$$

We have shown that $p = E(mx)$ implies a single beta representation using $m,\ x^*$ or R^* as factors. Let's ask the converse question: suppose we have an expected return - beta model (such as CAPM, APT, ICAPM, etc.), what discount factor model does this imply? I show that an expected return - beta model is equivalent to a model for the discount factor that is a linear function of the factors in the beta model. This is an important and central result. It gives the connection between the discount factor formulation emphasized in this book and the expected return/beta, factor model formulation common in empirical work.

One can write a linear factor model most compactly as $m = \mathbf{b}'\mathbf{f}$, letting one of the factors be a constant. However, it will be more transparent to treat a constant factor separately and explicitly, writing $m = a + \mathbf{b}'\mathbf{f}$.

Theorem. Given the model

$$m = a + \mathbf{b}'\mathbf{f},\ 1 = E(mR^i), \tag{59}$$

one can find α and λ such that

$$E(R^i) = \alpha + \boldsymbol{\lambda}'\boldsymbol{\beta}_i, \tag{60}$$

where $\boldsymbol{\beta}_i$ are the multiple regression coefficients of R^i on $\mathbf{f}$ plus a constant**.** Conversely, given a factor model of the form (60), one can find $a,\ \mathbf{b}$ such that (59) holds.

Proof: We just have to construct the relation between $(\alpha, \boldsymbol{\lambda})$ and $(a, \mathbf{b})$ and show that it works. Without loss of generality, fold the mean of the factors $\mathbf{b}'E(\mathbf{f})$ in the constant, so the factors are mean zero. Start with $m = a + \mathbf{b}'\mathbf{f}$, $1 = E(mR),$ and hence

$$E(R) = \frac{1}{E(m)} - \frac{cov(m, R)}{E(m)} = \frac{1}{a} - \frac{E(R\mathbf{f}')\mathbf{b}}{a}$$

$\boldsymbol{\beta}_i$ is the vector of the appropriate regression coefficients,

$$\boldsymbol{\beta}_i {\equiv} E\left(\mathbf{ff}'\right)^{-1} E(\mathbf{f}R^i),$$

so to get $\boldsymbol{\beta}$ in the formula, continue with

$$E(R) = \frac{1}{a} - \frac{E(R\mathbf{f}')E(\mathbf{ff}')^{-1}E(\mathbf{ff}')\mathbf{b}}{a} = \frac{1}{a} - \boldsymbol{\beta}'\frac{E(\mathbf{ff}')\mathbf{b}}{a}$$

Now, define α and $\boldsymbol{\lambda}$ to make it work,

$$\begin{aligned} \alpha &\equiv \frac{1}{E\left(m\right)} = \frac{1}{a} \\ \boldsymbol{\lambda} &\equiv -\frac{1}{a}E(\mathbf{ff}')\mathbf{b} = -\ \alpha E\left[m\mathbf{f}\right] \end{aligned} \tag{61}$$

Using (61) we can just as easily go backwards from the expected return-beta representation to $m = a + \mathbf{b}'\mathbf{f}$. ■

Given either model *there is* a model of the other form. They are not *unique*. We can add to m any random variable orthogonal to returns, and we can add risk factors with zero β and/or λ , leaving pricing implications unchanged. We can also express the multiple beta model as a single beta model with $m = a + \mathbf{b}'\mathbf{f}$ as the single factor, or use its corresponding R^*.

Equation (61) has an interesting interpretation. $\boldsymbol{\lambda}$ captures the price $E(m\mathbf{f})$ of the (demeaned) factors brought forward at the risk free rate. More specifically, if we start with underlying factors $\tilde{\mathbf{f}}$ such that the demeaned factors are $\mathbf{f} = \tilde{\mathbf{f}} {-} E(\tilde{\mathbf{f}})$,

$$\boldsymbol{\lambda} \equiv -\alpha\ p\left[\tilde{\mathbf{f}} {-} E(\tilde{\mathbf{f}})\right] = -\alpha\ \left[p(\tilde{\mathbf{f}}) {-} \frac{E(\tilde{\mathbf{f}})}{\alpha}\right]$$

λ represents the price of the of the factors less their risk-neutral valuation, i.e. the *factor risk premium*. If the factors are not traded, λ is the model's predicted price rather than a market price. Low prices are high risk premia, resulting in the negative sign.

Note that the "factors" need not be returns (though they may be); they need not be orthogonal, and they need not be serially uncorrelated or conditionally or unconditionally mean-zero. Such properties may occur as part of the economic derivation of the factor model, i.e. showing how factors proxy for marginal utility growth, but they are not required for the existence of a factor pricing representation.

Factor-mimicking portfolios.

It is often convenient to represent a factor pricing model in terms of portfolio returns rather than underlying factors that are not returns.

An old trick

One common trick in this regard is to find portfolios of assets whose means are equal to the factor risk premia. Construct a zero-cost portfolio R^{ea} that has beta 1 on factor a, $\beta_{aa} = 1$ and $\beta_{ax} = 0$ on all the other factors. The time series regression for this portfolio is

$$R_t^{ea} = a^a + 1 \times f_t^a + 0 \times f_t^b + ... + \nu_t^a; \; E(\nu_t^a f_t^i) = 0. \tag{62}$$

or, in a geometric language,

$$f_t^a = proj(R^{ea}|\text{space of factors}).$$

To construct such a portfolio, pick weights on basis assets to get the desired pattern of regression coefficients. The beta pricing model for R^{ea} now implies

$$E(R^{ea}) = \lambda_a.$$

Thus, even if a factor is not itself a return, we can find a portfolio, related to the factors, whose mean is the factor risk premium.

It would be nice to completely represent the asset pricing implications of the factor pricing model in terms of portfolios like R^a, but these returns do not do the trick, because the betas are still betas on the factor, not betas on the returns. We can interpret the return R^a from (62) as the factor plus measurement error, so the betas of a generic return R^i on R^{ea} are different from those of R^ion f^a. (Measurement error in right hand variables biases regression coefficients.) Thus, while it is true that

$$E(R^{ei}) = \beta_{i,f^a} E(R^{ea}) + \beta_{i,f^b} E(R^{eb}) + ...,$$

it is not true that

$$E(R^{ei}) = \beta_{i,R^{ea}} E(R^{ea}) + \beta_{i,R^{eb}} E(R^{eb}) + ...$$

A better idea

To find a set of returns that can fully stand in for the factors, we have to project the *factors* on the *payoffs* rather than vice versa, so that the error is orthogonal to payoffs. Recall that x^* is the projection of m on the space of payoffs and is therefore a payoff that can stand in for m. We have also used the idea that if $p = E(mx)$ then $p = E\left(proj(m|\underline{X})x\right)$ to generate a payoff in $\underline{X}$ that captures all of ms pricing implications. We just apply the same idea to the individual factors comprising m: project the *factors* on the space of payoffs. Since the discount factor m is linear in the pricing factors f, the projection is the same. I reserve the term *factor mimicking portfolios* for payoffs constructed in this way.

If, as is almost always the case, we only want to use excess returns, this projection particularly easy. Since $\underline{X} = \underline{R^e}$ we project the factors on the set of excess returns, and the result is itself an excess return

$$R^{ea} = proj(f^a|\underline{R^e}).$$

The factor-mimicking return R^{ea} then satisfies

$$f^a = R^{ea} + \xi^a; \;\; E(\xi^a R^e) = 0 \, \forall R^e \in \underline{R^e}.$$

From $E(\xi^a R^e) = 0$, if $m = a + \mathbf{b}'\mathbf{f} = a + \mathbf{b}'\left(\mathbf{R}^e + \boldsymbol{\xi}\right)$ prices assets, so will $m = a + \mathbf{b}'\mathbf{R}^e$. Thus we have found a set of excess returns that completely captures the pricing implications of the original factors. From the above theorem relating discount factors to expected return - beta representations, expected excess returns on all assets will obey

$$E(R^{ei}) = \beta_{i,R^{ea}} E(R^{ea}) + \beta_{i,R^{eb}} E(R^{eb}) + ...$$

The $\beta_{i,R^{ea}}$ are not equal to the β_{i,f^a} and $E(R^{ea}) \neq \lambda_a$, but the product explains expected returns as well as the original factor model.

If we want to use returns, retaining the factor model's predictions for the risk free rate, it is a bit more complicated. It's easy to define a factor mimicking *payoff* by $x^* = proj(m|\underline{X})$, and this would work exactly as in the last paragraph. But this payoff is not a return, since it need not have price 1. If we want factor-mimicking *returns*, we can proceed by analogy to R^*, which is created from x^* by $R^* = x^*/p(x^*)$. Define

$$x^a = proj(f^a|\underline{X})$$

and then

$$R^a = \delta \; x^a$$

$$\delta = 1/p(x^a),$$

δR^a satisfies

$$f^a = \delta R^a + \xi^a; \; E(\xi^a x) = 0 \, \forall x \in \underline{X}.$$

Still linear functions of δR^a are linear functions of R^a so m linear in the returns R^a has the same pricing implications as m linear in $\mathbf{f}$. By the above theorem we once again can express this as a beta pricing model,

$$E(R^i) = \alpha + \beta_{i,R^a} E(R^a - \alpha) + \beta_{i,R^b} E(R^b - \alpha) + ...$$

Again, the individual β and $\lambda = E(R^a - \alpha)$ terms are not the same as for the factors, but the product is, so these factor-mimicking portfolios capture the full implications of the factor model.

Why can't we project on the space of returns directly? Because that isn't a space: it does not contain zero. To project on returns, you can't just take linear combinations, you have to add the side constraint that the weights always sum to one.

5.4 Testing for priced factors: lambdas or b's?

---

b_j asks whether factor j *helps to price* assets given the other factors. b_j gives the *multiple* regression coefficient of m on f_j given the other factors.

λ_j asks whether factor j *is priced*, or whether its factor-mimicking portfolio carries a positive risk premium. λ_j gives the *single* regression coefficient of m on f_j.

Therefore, when factors are correlated, one should test $b_j = 0$ to see whether to include factor j given the other factors rather than test $\lambda_j = 0$.

Expected return-beta models defined with *single* regression betas give rise to λ with multiple regression interpretation that one can use to test factor pricing.

One is often not sure exactly what factors are important in pricing a cross-section of assets. There are two natural ways to ask whether we should include a given factor. We can ask whether the risk premium λ_j of a factor is zero, or we can ask whether b_j is zero, i.e. if the pricing factor enters in the discount factor. (The b's are *not* the same as the β's. $\mathbf{b}$ are the regression coefficient of m on $\mathbf{f}$, $\boldsymbol{\beta}$ are the regression coefficients of R^i on $\mathbf{f}$.)

Section 3.3 gave us the tools; In this section, I use those tools to compare the two approaches to testing factor pricing models.

$\mathbf{b}$ and $\boldsymbol{\lambda}$ are related. Recall from (61) that when the model is expressed $m = a + \mathbf{b}'\mathbf{f}$ and the factors de-meaned,

$$\boldsymbol{\lambda} \equiv -\alpha E(\mathbf{f}\,\mathbf{f}')\mathbf{b} = -\,\alpha E\left[m\mathbf{f}\right] = -\,\alpha p(\mathbf{f})$$

Thus, when the factors are orthogonal, each $\lambda_j = 0$ if and only if the corresponding $b_j = 0$. The distinction between b and λ only matters when the factors are correlated. Factors are often correlated however.

λ_j captures whether factor f_j *is priced.* b_j captures whether factor f_j is marginally useful in pric*ing* assets, given the presence of other factors. If $b_j = 0$, we can price assets just as well without factor f_j as with it.

λ_j is proportional to the *single* regression coefficient of m on f. $\lambda_j = -\alpha\ cov(m, f_j)$. $\lambda_j = 0$ asks the corresponding single regression coefficient question—"is factor j correlated with the true discount factor?"

b_j is the *multiple* regression coefficient of m on f_j given all the other factors. This just follows from $m = \mathbf{b}'\mathbf{f}$. (Regressions don't have to have error terms!) A *multiple* regression coefficient β_j in $y = \mathbf{x}\boldsymbol{\beta} + \varepsilon$ is the way to answer "does x_j help to explain variation in y *given* the presence of the other x's?" When you want to ask the question, "should I include factor j given the other factors?" you want to ask the *multiple* regression question. You want to know if factor j has *marginal* explanatory power for m and hence for pricing assets. When there is a difference — when the factors are correlated — you want to test b_j not λ_j.

Here is an example. Suppose the CAPM is true, which is the single factor model

$$m = a + bR^m$$

where R^m is the "market return." Consider any other return R^x, positively correlated with R^m (x for extra). If we try a factor model with the spurious factor R^x, the answer is

$$m = a + bR^m + 0 \times R^x,$$

the corresponding b_x is obviously zero, indicating that adding this factor does not help to price assets.

However, since the correlation of R^x with R^m and hence m is positive R^x earns a positive expected excess return, and $\lambda_x > 0$. In the expected return - beta model

$$E(R^i) = \alpha + \beta_{im}\lambda_m + \beta_{ix}\lambda_x$$

$\lambda_m = E(R^m) - \alpha$ is unchanged by the addition of the spurious factor. However, since the factors R^m, R^x are correlated, the multiple regression *betas* of R^i on the factors change when we add the extra factor x. For example, β_{im} may decline if β_{ix} is positive, so the new model explains the same expected return $E(R^i)$. Thus, the expected return - beta model will indicate a risk premium for β_x exposure, and many assets will have β_x exposure (R^x for example!) even though factor R^x is spurious.

So, as usual, the answer depends on the question. If you want to know whether factor i *is priced*, look at λ (or $E(mf^i)$). If you want to know whether factor i *helps to price other assets*, look at b_i. This is not an issue about sampling error or testing. All moments above are population values.

Chapter 6. Implications of existence and equivalence theorems

Existence of a discount factor means $p = E(mx)$ is innocuous, and all content flows from the discount factor model.

The theorems apply to sample moments too; the dangers of fishing up ex-post or sample mean-variance efficient portfolios.

Sources of discipline in factor fishing expeditions.

The joint hypothesis problem. How efficiency tests are the same as tests of economic discount factor models.

Factors vs. their mimicking portfolios.

Testing the number of factors.

The theorems on the existence of a discount factor, and the equivalence between the $p = E(mx)$, expected return - beta, and mean-variance views of asset pricing have important implications for how we approach and evaluate empirical work.

$p = E(mx)$ *is innocuous*

Before Roll (1976), expected return – beta representations had been derived in the context of special and explicit economic models, especially the CAPM. In empirical work, the success of any expected return - beta model seemed like a vindication of the whole structure. The fact that, for example, one might use the NYSE value-weighted index portfolio in place of the return on total wealth predicted by the CAPM seemed like a minor issue of empirical implementation.

When Roll first showed that mean-variance efficiency implies a single beta representation, all that changed. *Some* single beta representation always exists, since there is some mean-variance efficient return. The asset pricing model only serves to predict that a particular return (say, the "market return") will be mean-variance efficient. Thus, if one wants to "test the CAPM" it becomes much more important to be choosy about the reference portfolio, to guard against stumbling on something that happens to be mean-variance efficient and hence prices assets by construction. This insight led naturally to the use of broader wealth indices (Stambaugh 198x) in the reference portfolio.

(A very interesting and deep fact is that these attempts have been dismal failures. Using statistical measures, stocks are well priced by ad-hoc stock portfolios, bonds by bonds, foreign exchange by foreign exchange and so on. More recently, stocks sorted on size, book/market, and past performance characteristics are priced by portfolios sorted on those characteristics. Covariances with the returns on each form of wealth have very little explanatory power for expected returns of other forms of wealth. The fundamental idea that assets gain expected return premia by covariance with, and hence diversification of, the widest possible portfolio seems to fail against the alternative that expected return premia are determined by fairly narrow portfolios. This fact suggests that risks are not as well shared as in our models.)

The good news in this existence theorem is that you can always start by writing an expected return-beta model, knowing that almost no structure has been imposed in so doing. The bad news is that you haven't gotten very far by doing this. All the economic, statistical and predictive content comes in picking the factors.

The more modern statement of the same theorem (Ross 1978, Harrison and Kreps 1979) is that, from the law of one price, there exists *some* discount factor m such that $p = E(mx)$. The content is all in $m = f(\text{data})$ not in $p = E(mx)$. Again, an asset pricing framework that initially seemed to require a lot of completely unbelievable structure–the representative consumer consumption-based model in complete frictionless markets–turns out to require (almost) no structure at all. Again, the good news is that you can always start by writing $p = E(mx)$, and need not suffer criticism about hidden contingent claim or representative consumer assumptions in so doing. The bad news is that you haven't gotten very far by writing $p = E(mx)$ as all the economic, statistical and predictive content comes in picking the discount factor model $m = f(\text{data})$.

Ex-ante and ex-post.

I have been deliberately vague about the probabilities underlying expectations and other moments in the theorems. The fact is, the theorems hold for *any* set of probabilities[4]. Thus, the theorems work equally well *ex-ante* as *ex-post*: $E(mx), \beta, E(R)$ and so forth can refer to agent's subjective probability distributions, objective population probabilities, or to the moments realized in a given sample.

Thus, if the law of one price holds in a sample, one may form an x^* from *sample* moments that satisfies $p(x) = E(x^*x)$, *exactly,* in that sample, where $p(x)$ refers to observed prices and $E(x^*x)$ refers to the sample average. Equivalently, if the *sample* covariance matrix of a set of returns is nonsingular, there exists an *ex-post* mean-variance efficient portfolio for which sample average returns line up exactly with sample regression betas.

This observation, it seems to me, points to a great danger in the widespread exercise of searching for and statistically evaluating ad-hoc asset pricing models. Such models are *guaranteed* empirical success in a sample if one places little enough structure on what is included in the discount factor function. The only reason the model doesn't work *perfectly* is whatever restrictions the researcher has imposed on the number or identity factors included in m, or the parameters of the function relating the factors to m. Since these restrictions are the *entire* content of the model, they had better be interesting, carefully described and well motivated!

[4] Precisely, any set of probabilities that agree on agree on impossible (zero-probability) events.

Obviously, this is typically not the case or I wouldn't be making such a fuss about it. Most empirical asset pricing research posits an ad-hoc pond of factors, fishes around a bit in that set, and reports statistical measures that show "success," in that the model is not statistically rejected in pricing an ad-hoc set of portfolios. The set of discount factors is usually not large enough to give the zero pricing errors we know are possible, yet the boundaries are not clearly defined.

Discipline

What is wrong, you might ask, with finding an ex-post efficient portfolio or x^* that prices assets by construction? Perhaps the lesson we should learn from the existence theorems is to forget about economics, the CAPM, marginal utility and all that, and simply price assets with ex-post mean variance efficient portfolios that we know set pricing errors to zero!

The mistake is that a portfolio that is ex-post efficient in one sample, and hence prices all assets in that sample, is unlikely to be mean-variance efficient, ex-ante or ex-post, in the next sample, and hence is likely to do a poor job of pricing assets in the future. Similarly, the portfolio $x^* = p'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}$ (using the sample second moment matrix) that is a discount factor by construction in one sample is unlikely to be a discount factor in the next sample; the required portfolio weights $p'E(\mathbf{x}\mathbf{x}')^{-1}$ change, often drastically, from sample to sample.

For example, suppose the CAPM is true, the market portfolio is ex-ante mean-variance efficient, and sets pricing errors to zero if you use true or subjective probabilities. Nonetheless, the market portfolio is unlikely to be *ex-post* mean-variance efficient in any given sample. In any sample, there will be lucky winners and unlucky losers. An *ex-post* mean variance efficient portfolio will be a Monday-morning quarterback; it will tell you to put large weights on assets that happened to be lucky in a given sample, but are no more likely than indicated by their betas to generate high returns in the future. "Oh, if I had only bought Microsoft in 1982..." is not a useful guide to forming a mean-variance efficient portfolio today.

The only solution is to impose some kind of discipline in order to avoid dredging up spuriously good in-sample pricing.

The situation is the same as in traditional regression analysis. Regressions are used to forecast or to explain a variable y by other variables $\mathbf{x}$ in a regression $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. By blindly including right hand variables, one can produce models with arbitrarily good statistical measures of fit. But this kind of model is typically unstable out of sample or otherwise useless for explanation or forecasting. One has to carefully and thoughtfully limit the search for right hand variables $\mathbf{x}$ to produce good models.

What makes for an interesting set of restrictions? Econometricians wrestling with $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ have been thinking about this question for about 50 years, and the best answers are 1) use economic theory to carefully specify the right hand side and 2) use a battery of cross-sample and out-of-sample stability checks.

Alas, this advice is hard to follow. Economic theory is usually either silent on what variables to put on the right hand side, or allows a huge range of variables. The same is true in finance. “What are the fundamental risk factors?” is still an unanswered question. At the same time one can appeal to the APT and ICAPM to justify the inclusion of just about any desirable factor (Fama 1991 calls these theories a “fishing license.”) Thus, you will grow old waiting for theorists to provide useful answers to this kind of question.

Following the purely statistical advice, the battery of cross-sample and out-of-sample tests usually reveals the model is unstable, and needs to be changed. Once it is changed, there is no more out-of-sample left to check it. Furthermore, even if one researcher is pure enough to follow the methodology of classical statistics, and wait 50 years for another fresh sample to be available before contemplating another model, his competitors and journal editors are unlikely to be so patient. In practice, then, out of sample validation is not a strong guard against fishing.

Nonetheless, these are the only standards we have to guard against fishing. In my opinion, the best hope for finding pricing factors that are robust out of samples and across different markets, is to try to understand the fundamental macroeconomic sources of risk. By this I mean, tying asset prices to macroeconomic events, in the way the ill-fated consumption based model does via $m_{t+1} = \beta u'(c_{t+1})/u'(c_t)$. The difficulties of the consumption-based model has made this approach lose favor in recent years. However, the alternative approach is also running into trouble: every time a new anomaly or data set pops up, a new set of ad-hoc factors gets created to explain them! Also models specified with economic fundamentals will always seem to do poorly in a given sample against ad-hoc variables (especially if one fishes an ex-post mean-variance efficient portfolio out of the latter!). But what other source of discipline do we have?

In any case, one should always ask of a factor model, “what is the compelling economic story that restricts the range of factors used?” and / or “what *statistical* restraints are used to keep from discovering ex-post mean variance efficient portfolios, or to ensure that the results will be robust across samples?” The existence theorems tell us that the answers to these questions are the *only* content of the exercise. If the purpose of the model is not just to *predict* asset prices but also to *explain* them, this puts an additional burden on careful *economic* motivation of the risk factors.

There is a natural resistance to such discipline built in to our current statistical methodology for evaluating models (and papers). When the last author fished around and produced a popular though totally ad-hoc factor pricing model that generates 1% average pricing errors, it is awfully hard to persuade readers, referees, journal editors, and clients that your economically motivated factor pricing model is better despite 2% average pricing errors. Your model may really be better and will therefore continue to do well out of sample when the fished model falls by the wayside of financial fashion, but it is hard to get past statistical measures of in-sample fit. One hungers for a formal measurement of the number of hurdles imposed on a factor fishing expedition, like the degrees of freedom correction in $\bar{R}^2$. Absent a numerical correction, we have to use judgment to scale back apparent statistical successes by the amount of economic and statistical fishing that produced them.

Irrationality and Joint Hypothesis

Finance contains a long history of fighting about “rationality” vs. “irrationality” and “efficiency” vs. “inefficiency” of asset markets. The results of many empirical asset pricing papers are sold as evidence that markets are “inefficient” or that investors are “irrational.” For example, the crash of October 1987, and various puzzles such as the small-firm, book/market, seasonal effects or long-term predictability (discussed below) have all been sold this way.

However, none of these puzzles documents an arbitrage opportunity[5]. Therefore, we know that there is *a* “rational model”–*a* stochastic discount factor, *an* efficient portfolio to use in a single-beta representation—that rationalizes them all. And we can confidently predict this situation to continue; real arbitrage opportunities do not last long! Fama (1970) contains a famous statement of the same point. Fama emphasized that any test of “efficiency” is a *joint* test of efficiency and a “model of market equilibrium.” Translated, an asset pricing model, or a model of m.

But surely markets can be “irrational” or “inefficient” without requiring *arbitrage* opportunities? Yes, they can, if the discount factors that generate asset prices are disconnected from marginal rates of substitution or transformation in the real economy. But now we are right back to specifying and testing economic models of the discount factor! At best, an asset pricing puzzle might be so severe that we can show that the required discount factors are completely “unreasonable” (by some standard) measures of marginal rates of substitution and/or transformation, but we still have to say *something* about what a reasonable marginal rate looks like.

In sum, the existence theorems mean that there are no quick proofs of “rationality” or “irrationality.” The only game in town for the purpose of *explaining* asset prices is thinking about economic models of the discount factor.

Mimicking portfolios

The theorem $x^* = proj(m|\underline{X})$ also has interesting implications for empirical work. The pricing implications of any model can be equivalently represented by its factor-mimicking portfolio. If there is any measurement error in a set of economic variables driving m, the factor-mimicking portfolios will price assets better.

Thus, it is probably not a good idea to evaluate economically interesting models with statistical horse races against models that use portfolio returns as factors. Economically interesting models, even if true and perfectly measured, will just equal the performance of their own factor-mimicking portfolios, even in large samples. They will always lose in sample against ad-hoc factor models that find nearly ex-post efficient portfolios.

This said, there is an important place for models that use returns as factors. *After* we have found the underlying true macro factors, practitioners will be well advised to look at the factor-mimicking portfolio on a day-by-day basis. Good data on the factor-mimicking portfolios will be available on a minute-by-minute basis. For many purposes, one does not

[5] The closed-end fund puzzle comes closest since it documents an apparent violation of the law of one price. However, you can’t costlessly short closed end funds, and we have ignored short sales constraints so far.

have to understand the *economic* content of a model. But this fact does not tell us to circumvent the process of understanding the true macroeconomic factors by simply fishing for factor-mimicking portfolios. The experience of practitioners who use factor models seems to bear out this advice. Large commercial factor models resulting from extensive statistical analysis (otherwise known as fishing) perform poorly out of sample, as revealed by the fact that the factors and loadings (β) change all the time.

The number of factors.

Many assets pricing tests focus on the *number* of factors required to price a cross-section of assets. The equivalence theorems imply that this is a silly question. A linear factor model $m = \mathbf{b}'\mathbf{f}$ or its equivalent expected return / beta model $E(R^i) = \alpha + \boldsymbol{\beta}'_{i\mathbf{f}}\boldsymbol{\lambda}_{\mathbf{f}}$ are not unique representations. In particular, given any multiple-factor or multiple-beta representation we can easily find a single-beta representation. The single factor $m = \mathbf{b}'\mathbf{f}$ will price assets just as well as the original factors $\mathbf{f}$, as will $x^* = proj(\mathbf{b}'\mathbf{f} \mid \underline{X})$ or the corresponding R^*. All three options give rise to single-beta models with exactly the same pricing ability as the multiple factor model. We can also easily find equivalent representations with different numbers (greater than one) of factors. For example, write

$$m = a + b_1 f_1 + b_2 f_2 + b_3 f_3 = a + b_1 f_1 + b_2 \left(f_2 + \frac{b_3}{b_2} f_3 \right) = a + b_1 f_1 + b_2 \hat{f}_2$$

to reduce a "three factor" model to a "two factor" model. In the ICAPM language, consumption itself could serve as a single state variable, in place of the S state variables presumed to drive it.

There is a reason to be interested in a multiple factor representation. Sometimes the factors have an economic interpretation that is lost on taking a linear combination. But the pure *number* of pricing factors is not a meaningful question.

6.1 Discount factors vs. mean, variance and beta.

The essential difference is contingent claims as the commodity space rather than portfolio return moments.

The point of the previous chapter was to show how the discount factor, mean-variance, and expected return- beta models are all equivalent representations of asset pricing. It seems a good moment to contrast them as well; to understand why the mean-variance and beta language developed first, and to think about why the discount factor language seems to be taking over.

Asset pricing started by putting mean and variance of returns on the axes, rather than payoff in state 1 payoff in state 2, etc. as we do now. The early asset pricing theorists posed the question just right: they wanted to treat assets in the apples-and-oranges, indifference curve and budget set framework of macroeconomics. The problem was, what labels to put on the axis? Clearly, "IBM stock" and "GM stock" is not a good idea; investors do not value securities per se, but value some aspects of the stream of random cash flows that those securities give rise to.

Mean and variance of portfolios returns is a natural specification of two characteristics to be traded off. Investors plausibly want more mean and less variance. Thus, the early theorists put portfolio mean and variance on the axes. They gave investors "utility functions" defined over this mean and variance the way standard utility functions are defined over apples and oranges. The mean-variance frontier is the "budget set".

With this focus on portfolio mean and standard deviation, the next step was to realize that each security's mean return measures its contribution to the portfolio mean, and that regression betas on the overall portfolio give each security's contribution to the portfolio variance. Mean return vs. beta descriptions for each security (hence, R^i) was born.

In a deep sense, the transition from mean-variance frontiers and beta models to discount factors represents the realization that putting consumption in state 1 and consumption in state 2 on the axes — specifying preferences and budget constraints over state-contingent consumption — is a much more natural mapping of standard microeconomics into finance than putting mean, variance, etc. on the axes. If for no other reason, the contingent claim budget constraints are linear, while the mean-variance frontier is not. Thus, I think, the focus on means and variance, the mean-variance frontier and expected return/beta models is all due to an accident of history, that the early asset pricing theorists happened to put mean and variance on the axes rather than state contingent consumption. If Arrow or Debreu (195x), who invented state-contingent claims, had taken on asset pricing, we might never have heard of these constructs.

Well, here we are, why prefer one language over another? I prefer the discount factor language for its simplicity, generality, mathematical convenience, and elegance. These virtues are to some extent in the eye of the beholder, but to this beholder, it is inspiring to be able to start *every* asset pricing calculation with one equation, $p = E(mx)$. $p = E(mx)$ covers all assets, including bonds, options, and real investment opportunities, while the expected return/beta formulation is not useful or very cumbersome in the latter applications. Thus, it has seemed that there are several different asset pricing theories: expected return/beta for stocks, yield-curve models for bonds, arbitrage models for options. In fact all three are just cases of $p = E(mx)$. As a particular example, *arbitrage*, in the precise sense of positive payoffs with negative prices, has not entered the equivalence discussion at all. I don't know of any way to cleanly graft absence of arbitrage on to expected return/beta models. You have to tack it on after the fact – "by the way, make sure that every portfolio with positive payoffs has a positive price." It is trivially easy to graft it on to a discount factor model: just add $m > 0$.

The choice of language is *not* about normality or return distributions. There is a lot of confusion about where return distribution assumptions show up in finance. I have made *no* distributional assumptions in any of the discussion so far. Second moments show up because $p = E(mx)$ involves a second moment. One does not need to assume normality to talk about the mean-variance frontier, or for returns on the frontier to price other assets.

Chapter 7. Conditioning information

The asset pricing theory I have sketched so far really describes prices at time t in terms of *conditional* moments. The investor's first order conditions are

$$p_t u'(c_t) = \beta E_t \left[u'(c_{t+1})x_{t+1}\right]$$

where E_t means expectation *conditional* on the investor's time t information. Sensibly, the price at time t should be higher if there is information at time t that the discounted payoff is likely to be higher than usual at time $t+1.$ The basic asset pricing equation should be

$$p_t = E_t(m_{t+1}x_{t+1}).$$

(Conditional expectation can also be written

$$p_t = E\left[m_{t+1}x_{t+1}|I_t\right]$$

when it is important to specify the *information set* I_t.).

If payoffs and discount factors were independent and identically distributed (i.i.d.) over time, then conditional expectations would be the same as unconditional expectations and we would not have to worry about the distinction between the two concepts. But stock price/dividend ratios, bond and option prices all change over time, which must reflect changing conditional moments of something on the right hand side.

One approach is to specify and estimate explicit statistical models of conditional distributions of asset payoffs and discount factor variables (e.g. consumption growth). This approach is sometimes used, and is useful in some applications, but it is usually cumbersome. As we make the conditional mean, variance covariance and other parameters of the distribution of (say) N returns depend flexibly on M information variables, the number of required parameters can quickly exceed the number of observations.

More importantly, this explicit approach typically requires us to assume that investors use the same model of conditioning information that we do. We obviously don't even observe all the conditioning information used by economic agents, and we can't include even a fraction of observed conditioning information in our models. The basic feature and beauty of asset prices (like all prices) is that they summarize an enormous amount of information that only individuals see. The events that make the price of IBM stock change by a dollar, like the events that make the price of tomatoes change by 10 cents, are inherently unobservable to economists or would-be social planners (Hayek 194x). Whenever possible, our treatment of conditioning information should allow agents to see more than we do.

If we don't want to model conditional distributions explicitly, and if we want to avoid assuming that investors only see the variables that we include in an empirical investigation, we eventually have to think about unconditional moments, or at least moments conditioned on less information than agents see. Unconditional implications are also interesting in and of themselves. For example, we may be interested in finding out why the unconditional mean returns on some stock portfolios are higher than others, even if every agent fundamentally seeks high conditional mean returns. Most statistical estimation essentially amounts to characterizing unconditional means, as we will see in the chapter on GMM. Thus, rather than *model conditional distributions*, this chapter focuses on what implications for *unconditional* moments we can derive from the *conditional* theory.

7.1 Scaled payoffs

$$p_t = E_t(m_{t+1}x_{t+1}) \Rightarrow E(p_t z_t) = E(m_{t+1}x_{t+1}z_t)$$

One can incorporate conditioning information by adding *scaled payoffs* and doing everything unconditionally. I interpret scaled returns as payoffs to *managed portfolios.*

7.1.1 Conditioning down

The unconditional implications of any pricing model are pretty easy to state. From

$$p_t = E_t(m_{t+1}x_{t+1})$$

we can take unconditional expectations to obtain[6]

$$E(p_t) = E(m_{t+1}x_{t+1}). \tag{63}$$

Thus, if we just interpret p to stand for $E(p_t)$, everything we have done above applies to unconditional moments. In the same way, we can also condition down from agents' fine information sets to coarser sets that we observe,

$$\begin{aligned} p_t &= E(m_{t+1}R_{t+1} \mid \Omega) \Rightarrow E(p_t|I \subset \Omega) = E(m_{t+1}R_{t+1} \mid I \subset \Omega) \\ &\Rightarrow p_t = E(m_{t+1}R_{t+1} \mid I_t \subset \Omega_t) \text{ if } p_t \in I_t. \end{aligned}$$

In making the above statements I used the *law of iterated expectations*, which is important enough to highlight it. This law states that if you take an expected value using less information of an expected value that is formed on more information, you get back the expected value using less information. Your best forecast today of your best forecast tomorrow is the same

[6] We need a small technical assumption that the unconditional moment or moment conditioned on a coarser information set *exists.* For example, if X and Y are normal $(0,1)$, then $E\left(\frac{X}{Y}|Y\right) = 0$ but $E\left(\frac{X}{Y}\right)$ is infinite.

as your best forecast today. In various useful guises,

$$E(E_t(x)) = E(x),$$

$$E_{t-1}(E_t(x_{t+1})) = E_{t-1}(x_{t+1})$$

$$E\left[E(x|\Omega) \mid I \subset \Omega\right] = E\left[x|I\right]$$

7.1.2 Instruments and managed portfolios

We can do more than just condition down. Suppose we multiply the payoff and price by an instrument z_t observed at time t. Then,

$$z_t p_t = E_t(m_{t+1} x_{t+1} z_t)$$

and, taking unconditional expectations,

$$E(p_t z_t) = E(m_{t+1} x_{t+1} z_t). \tag{64}$$

This is an *additional* implication of the conditional model, not captured by just conditioning down as in (63). This trick originates from the GMM method of estimating asset pricing models, discussed below. The word *instruments* for the z variables comes from the *instrumental variables estimation* heritage of GMM.

To think about equation (64), group $(x_{t+1} z_t)$. Call this product a *payoff* $x = x_{t+1} z_t$, with *price* $p = E(p_t z_t)$. Then 64 reads

$$p = E(mx)$$

once again. Rather than thinking about (64) as a instrumental variables estimate of a conditional model, we can think of it as a price and a payoff, and apply all the asset pricing theory directly.

This interpretation is not as artificial as it sounds. $z_t R_{t+1}$ are the payoffs to *managed portfolios*. An investor who observes z_t can, rather than "buy and hold," invest in an asset according to the value of z_t. For example, if a high value of z_t forecasts that asset returns are likely to be high the next period, the investor might buy more of the asset when z_t is high and vice-versa. If the investor follows a linear rule, he puts z_t dollars into the asset each period and receives $z_t R_{t+1}$ dollars the next period. If he does this, z_t and $z_t R_{t+1}$ really are prices and payoffs.

This all sounds new and different, but practically every test uses managed portfolios. For example, the size, beta, industry, book/market and so forth portfolios of stocks are all managed portfolios, since their composition changes every year in response to conditioning information – the size, beta, etc. of the individual stocks. This idea is also closely related to the deep idea of *dynamic spanning*. Markets that are apparently very incomplete can in reality provide many more state-contingencies through dynamic (conditioned on information) trading strategies.

Equation (64) offers a very simple view of how to incorporate the extra information in conditioning information: *Add managed portfolio payoffs, and proceed with unconditional moments as if conditioning information didn't exist!*

Linearity is not important. If the investor wanted to place, say, $2 + 3z^2$ dollars in the asset, we could capture this desire with an instrument $z_2 = 2 + 3z^2$. Nonlinear (measurable) transformations of time$-t$ random variables are again random variables.

We can thus incorporate conditioning information while still looking at unconditional moments instead of conditional moments, without any of the statistical machinery of explicit models with time-varying moments. The only subtleties are 1) The set of asset payoffs expands dramatically, since we can consider all managed portfolios as well as basic assets, potentially multiplying every asset return by every information variable. 2) Expected prices of managed portfolios show up for p instead of just $p = 0$ and $p = 1$ if we started with basic asset returns and excess returns.

7.2 Sufficiency of adding scaled returns

Checking the expected price of all managed portfolios is, in principle, sufficient to check *all* the implications of conditioning information.

$$E(z_t) = E(m_{t+1}R_{t+1}z_t)\ \forall z_t \in I_t \Rightarrow 1 = E_t(m_{t+1}R_{t+1})$$

$$E(p_t) = E(m_{t+1}x_{t+1})\ \forall\ \ x_{t+1} \in \underline{X}_{t+1} \Rightarrow p_t = E_t(m_{t+1}x_{t+1})$$

We have shown that we can derive *some* extra implications from the presence of conditioning information by adding scaled returns. But does this exhaust the implications of conditioning information? Are we missing something important by relying on this trick? The answer is, in principle *no*.

I rely on the following mathematical fact: The conditional expectation of a variable y_{t+1} given an information set I_t, $E(y_{t+1} \mid I_t)$ is equal to a regression forecast of y_{t+1} using every variable $z_t \in I_t$. Now, "every random variable" means every variable and every nonlinear (measurable) transformation of every variable, so there are a lot of variables in this regression! (The word *projection* and $proj(y_{t+1}|z_t)$ is used to distinguish the best forecast of y_{t+1} using only *linear* combinations of z_t from the conditional expectation.) Applying this fact to our case, let $y_{t+1} = m_{t+1}R_{t+1} - 1$. Then $E\left[(m_{t+1}R_{t+1} - 1)\ z_t\right] = 0$ for every $z_t \in I_t$ implies $1 = E(m_{t+1}R_{t+1} \mid I_t)$. Thus, no implications are lost in principle by looking at scaled returns.

"All linear and nonlinear transformations of all variables observed at time t" sounds like a lot of instruments, and it is. But there is a practical limit to the number of instruments z_t one needs to scale by, since only variables that forecast returns or m (or their higher moments) add any information.

Since adding instruments is the same thing as including potential managed portfolios, thoughtfully choosing a few instruments is the *same* thing as the thoughtful choice of a few assets or portfolios that one makes in any test of an asset pricing model. Even when evaluating completely unconditional asset pricing models, one always forms portfolios and omits many possible assets from analysis. Few studies, in fact, go beyond checking whether a model correctly prices 10-25 stock portfolios and a few bond portfolios. Implicitly, one feels that the chosen payoffs do a pretty good job of spanning the set of available risk-loadings (mean returns) and hence that adding additional assets will not affect the results. Nonetheless, since data are easily available on all 2000 or so NYSE stocks, plus AMEX and NASDAQ stocks, to say nothing of government and corporate bonds, returns of mutual funds, foreign exchange, foreign equities, real investment opportunities, etc., the use of a few portfolios means that a tremendous number of potential asset payoffs are left out in an ad-hoc manner.

In a similar manner, if one had a small set of instruments that capture all the predictability of discounted returns $m_{t+1}R_{t+1}$, then there would be no need to add more instruments. Thus, we carefully but arbitrarily select a few instruments that we think do a good job of characterizing the conditional distribution of returns. Exclusion of potential instruments is exactly the same thing as exclusion of assets. It is no better founded, but the fact that it is a common sin may lead one to worry less about it.

There is nothing special about unscaled returns, and no economic reason to place them above scaled returns. A mutual fund might come into being that follows the managed portfolio strategy and then its *unscaled* returns would be the same as an original scaled return. Models that cannot price scaled returns are no more interesting than models that can only price (say) stocks with first letter A through L. (There may be econometric reasons to trust results for nonscaled returns a bit more, but we haven't gotten to statistical issues yet.)

Of course, the other way to incorporate conditioning information is by constructing explicit parametric models of conditional distributions. With this procedure one can in fact check *all* of a model's implications about conditional moments. However, the parametric model may be incorrect, or may not reflect some variable used by investors. Including instruments may not be as efficient, but it is still consistent if the parametric model is incorrect. The wrong parametric model of conditional distributions may lead to inconsistent estimates. In addition, one avoids estimating nuisance parameters of the parametric distribution model.

7.3 Conditional and unconditional models

A conditional factor model does not imply a fixed-weight or unconditional factor model:

$m_{t+1} = \mathbf{b}_t'\mathbf{f}_{t+1},\ p_t = E_t(m_{t+1}x_{t+1})$ does not imply that $\exists \mathbf{b}\ s.t.\ m_{t+1} = \mathbf{b}'\mathbf{f}_{t+1},\ E(p_t) = E(m_{t+1}x_{t+1})$.

$E_t(R_{t+1}) = \boldsymbol{\beta}_t'\boldsymbol{\lambda}_t$ does not imply $E(R_{t+1}) = \boldsymbol{\beta}'\boldsymbol{\lambda}$.

Conditional mean-variance efficiency does not imply unconditional mean-variance efficiency.

The converse statements are true, if managed portfolios are included.

For explicit discount factor models—models whose parameters are constant over time—the fact that one looks at a conditional vs. unconditional implications makes no difference to the statement of the model.

$$p_t = E_t(m_{t+1}x_{t+1}) \Rightarrow E(p_t) = E(m_{t+1}x_{t+1})$$

and that's it. Examples include the consumption-based model with power utility, $m_{t+1} = \beta(c_{t+1}/c_t)^{-\gamma}$, and the log utility CAPM, $m_{t+1} = 1/R^W_{t+1}$.

However, linear factor models include parameters that may vary over time. In these cases the transition from conditional to unconditional moments is much more subtle. We cannot easily condition down the model at the same time as the prices and payoffs.

7.3.1 Conditional vs. unconditional factor models in discount factor language

As an example, consider the CAPM

$$m = a - bR^W$$

where R^W is the return on the market or wealth portfolio. We can find a and b from the condition that this model correctly price any two returns, for example R^W itself and a risk-free rate:

$$\left\{ \begin{array}{l} 1 = E_t(m_{t+1}R^W_{t+1}) \\ 1 = E_t(m_{t+1})R^f_t \end{array} \right. \Rightarrow \left\{ \begin{array}{l} a = \frac{1}{R^f_t} + bE_t(R^W_{t+1}) \\ b = \frac{E_t(R^W_{t+1}) - R^f_t}{R^f_t \sigma^2_t(R^W_{t+1})} \end{array} \right. . \tag{65}$$

As you can see, $b > 0$ and $a > 0$: to make a payoff proportional to the minimum second-moment return (on the inefficient part of the mean-variance frontier) we need a portfolio long the risk free rate and short the market R^W.

More importantly for our current purposes, *a and b vary over time, as $E_t(R^W_{t+1}), \sigma^2_t(R^W_{t+1})$, and R^f_t vary over time.* If it is to price assets conditionally, the CAPM must be a linear factor model with time-varying weights, of the form

$$m_{t+1} = a_t + b_t R^W_{t+1}.$$

This fact means that we can no longer transparently condition down. The statement that

$$1 = E_t \left[(a_t + b_t R^W_{t+1}) R_{t+1} \right]$$

does not imply that we can find constants a and b so that

$$1 = E \left[(a + b R^W_{t+1}) R_{t+1} \right].$$

Just try it. Taking unconditional expectations,

$$1 = E \left[(a_t + b_t R^W_{t+1}) R_{t+1} \right] = E \left[a_t R_{t+1} + b_t R^W_{t+1} R_{t+1} \right]$$

$$= E(a_t)E(R_{t+1}) + E(b_t)E(R^W_{t+1}R_{t+1}) + cov(a_t, R_{t+1}) + cov(b_t, R^W_{t+1}R_{t+1})$$

Thus, the unconditional model

$$1 = E \left[\left(E(a_t) + E(b_t) R^W_{t+1} \right) R_{t+1} \right]$$

only holds if the covariance terms above happen to be zero. Since a_t and b_t are formed from conditional moments of returns, the covariances will not, in general be zero. (To be a little more precise, I have shown that *one* choice of a and b, $a = E(a_t)$ and $b = E(b_t)$, will not work. However, if there is *any* a and b that work, they must be $a = E(a_t)$ and $b = E(b_t)$. Thus, in fact, we have shown that there is *no* a and b that work, unless the covariance terms are zero.)

On the other hand, suppose it *is* true that a_t and b_t are constant over time. Then

$$1 = E_t \left[(a + b R^W_{t+1}) R_{t+1} \right]$$

does imply

$$1 = E \left[(a + b R^W_{t+1}) R_{t+1} \right],$$

just like any other constant-parameter factor pricing model. Furthermore, the latter unconditional model implies the former conditional model, if the latter holds for all managed portfolios.

7.3.2 Conditional vs. unconditional in an expected return / beta model

To put the same observation in beta-pricing language,

$$E_t(R^i) = R^f_t + \beta_t \lambda_t \tag{66}$$

does *not* imply that

$$E(R^i) = \alpha + \beta \lambda \tag{67}$$

The reason is that β_t and β represent conditional and unconditional regression coefficients respectively.

Again, if returns and factors are i.i.d., the unconditional model can go through. In that case, $cov(\cdot) = cov_t(\cdot)$, $var(\cdot) = var_t(\cdot)$, so the unconditional regression beta is the same as the conditional regression beta, $\beta = \beta_t$. Then, we can take expectations of (66) to get (67), with $\lambda = E(\lambda_t)$. If the betas do not vary over time, the λ may still vary and $\lambda = E(\lambda_t)$.

To condition down, the covariance and variance must *each* be constant over time. It is not enough that their ratio, or conditional betas are constant. If cov_t and var_t change over time, then the unconditional regression beta, $\beta = cov/var$ is not equal to the average conditional regression beta, $E(\beta_t)$ or $E(cov_t/var_t)$. Some models specify that cov_t and var_t vary over time, but cov_t/var_t is a constant. This specification still does not imply that the unconditional regression beta $\beta \equiv cov/var$ is equal to the constant cov_t/var_t. Similarly, it is not enough that λ be constant, since $E(\beta_t) \neq \beta$. The betas must be regression coefficients, not just numbers.

7.3.3 A precise statement.

Let's formalize these observations somewhat. Let $\underline{X}$ denote the space of all portfolios of the primitive assets, *including* managed portfolios in which the weights may depend on conditioning information, i.e. scaled returns.

A *conditional factor pricing model* is a model $m_{t+1} = a_t + \mathbf{b}_t'\mathbf{f}_{t+1}$ that satisfies $p_t = E_{t+1}(m_{t+1}x_{t+1})$ for all $x_{t+1} \in \underline{X}$.

An *unconditional factor pricing model* is model $m_{t+1} = a + \mathbf{b}'\mathbf{f}_{t+1}$ satisfies $E(p_t) = E(m_{t+1}x_{t+1})$ for all $x_{t+1} \in \underline{X}$. It might be more appropriately called a *fixed-weight factor pricing model.*

Given these definitions, and the fact that the unconditional moment conditions are equivalent to the conditional moments since all managed portfolios are in $\underline{X}$ it's almost trivial that the unconditional model is just a special case of the conditional model, one that happens to have fixed weights. Thus, *a conditional factor model does not imply an unconditional factor model* (because the weights may vary) but *an unconditional factor model does imply a conditional factor model*.

It's important to remember that the unconditional model must price must price the managed portfolios too. For example, we might simply check that the static (constant a, b) CAPM captures the unconditional mean returns of a set of assets. If this model does not also price those assets *scaled* by instruments, then it is not a conditional model, or, as I argued above, really a model at all.

Of course, everything applies for the relation between a conditional factor pricing model using a fine information set (like investors' information sets) and conditional factor pricing models using coarser information sets (like ours). If you think a set of factors prices assets with respect to investors' information, that does not mean the same set of factors prices assets with respect to our, coarser, information sets.

7.3.4 Mean-variance frontiers

Define the *conditional mean-variance frontier* as the set of returns that minimize $var_t(R_{t+1})$ given $E_t(R_{t+1})$ (including the "inefficient" lower segment as usual). Define the *unconditional mean-variance frontier* as the set of returns *including managed portfolio returns* that minimize $var(R_{t+1})$ given $E(R_{t+1})$. These two frontiers are related by:

If a return is on the unconditional mean-variance frontier, it is on the conditional mean-variance frontier.

However,

If a return is on the conditional mean-variance frontier, it need not be on the unconditional mean-variance frontier.

These statements are exactly the opposite of what you first expect from the language. The law of iterated expectations $E(E_t(x)) = E(x)$ leads you to expect that "conditional" should imply "unconditional." But we are studying the conditional vs. unconditional mean-variance *frontier*, not raw conditional and unconditional expectations, and it turns out that exactly the opposite words apply.

Again, keep in mind that the unconditional mean variance frontier *includes* returns on managed portfolios. This definition is eminently reasonable. If you're trying to minimize variance for given mean, why tie your hands to fixed weight portfolios? Equivalently, why not allow yourself to include in your portfolio the returns of mutual funds whose advisers promise the ability to adjust portfolios based on conditioning information?

You could form a mean-variance frontier of fixed-weight portfolios of a basis set of assets, and this is what many people often mean by "unconditional mean-variance frontier." The return on the true unconditional mean-variance frontier will, in general, include some managed portfolio returns, and so will lie outside this *mean-variance frontier of fixed-weight portfolios*. Conversely, a return on the fixed-weight portfolio MVF is, in general, *not* on the unconditional or conditional mean-variance frontier. All we know is that the fixed-weight frontier lies inside the other two. It may touch, but it need not. This is not to say the fixed-weight unconditional frontier is uninteresting. For example, returns on this frontier will price fixed-weight portfolios of the basis assets. The point is that this frontier has no connection to the other two frontiers. In particular, a conditionally mean-variance efficient return (conditional CAPM) need not unconditionally price the fixed weight portfolios.

I offer several ways to see this important statement.

Using the connection to factor models

We have seen that the conditional CAPM $m_{t+1} = a_t - b_t R^W_{t+1}$ does not imply an unconditional CAPM $m_{t+1} = a - bR^W_{t+1}$. We have seen that the existence of such a conditional factor model is equivalent to the statement that the return R^W_{t+1} lies on the conditional mean-variance frontier, and the existence of an unconditional factor model $m_{t+1} = a - bR^W_{t+1}$ is equivalent to the statement that R^W is on the unconditional mean-variance frontier. Then, from the "trivial" fact that an unconditional factor model is a special case of a conditional one, we know that R^W on the unconditional frontier implies R^W on the conditional frontier but not vice-versa.

Using the orthogonal decomposition

We can see the relation between conditional and unconditional mean-variance frontiers using the orthogonal decomposition characterization of mean-variance efficiency given above. This beautiful proof is the main point of Hansen and Richard (1987).

By the law of iterated expectations, x^* and R^* generate expected prices and R^{e*} generates unconditional means as well as conditional means:

$$E\left[p = E_t(x^*x)\right] \Rightarrow E(p) = E(x^*x)$$

$$E\left[E_t(R^{*2}) = E_t(R^*R)\right] \Rightarrow E(R^{*2}) = E(R^*R)$$

$$E\left[E_t(R^{e*}R^e) = E_t(R^e)\right] \Rightarrow E(R^{e*}R^e) = E(R^e)$$

This fact is subtle and important. For example, starting with $x^* = \mathbf{p}_t' E_t(\mathbf{x}_{t+1}\mathbf{x}_{t+1}')^{-1}\mathbf{x}_{t+1}$, you might think we need a different x^*, R^*, R^{e*} to represent expected prices and unconditional means, using unconditional probabilities to define inner products. The three lines above show that this is not the case. The same old x^*, R^*, R^{e*} represent conditional as well as unconditional prices and means.

Recall that a return is mean-variance efficient if and only if it is of the form

$$R^{mv} = R^* + wR^{e*}.$$

Thus, R^{mv} is conditionally mean-variance efficient if w is any number in the time t information set.

$$\text{conditional frontier: } R^{mv}_{t+1} = R^*_{t+1} + w_t R^{e*}_{t+1},$$

and R^{mv} is unconditionally mean-variance efficient if w is any constant.

$$\text{unconditional frontier: } R^{mv}_{t+1} = R^*_{t+1} + wR^{e*}_{t+1}.$$

Constants are in the t information set; time t random variables are not necessarily constant.

Thus unconditional efficiency (including managed portfolios) implies conditional efficiency but not vice versa. As with the factor models, once you see the decomposition, it is a trivial argument about whether a weight is constant or time-varying.

Brute force and examples.

If you're still puzzled, an additional argument by brute force may be helpful.

If a return is on the unconditional MVF it must be on the conditional MVF at each date. If not, you could improve the unconditional mean-variance trade-off by moving to the conditional MVF at each date. Minimizing unconditional variance given mean is the same as minimizing unconditional second moment given mean,

$$\min E(R^2)\ s.t.\ E(R)=\mu$$

Writing the unconditional moment in terms of conditional moments, the problem is

$$\min E\left[E_t(R^2)\right]\ s.t.\ E\left[E_t(R)\right]=\mu$$

Now, suppose you could lower $E_t(R^2)$ at one date t without affecting $E_t(R)$ at that date. This change would lower the objective, without changing the constraint. Thus, you should have done it: you should have picked returns on the *conditional* mean variance frontiers.

It almost seems that reversing the argument we can show that conditional efficiency implies unconditional efficiency, but it doesn't. Just because you have minimized $E_t(R^2)$ for given value of $E_t(R)$ at each date t *does not* imply that you have minimized $E(R^2)$ for a given value of $E(R)$. In showing that unconditional efficiency implies conditional efficiency we held fixed $E_t(R)$ at each date at μ, and showed it is a good idea to minimize $\sigma_t(R)$. In trying to go backwards, the problem is that a given value of $E(R)$ does not specify what $E_t(R)$ should be at each date. We can increase $E_t(R)$ in one conditioning information set and decrease it in another, leaving the return on the conditional MVF.

Figure 16 presents an example. Return B is conditionally mean-variance efficient. It also has zero unconditional variance, so it is the unconditionally mean-variance efficient return at the expected return shown. Return A is on the conditional mean-variance frontiers, and has the same unconditional expected return as B. But return A *has* some unconditional variance, and so is inside the unconditional mean-variance frontier.

As a second example,the riskfree rate is only on the unconditional mean-variance frontier if it is a constant. Remember the expression (49) for the risk free rate,

$$R^f = R^* + R^f R^{e*}.$$

The unconditional mean-variance frontier is $R^* + wR^{e*}$ with w a constant. Thus, the riskfree rate is only unconditionally mean-variance efficient if it is a constant.

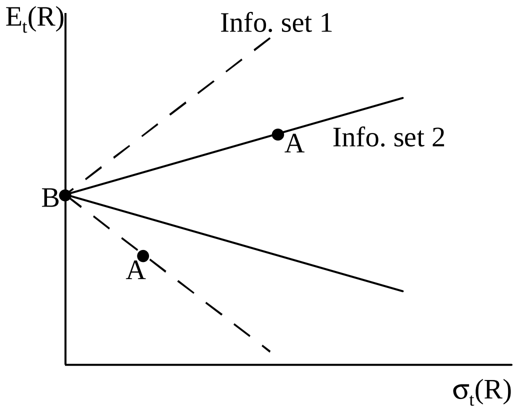


Figure 16. Return A is on the conditional mean-variance frontiers but not on the unconditional mean variance frontier.

7.3.5 Implications: Hansen-Richard Critique.

Many models, such as the CAPM, imply a *conditional* linear factor model $m_{t+1} = a_t + \mathbf{b}_t'\mathbf{f}_{t+1}$. These theorems show that such a model *does not* imply an unconditional model. Equivalently, if the model predicts that the market portfolio is conditionally mean-variance efficient, this does *not* imply that the market is unconditionally mean-variance efficient. We often test the CAPM by seeing if it explains the average returns of some portfolios or (equivalently) if the market is on the unconditional mean-variance frontier. The CAPM may quite well be true (conditionally) and fail these tests; many assets may do better in terms of *unconditional* mean vs. *unconditional* variance.

The situation is even worse than these comments seem, and are not repaired by simple inclusion of some conditioning information. Models such as the CAPM imply a conditional linear factor model with respect to *investors'* information sets. However, the best we can hope to do is to test implications conditioned down on variables that we can observe and include in a test. Thus, a conditional linear factor model *is not testable!*

I like to call this observation the "Hansen-Richard critique" by analogy to the "Roll Critique." Roll pointed out, among other things, that the wealth portfolio might not be observable, making tests of the CAPM impossible. Hansen and Richard point out that the conditioning information of agents might not be observable, and that one cannot omit it in testing a conditional model. Thus, even if the wealth portfolio *was* observable, the fact that we cannot observe agents' *information sets* dooms tests of the CAPM.

7.4 Scaled factors: a partial solution

You can expand the set of factors to test conditional factor pricing models

$$\text{factors} = \mathbf{f}_{t+1} \otimes z_t$$

The problem is that the parameters of the factor pricing model $m_{t+1} = a_t + b_t f_{t+1}$ may vary over time. A partial solution is to *model* the dependence of parameters a_t and b_t on variables in the time$-t$ information set; let $a_t = a(\mathbf{z}_t)$, $b_t = b(\mathbf{z}_t)$ where $\mathbf{z}_t$ is a vector of variables observed at time t (including a constant). In particular, why not try *linear* models

$$a_t = \mathbf{a}'\mathbf{z}_t,\ b_t = \mathbf{b}'\mathbf{z}_t$$

Linearity is not restrictive: z_t^2 is just another instrument. The only criticism one can make is that some instrument z_{jt} is important for capturing the variation in a_t and b_t, and was omitted. For instruments on which we have data, we can meet this objection by trying z_{jt} and seeing whether it does, in fact, enter significantly. However, for instruments z_t that are observed by agents but not by us, this criticism remains valid.

Linear discount factor models lead to a nice interpretation as *scaled factors,* in the same way that linearly managed portfolios are scaled returns. With a single factor and instrument, write

$$m_t = a(z_t) + b(z_t) f_{t+1} \tag{68}$$

$$= a_0 + a_1 z_t + (b_0 + b_1 z_t) f_{t+1}$$

$$= a_0 + a_1 z_t + b_0 f_{t+1} + b_1 \left(z_t f_{t+1} \right). \tag{69}$$

Thus, in place of the one-factor model with time-varying coefficients (68), we have a four-factor model (constant, z_t , f_{t+1}, $z_t f_{t+1}$) with fixed coefficients, 69.

Since the coefficients are now fixed, we *can* use the scaled-factor model with unconditional moments.

$$p_t = E_t \left[\left(a_0 + a_1 z_t + b_0 f_{t+1} + b_1 \left(z_t f_{t+1} \right) \right) \; x_{t+1} \right] \Rightarrow$$

$$E(p_t) = E\left[\left(a_0 + a_1 z_t + b_0 f_{t+1} + b_1 (z_t f_{t+1}) \right) \; x_{t+1} \right]$$

For example, in standard derivations of CAPM, the market (wealth portfolio) return is *conditionally* mean-variance efficient; investors want to hold portfolios on the *conditional* mean-variance frontier; *conditionally* expected returns follow a *conditional* single-beta representation, or the discount factor m follows a *conditional* linear factor model

$$m_{t+1} = a_t - b_t R^W_{t+1}$$

as we saw above.

But none of these statements mean that we can use the CAPM *unconditionally*. Rather than throw up our hands, we can add some scaled factors. Thus, if, say, the dividend/price ratio and term premium do a pretty good job of summarizing variation in conditional moments, the *conditional* CAPM implies an *unconditional, five-factor (plus constant) model*. The factors are a constant, the market return, the dividend/price ratio, the term premium, and the market return *times* the dividend-price ratio and the term premium.

The unconditional pricing implications of such a five-factor model could, of course, be summarized by a single$-\beta$ representation. (See the caustic comments in the section on implications and equivalence.) The reference portfolio would not be the market portfolio, of course, but a mimicking portfolio of the five factors. However, the single mimicking portfolio would not be easily interpretable in terms of a single factor conditional model and two instruments. In this case, it might be more interesting to look at a multiple $-\beta$ or multiple-factor representation.

If we have many factors f and many instruments z, we should in principle multiply every factor by every instrument,

$$m = b_1 f_1 + b_2 f_1 z_1 + b_3 f_1 z_2 + ... + b_{N+1} f_2 + b_{N+2} f_2 z_1 + b_{N+3} f_2 z_2 + ...$$

This operation can be compactly summarized with the *Kronecker product* notation, $a \otimes b$, which means “multiply every element in vector a by every element in vector b, or

$$m_{t+1} = \mathbf{b}'(f_{t+1} \otimes z_t).$$

7.5 Summary

When you first think about it, conditioning information sounds scary – how do we account for time-varying expected returns, betas, factor risk premia, variances, covariances, etc. However, the methods outlined in this chapter allow a very simple and beautiful solution to the problems raised by conditioning information. To express the conditional implications of a given model, all you have to do is include some scaled or managed portfolio returns, and then pretend you never heard about conditioning information.

Some factor models are conditional models, and have coefficients that are functions of investors’ information sets. In general, there is no way to test such models, but if you are willing to assume that the relevant conditioning information is well summarized by a few variables, then you can just add new factors, equal to the old factors scaled by the conditioning variables, and again forget that you ever heard about conditioning information.

You may want to remember conditioning information as a diagnostic and in economic interpretation of the results. It may be interesting to take estimates of a many factor model, $m_t = a_0 + a_1 z_t + b_0 f_{t+1} + b_1 z_t f_{t+1}$, and see what they say about the implied conditional model, $m_t = (a_0 + a_1 z_t) + \ (b_0 + b_1 z_t) f_{t+1}$. You may want to make plots of conditional bs, betas, factor risk premia, expected returns,etc. But you don’t have to worry about it in estimation and testing.

Chapter 8. Factor pricing models

In the second chapter, I noted that the consumption-based model, while a complete answer to most asset pricing questions in principle, does not (yet) work well in practice. This observation motivates efforts to tie the discount factor m to other data. Linear factor pricing models are the most popular models of this sort in finance. They dominate discrete time empirical work.

Factor pricing models replace the consumption-based expression for marginal utility growth with a linear model of the form

$$m_{t+1} = a + \mathbf{b}'\mathbf{f}_{t+1}$$

a and $\mathbf{b}$ are free parameters. As we have seen above, this specification is equivalent to a multiple-beta model

$$E(R_{t+1}) = \alpha + \beta'\lambda$$

where β are multiple regression coefficients of returns R on the factors f. Here, α and λ are the free parameters.

The big question is, what should one use for factors $\mathbf{f}_{t+1}$? Factor pricing models look for variables that are good proxies for aggregate marginal utility growth, i.e., variables for which

$$\beta \frac{u'(c_{t+1})}{u'(c_t)} \approx a + \mathbf{b}'\mathbf{f}_{t+1} \tag{70}$$

is a sensible and economically interpretable approximation.

The factors that result from this search are and should be intuitively sensible. In any sensible economic model, as well as in the data, consumption is related to returns on broad-based portfolios, to interest rates, to growth in GNP, investment, or other macroeconomic variables, and to returns on production processes. All of these variables measure “wealth” or the state of the economy. Consumption is and should be high in “good times” and low in “bad times.”

Furthermore, consumption and marginal utility respond to *news*: if a change in some variable today signals high income in the future, then consumption rises *now*, by permanent income logic. This fact opens the door to *forecasting* variables: any variable that forecasts asset returns (“changes in the investment opportunity set”) or macroeconomic variables is a candidate factor. Variables such as the term premium, dividend/price ratio, stock returns, etc. can be defended as pricing factors on this logic. Though they themselves are not measures of aggregate good or bad times, they *forecast* such times.

Should factors be independent over time? The answer is, sort of. If there is a constant real interest rate, then marginal utility growth should be unpredictable. (“Consumption is a random walk” in the quadratic utility permanent income model.) To see this, just look at the first order condition with a constant interest rate,

$$u'(c_t) = \beta R^f E_t \left[u'(c_{t+1}) \right]$$

or in a more time-series notation,

$$\frac{u'(c_{t+1})}{u'(c_t)} = \frac{1}{\beta R^f} + \varepsilon_{t+1}; \quad E_t(\varepsilon_{t+1}) = 0.$$

The real risk free rate is not constant, but it does not vary a lot, especially compared to asset returns. Measured consumption growth is not exactly unpredictable but it is the least predictable macroeconomic time series, especially if one accounts properly for temporal aggregation (consumption data are quarterly averages). Thus, factors that proxy for marginal utility growth, though they don't have to be totally unpredictable, should not be highly predictable. If one chooses highly predictable factors, the model will counterfactually predict large interest rate variation.

In practice, this consideration means that one should choose the right units: Use GNP *growth* rather than level, portfolio *returns* rather than prices or price/dividend ratios, etc. However, unless one wants to impose an exactly constant risk free rate, one does not have to filter or prewhiten factors to make them exactly unpredictable.

This view of factors as intuitively motivated proxies for marginal utility growth is sufficient to carry the reader through current empirical tests of factor models. The extra constraints of a formal exposition of theory in this part have not yet constrained the factor-fishing expedition.

The precise derivations all proceed in the way I have motivated factor models: One writes down a general equilibrium model, in particular a specification of the production technology by which real investment today results in real output tomorrow. This general equilibrium produces relations that express the determinants of consumption from exogenous variables, and relations linking consumption and other endogenous variables; equations of the form $c_t = g(\mathbf{f}_t)$. One then uses this kind of equation to substitute out for consumption in the basic first order conditions. (You don't have to know more than this about general equilibrium to follow the derivations in this chapter. I discuss the economics and philosophy of general equilibrium models in some depth later, in Chapter 15.)

The formal derivations accomplish two things: they determine one particular *list of factors* that can proxy for marginal utility growth, and they prove that the relation should be *linear*. Some assumptions can often be substituted for others in the quest for these two features of a factor pricing model.

This is a point worth remembering: *all factor models are derived as specializations of the consumption-based model*. Many authors of factor model papers disparage the consumption-based model, forgetting that their factor model *is* the consumption-based model plus extra assumptions that allow one to proxy for marginal utility growth from some other variables.

Above, I argued that clear economic foundation was important for factor models, since it is the only guard against fishing. Alas, we discover here that the current state of factor pricing models is not a particularly good guard against fishing. One can call for better theories or derivations, more carefully aimed at limiting the list of potential factors and describing the fundamental macroeconomic sources of risk, and thus providing more discipline for empirical work. The best minds in finance have been working on this problem for 40 years though, so a ready solution is not immediately in sight. On the other hand, we will see that even current theory can provide much more discipline than is commonly imposed in empirical work. For example, the derivations of the CAPM and ICAPM do leave predictions for the risk free rate and for factor risk premia that are often ignored. The ICAPM gives tighter restrictions on state variables than are commonly checked: "State variables" do have to forecast something! We also see how special and unrealistic are the general equilibrium setups necessary to derive popular specifications such as CAPM and ICAPM. This observation motivates a more serious look at real general equilibrium models below.

8.1 Capital Asset Pricing Model (CAPM)

The CAPM is the model $m = a + bR^w$; R^w = wealth portfolio return. I derive it from the consumption based model by 1) Two period quadratic utility; 2) Two periods, exponential utility and normal returns; 3) Infinite horizon, quadratic utility and i.i.d. returns; 4) Log utility and normal distributions.

The CAPM is the first, most famous and (so far) most widely used model in asset pricing, as the related consumption-based model is in macroeconomics. It ties the discount factor m to the return on the "wealth portfolio." The function is linear,

$$m_{t+1} = a + bR^W_{t+1}.$$

a and b are free parameters. One can find theoretical values for the parameters a and b by requiring the discount factor m to price any two assets, such as the wealth portfolio return and risk-free rate, $1 = E(mR^W)$ and $1 = E(m)R^f$. (As an example, we did this in equation (65) above.) In empirical applications, we can also pick a and b to "best" price larger cross-sections of assets. We do not have good data on, or even a good empirical definition for, the return on total wealth. It is conventional to proxy R^W by the return on a broad-based stock portfolio such as the value- or equally-weighted NYSE, S&P500, etc.

The CAPM is of course most frequently stated in equivalent expected return / beta language,

$$E(R^i) = \alpha + \beta_{i,R^W}\left[E(R^w) - \alpha\right].$$

SECTION 8.1 CAPITAL ASSET PRICING MODEL (CAPM)

This section briefly describes some classic derivations of the CAPM. Again, we need to find assumptions that defend *which* factors proxy for marginal utility (R^W here), and assumptions to defend the *linearity* between m and the factor.

I present several derivations of the same model. Many of these derivations use classic modeling assumptions which are important in their own sake. This is also an interesting place in which to see that various sets of assumptions can often be used to get to the same place. The CAPM is often criticized for one or another assumption. By seeing several derivaitons, we can see how one assumption can be traded for another. For example, the CAPM does not in fact require normal distributions, if one is willing to swallow quadratic utility instead.

8.1.1 Two-period quadratic utility

Two period investors with no labor income and quadratic utility imply the CAPM.

Investors have quadratic preferences and only live two periods,

$$U(c_t, c_{t+1}) = -\frac{1}{2}(c_t - c^*)^2 - \frac{1}{2}\beta E[(c_{t+1} - c^*)^2]. \tag{71}$$

Their marginal rate of substitution is thus

$$m_{t+1} = \beta \frac{u'(c_{t+1})}{u'(c_t)} = \beta \frac{(c_{t+1} - c^*)}{(c_t - c^*)}.$$

The quadratic utility assumption means marginal utility is *linear* in consumption. Thus, the first target of the derivation, linearity.

Investors are born with wealth W_t in the first period and earn no labor income. They can invest in lots of assets with prices p_t^i and payoffs x_{t+1}^i, or, to keep the notation simple, returns R_{t+1}^i. They choose how much to consume at the two dates, c_t and c_{t+1}, and the *portfolio weights* α_i for their investment portfolio. Thus, the budget constraint is

$$c_{t+1} = W_{t+1} \tag{72}$$

$$W_{t+1} = R_{t+1}^W (W_t - c_t)$$

$$R^W = \sum_{i=1}^N \alpha_i R^i;\ \sum_{i=1}^N \alpha_i = 1.$$

R^W is the rate of return on total wealth.

The two-period assumption means that investors consume everything in the second period, by constraint (72). This fact allows us to substitute wealth and the return on wealth for consumption, achieving the second goal of the derivation, naming the factor that proxies for consumption or marginal utility:

$$m_{t+1} = \beta \frac{R^W_{t+1}(W_t - c_t) - c^*}{c_t - c^*} = \frac{-\beta c^*}{c_t - c^*} + \frac{\beta (W_t - c_t)}{c_t - c^*} R^W_{t+1}$$

i.e.

$$m_{t+1} = a_t + b_t R^W_{t+1}.$$

8.1.2 Exponential utility, normal distributions

$Eu(c) = e^{-\alpha c}$ and a normally distributed set of returns also produces the CAPM.

Exponential utility and normal distributions is another set of assumptions that deliver the CAPM in a one period model. This is a particularly convenient analytical form. Since it gives rise to linear demand curves, it is very widely used in models that complicate the trading structure, by introducing incomplete markets or asymmetric information.

Let utility be

$$Eu(c) = e^{-\alpha c}.$$

α is known as the *coefficient of absolute risk aversion.* If consumption is normally distributed, we have

$$Eu(c) = e^{-\alpha E(c) + \frac{\alpha^2}{2}\sigma^2(c)}.$$

Suppose this investor has initial wealth W which can be split between a riskfree asset paying R^f and a set of risky assets paying return R. Let y denote the amount of this wealth W (amount, not fraction) invested in each security. Then, the budget constraint is

$$\begin{aligned} c &= y^f R^f + y'R \\ W &= y^f + y'1 \end{aligned}$$

Plugging the first constraint into the utility function we obtain

$$Eu(c) = e^{-\alpha\left[y^f R^f + y'E(R)\right] + \frac{\alpha^2}{2} y'\Sigma y}. \tag{73}$$

As with quadratic utility, the two-period model is what allows us to set consumption to wealth and then substitute the return on the wealth portfolio for consumption growth in the discount factor.

Maximizing (73) with respect to y, y^f, we obtain the first order condition descrbing the optimal amount to be invested in the risky asset,

$$y = \Sigma^{-1}\frac{E(R) - R^f}{\alpha}$$

Sensibly, the consumer invests more in risky assets if their expected return is higher, less if his risk aversion coefficient is higher, and less if the assets are riskier. Notice that total wealth does not appear in this expression. With this setup, the amount invested in risky assets is independent of the level of wealth. This is why we say that this investor has an aversion to *absolute* rather than *relative* (to wealth) risk aversion. Note also that these "demands" for the risky assets are linear in expected returns, which is a very convenient property.

Inverting the first order conditions, we obtain

$$E(R) - R^f = \alpha\Sigma y = \alpha\ cov(R, R^m). \tag{74}$$

The consumer's total risky portfolio is $y'R$. Hence, Σy gives the covariance of each return with $y'R$, and also with the investor's overall portfolio $y^f R^f + y'R$. If all investors are identical, then the market portfolio is the same as the individual's portfolio so Σy also gives the correlation of each return with $R^m = y^f R^f + y'R$. (If investors differ in risk aversion α, the same thing goes through but with an aggregate risk aversion coefficient.)

Thus, we have the CAPM. This version is especially interesting because it ties the market price of risk to the risk aversion coefficient. Applying (74) to the market return itself, we have

$$\frac{E(R^m) - R^f}{\sigma^2(R^m)} = \alpha.$$

8.1.3 Quadratic value function, dynamic programming.

We can let consumers live forever in the quadratic utility CAPM so long as we assume that the environment is independent over time. Then the *value function* is quadratic, taking the place of the quadratic second-period utility function. This case is a nice first introduction to *dynamic programming*.

The two-period structure given above is unpalatable, since (most) investors do in fact live longer than two periods. It is natural to try to make the same basic ideas work with less restrictive and more palatable assumptions.

We can derive the CAPM in a multi-period context by replacing the second-period quadratic utility function with a quadratic *value* function. However, the quadratic value function requires the additional assumption that returns are i.i.d. (no “shifts in the investment opportunity set”). This famous observation is due to Fama (1970). It is also a nice introduction to *dynamic programming*, which is a powerful way to handle multiperiod problems by expressing them as two period problems. Finally, I think this derivation makes the CAPM more realistic, transparent and intuitively compelling. Buying stocks amounts to taking bets over *wealth;* really the fundamental assumption driving the CAPM is that marginal utility of *wealth* is linear in wealth and does not depend on other state variables.

Let’s start in a simple ad-hoc manner by just writing down a “utility function” defined over this period’s consumption and next period’s *wealth,*

$$U = u(c_t) + \beta E_t V(W_{t+1}).$$

This is a reasonable objective for an investor, and does not require us to make the very artificial assumption that he will die tomorrow. If an investor with this “utility function” can buy an asset at price p_t with payoff x_{t+1}, his first order condition (buy a little more, then x contributes to wealth next period) is

$$p_t u'(c_t) = \beta E_t \left[V'(W_{t+1}) x_{t+1} \right].$$

Thus, the discount factor uses next period’s marginal value of wealth in place of the more familiar marginal utility of consumption

$$m_{t+1} = \beta \frac{V'(W_{t+1})}{u'(c_{t+1})}$$

Now, suppose the value function were quadratic,

$$V(W_{t+1}) = -\frac{\eta}{2}(W_{t+1} - W^*)^2.$$

Then, we would have

$$\begin{aligned} m_{t+1} &= -\beta\eta \frac{W_{t+1} - W^*}{u'(c_{t+1})} = -\beta\eta \frac{R^W_{t+1}(W_t - c_t) - W^*}{u'(c_{t+1})} \\ &= \left[\frac{\beta\eta W^*}{u'(c_{t+1})} \right] - \left[-\frac{\beta\eta (W_t - c_t)}{u'(c_{t+1})} \right] R^W_{t+1}, \end{aligned}$$

or, once again,

$$m_{t+1} = a_t + b_t R^W_{t+1},$$

the CAPM!

Let’s be clear about the assumptions and what they do. 1) *The value function only depends on wealth.* If other variables entered the value function, then $\partial V / \partial W$ would depend on those other variables, and so would m. This assumption bought us the first objective of any derivation: the identity of the factors. The ICAPM, below, allows other variables in the value function, and obtains more factors. (Actually, other variables could enter so long as they don't affect the *marginal* value of wealth. The weather is an example: You like me might be happier on sunny days, but you do not value additional wealth more on sunny than on rainy days. Hence, covariance with weather does not affect how you value stocks.)

2) *The value function is quadratic*. We wanted the *marginal* value function $V'(W)$ be linear, to buy us the second objective, showing m is *linear* in the factor. Quadratic utility and value functions deliver a globally linear marginal value function $V'(W)$. By the usual Taylor series logic, linearity of $V'(W)$ is probably not a bad assumption for small perturbations, and not a good one for large perturbations.

Why is the value function quadratic?

You might think we are done. But economists are unhappy about a utility function that has *wealth* in it. Few of us are like Disney's Uncle Scrooge, who got pure enjoyment out of a daily swim in the coins in his vault. Wealth is valuable because it gives us access to more consumption. Utility functions should always be written over *consumption*. One of the few real rules in economics that keep our theories from being vacuous is that ad-hoc "utility functions" over other objects like wealth (or means and variances of portfolio returns, or "status" or "political power") should be defended as arising from a more fundamental desire for consumption.

More practically, being careful about the derivation makes clear that the superficially plausible assumption that the value function is only a function of wealth derives from the much less plausible, in fact certainly false, assumption that interest rates are constant, the distribution of returns is i.i.d., and that the investor has no risky labor income. So, let us see what it takes to defend the quadratic *value* function in terms of some *utility* function.

Suppose investors last forever, and have the standard sort of utility function

$$U = -\frac{1}{2}E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}).$$

Again, investors start with wealth W_0 which earns a random return R^W and they have no other source of income. In addition, suppose that interest rates are constant, and stock returns are i.i.d. over time.

Define the *value function* as the *maximized value* of the utility function in this environment. Thus, define $V(W)$ as[7]

$$\begin{aligned} V(W_t) &\equiv max_{\{c_t, c_{t+1}, c_{t+2} \dots \boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1}, \dots\}} E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}) \qquad (75) \\ \text{s.t. } W_{t+1} &= R^W_{t+1}(W_t - c_t);\ R^W_t = \boldsymbol{\alpha}'_t \mathbf{R}_t;\ \boldsymbol{\alpha}'_t \mathbf{1} = 1 \end{aligned}$$

(I used vector notation to simplify the statement of the portfolio problem; $\mathbf{R} \equiv [R^1$The value function is the total level of utility the investor can achieve, given how much wealth he has and any other variables constraining him. This is where the assumptions of no labor income, a constant interest rate, and i.i.d. returns come in. Without these assumptions, the value function as defined above might depend on these other characteristics of the investor's environment. For example, if there were some variable, say, "DP" that indicated returns would be high or low for a while, then the consumer would be happier, and have a high value, when DP is high, for a given level of wealth. Thus, we would have to write $V(W_t, DP_t)$

Value functions allow you to express an infinite period problem as a two period problem. Break up the maximization into the first period and all the remaining periods, as follows

$$V(W_t) = max_{\{c_t, \boldsymbol{\alpha}_t\}} \left\{ u(c_t) + \beta E_t \left[\max_{\{c_{t+1}, c_{t+2}.., \boldsymbol{\alpha}_{t+1}, \boldsymbol{\alpha}_{t+2} \dots\}} E_{t+1} \sum_{j=0}^{\infty} \beta^j u(c_{t+1+j}) \right] \right\} \quad s.\ t.\ ..$$

or

$$V(W_t) = max_{\{c_t, \boldsymbol{\alpha}_t\}} \left\{ u(c_t) + \beta E_t V(W_{t+1}) \right\} \quad s.t.\ ... \qquad (76)$$

Thus, we have defended the *existence* of a value function. Writing down a two period "utility function" over this period's consumption and next period's *wealth* is not as crazy as it might seem.

The value function is also an attractive view of how people actually make decisions. You don't think "If I buy a new car today I won't be able to buy a restaurant dinner 20 years from now" – trading off goods directly as expressed by the utility function. You think "I can't afford a new car" meaning that the decline in the value of wealth is not worth the increase in the marginal utility of consumption. Thus, the maximization in (76) describes your psychological approach to utility maximization.

The remaining question is, can the value function be quadratic? What utility function assumption leads to a quadratic value function? Here is the fun fact: *A quadratic utility function leads to a quadratic value function in this environment.* This is not a law of nature; it is not true that for any $u(c)$, $V(W)$ has the same functional form. But it is true here and a few other special cases. The "in this environment" clause is not innocuous. The value

[7] There is also a transversality condition or a lower limit on wealth in the budget constraints. This keeps the consumer from consuming a bit more and rolling over more and more debt, and it means we can write the budget constraint in present value form.

function – the achieved level of expected utility – is a result of the utility function *and* the constraints.

How could we show this fact? One way would be to try to calculate the value function by brute force from its definition, equation (75). This approach is not fun, and it does not exploit the beauty of dynamic programming, which is the reduction of an infinite period problem to a two period problem.

Instead solve (76) as a functional equation. *Guess* that the value function $V(W_{t+1})$ is quadratic, with some unknown parameters. Then use the *recursive* definition of $V(W_t)$ in (76), and solve a *two* period problem–find the optimal consumption choice, plug it into (76) and calculate the value function $V(W_t)$. If the guess was right, you obtain a quadratic function for $V(W_t)$, and determine any free parameters.

Let's do it. Specify

$$u(c_t) = -\frac{1}{2}\left(c_t - c^*\right)^2.$$

Guess

$$V(W_{t+1}) = -\frac{\gamma}{2}(W_{t+1} - W^*)^2$$

with γ and W^* parameters to be determined later. Then the problem (76) is (I don't write the portfolio choice $\boldsymbol{\alpha}$ part for simplicity; it doesn't change anything)

$$V(W_t) = \max_{\{c_t\}}\left[-\frac{1}{2}(c_t - c^*)^2 - \beta\frac{\gamma}{2}E(W_{t+1} - W^*)^2\right] \quad s.\ t.\ W_{t+1} = R^W_{t+1}(W_t - c_t).$$

(E_t is now E since I assumed i.i.d.) Substituting the constraint into the objective,

$$V(W_t) = \max_{\{c_t\}}\left[-\frac{1}{2}(c_t - c^*)^2 - \beta\frac{\gamma}{2}E\left[R^W_{t+1}(W_t - c_t) - W^*\right]^2\right]. \tag{77}$$

The first order condition with respect to c_t, using $\hat{c}$ to denote the optimal value, is

$$\hat{c}_t - c^* = \beta\gamma E\left\{\left[R^W_{t+1}(W_t - \hat{c}_t) - W^*\right]R^W_{t+1}\right\}$$

Solving for $\hat{c}_t$,

$$\hat{c}_t = c^* + \beta\gamma E\left\{\left[R^{W2}_{t+1}W_t - \hat{c}_t R^{W2}_{t+1} - W^* R^W_{t+1}\right]\right\}$$

$$\hat{c}_t\left[1 + \beta\gamma E(R^{W2}_{t+1})\right] = c^* + \beta\gamma E(R^{W2}_{t+1})W_t - \beta\gamma W^* E(R^W_{t+1})$$

$$\hat{c}_t = \frac{c^* - \beta\gamma E(R^W_{t+1})W^* + \beta\gamma E(R^{W2}_{t+1})W_t}{1 + \beta\gamma E(R^{W2}_{t+1})} \tag{78}$$

This is a *linear* function of W_t. Writing (77) in terms of the optimal value of c, we get

$$V(W_t) = -\frac{1}{2}(\hat{c}_t - c^*)^2 - \beta\frac{\gamma}{2}E\left[R^W_{t+1}(W_t - \hat{c}_t) - W^*\right]^2 \tag{79}$$

This is a *quadratic* function of W_t and $\hat{c}$. A quadratic function of a linear function is a quadratic function, so *the value function is a quadratic function of* W_t. If you want to spend a pleasant few hours doing algebra, plug (78) into (79), check that the result really is quadratic in W_t, and determine the coefficients γ, W^* in terms of fundamental parameters $\beta, c^*, E(R^W), E(R^{W2})$ (or $\sigma^2(R^W)$). The expressions for γ, W^* do not give much insight, so I don't do the algebra here.

8.1.4 Log utility

Log utility rather than quadratic utility also implies a CAPM. Log utility implies that consumption is proportional to wealth, allowing us to substitute the wealth return for consumption data.

The point of the CAPM is to avoid the use of consumption data, and so to use wealth or the rate of return on wealth instead. Log utility is another special case that allows this substitution. Log utility is much more plausible than quadratic utility.

Suppose that the investor has log utility

$$u(c) = \ln(c).$$

Define the wealth portfolio as a claim to all future consumption. Then, *with log utility, the price of the wealth portfolio is proportional to consumption itself.*

$$p^W_t = E_t\sum_{j=1}^{\infty}\beta^j\frac{u'(c_{t+j})}{u'(c_t)}c_{t+j} = E_t\sum_{j=1}^{\infty}\beta^j\frac{c_t}{c_{t+j}}c_{t+j} = \frac{\beta}{1-\beta}c_t$$

The return on the wealth portfolio is proportional to consumption growth,

$$R^W_{t+1} = \frac{p^W_{t+1} + c_{t+1}}{p^W_t} = \frac{\frac{\beta}{1-\beta}+1}{\frac{\beta}{1-\beta}}\frac{c_{t+1}}{c_t} = \frac{1}{\beta}\frac{c_{t+1}}{c_t} = \frac{1}{\beta}\frac{u'(c_t)}{u'(c_{t+1})}.$$

Thus, the log utility discount factor equals the *inverse* of the wealth portfolio return,

$$m_{t+1} = \frac{1}{R^W_{t+1}}. \tag{80}$$

Equation (80) could be used by itself: it attains the goal of replacing consumption data by some other variable. (Brown and Gibbons 1982 test a CAPM in this form.) Note that log utility is the *only* assumption so far. We do *not* assume constant interest rates, i.i.d. returns or the absence of labor income.

8.1.5 Linearizing any model: Taylor approximations and normal distributions.

Any nonlinear model $m = f(z)$ can be turned into a linear model $m = a+bz$ by assuming normal returns.

It is traditional in the CAPM literature to try to derive a *linear* relation between m and the wealth portfolio return. We could always do this by a Taylor approximation,

$$m_{t+1} \cong a_t + b_t R^W_{t+1}.$$

We can make this approximation exact in a special case, that the factors and all asset returns are normally distributed. First, I quote without proof the central mathematical trick as a lemma

Lemma 1 *(*Stein's lemma) *If f, R are bivariate normal, $g(f)$ is differentiable and $E \mid g'(f) \mid < \infty$, then*

$$cov\left[g(f), R\right] = E[g'(f)]\ cov(f, R). \tag{81}$$

Now we can use the lemma to state the theorem.

Theorem 2 *If $m = g(f)$, if f and a set of the payoffs priced by m are normally distributed returns, and if $|E[g'(f)]| < \infty$, then there is a linear model $m = a + bf$ that prices the normally distributed returns.*

Proof: First, the definition of covariance means that the pricing equation can be rewritten as a restriction between mean returns and the covariance of returns with m:

$$1 = E(mR) \Leftrightarrow 1 = E(m)E(R) + cov(m, R). \tag{82}$$

Now, given $m = g(f)$, f and R jointly normal, apply Stein's lemma (81) and (82),

$$1 = E[g(f)]E(R) + E[g'(f)]cov(f, R)$$

$$1 = E[g(f)]E(R) + cov(E[g'(f)], R)$$

Exploiting the $\Leftarrow$ part of (82), we obtain a model linear in f,

$$m = E[g(f)] + E[g'(f)][f - E(f)].$$

■

Using this trick, and recalling that we have not assumed i.i.d. so all these moments are conditional, *the log utility CAPM implies the linear model*

$$m_{t+1} = E_t\left(\frac{1}{R^W_{t+1}}\right) - E_t\left[\left(\frac{1}{R^W_{t+1}}\right)^2\right]\left[R^W_{t+1} - E_t(R^W_{t+1})\right] \tag{83}$$

if R^W_{t+1} *and all asset returns to be priced are normally distributed*. From here it is a short step to an expected return-beta representation using the wealth portfolio return as the factor.

In the same way, we can trade the quadratic utility function for normal distributions in the dynamic programming derivation of the CAPM. Starting from

$$m_{t+1} = \beta\frac{V'(W_{t+1})}{u'(c_t)} = \beta\frac{V'\left[R^W_{t+1}(W_t - c_t)\right]}{u'(c_t)}$$

we can derive an expression that links m *linearly* to R^W_{t+1} by assuming normality.

Using the same trick, the consumption-based model can be written in linear fashion, i.e. expected returns can be expressed as a linear function of betas on consumption growth rather than betas on consumption growth raised to a power. However, for large risk aversion coefficients (more than about 10 in postwar consumption data) or other transformations, the inaccuracies due to the normal or lognormal approximation can be very significant in discrete data.

The normal distribution assumption seems rather restrictive, and it is. However, the most popular class of continuous-time models specify instantaneously normal distributions even for things like options that have very non-normal discrete distributions. Therefore, one can think of the Stein's lemma tricks as a way to get to continuous time approximations without doing it in continuous time. The ICAPM, discussed next is an example.

8.2 Intertemporal Capital Asset Pricing Model (ICAPM)

Any "state variable" $\mathbf{z}_t$ can be a factor. The ICAPM is a linear factor model with wealth and state variables that forecast changes in the distribution of future returns or income.

The ICAPM generates linear discount factor models

$$m_{t+1} = a + \mathbf{b}'\mathbf{f}_{t+1}$$

In which the factors are "state variables" for the investor's consumption-portfolio decision.

The "state variables" are the variables that determine how well the investor can do in his maximization. State variables include current wealth, and also variables that describe the conditional distribution of income and asset returns the agent will face in the future or "shifts in the investment opportunity set." Therefore, optimal consumption decisions are a functions of the state variables, $c_t = g(\mathbf{z}_t)$. We can use this fact once again to substitute out consumption, and write

$$m_{t+1} = \beta \frac{u'\left[g(\mathbf{z}_{t+1})\right]}{u'\left[g(\mathbf{z}_t)\right]}.$$

Alternatively, the *value* function depends on the state variables

$$V(W_{t+1}, \mathbf{z}_{t+1}),$$

so we can write

$$m_{t+1} = \beta \frac{V_W(W_{t+1}, \mathbf{z}_{t+1})}{V_W(W_t, \mathbf{z}_t)}$$

(The marginal value of a dollar must be the same in any use, so I made the denominator pretty by writing $u'(c_t) = V_W(W_t, \mathbf{z}_t)$. This fact is known as the *envelope condition.*)

This completes the first step, naming the proxies. To obtain a linear relation, we can take a Taylor approximation, assume normality and use Stein's lemma, or, most conveniently, move to continuous time (which is really just a more convenient way of making the normal approximation.) We saw above that we can write the basic pricing equation in continuous time as

$$E\frac{dp}{p} - r^f dt = -E\left(\frac{d\Lambda}{\Lambda}\frac{dp}{p}\right).$$

(for simplicity of the formulas, I'm folding any dividends into the price process). The discount factor is marginal utility, which is the same as the marginal value of wealth,

$$\frac{d\Lambda_t}{\Lambda_t} = \frac{du'(c_t)}{u'(c_t)} = \frac{dV_W(W_t, z_t)}{V_W}$$

Our objective is to express the model in terms of factors z rather than marginal utility or value, and Ito's lemma makes this easy

$$\frac{dV_W}{V_W} = \frac{WV_{WW}}{V_W}\frac{dW}{W} + \frac{V_{Wz}}{V_W}dz + \frac{1}{2}(\text{second derivative terms})$$

(We don't have to grind out the second derivative terms if we are going to take $r^f dt = E_t\left(d\Lambda/\Lambda\right)$, though this approach removes a potentially interesting and testable implication of the model). The elasticity of marginal value with respect to wealth is often called the *coefficient of relative risk aversion,*

$$rra \equiv -\frac{WV_{WW}}{V_W}.$$

Substituting, we obtain the ICAPM, which relates expected returns to the covariance of returns with wealth, and also with the other state variables,

$$E\frac{dp}{p} - r^f dt = rra\ E\left(\frac{dW}{W}\frac{dp}{p}\right) + \frac{V_{Wz}}{V_W}E\left(dz\frac{dp}{p}\right).$$

From here, it is fairly straightforward to express the ICAPM in terms of betas rather than covariances, or as a linear discount factor model. Most empirical work occurs in discrete time; we often simply approximate the continuous time result as

$$E(R) - R^f \approx rra\ cov(R, \Delta W) + \lambda_z cov(R, \Delta z).$$

One often substitutes covariance with the wealth portfolio for covariance with wealth, and one uses factor-mimicking portfolios for the other factors dz as well. The factor-mimicking portfolios are interesting for portfolio advice as well, as they give the purest way of hedging against or profiting from state variable risk exposure.

8.3 Comments on the CAPM and ICAPM

Conditional vs. unconditional models.

Do they price options?

Why bother linearizing?

The wealth portfolio.

Ex post returns.

The implicit consumption-based model.

What are the ICAPM state variables?

CAPM and ICAPM as general equilibrium models

Is the CAPM conditional or unconditional?

Is the CAPM a conditional or an unconditional factor model? I.e., are the parameters a and b in $m = a - bR^W$ constants, or do they change at each time period, as conditioning information changes? We saw above that a conditional CAPM does not imply an unconditional CAPM, so additional steps must be taken to say anything about observed average returns.

The two period quadratic utility based derivation results in a *conditional* CAPM, since the parameters a and b can (must) change over time if the conditional moments of returns change over time. Equivalently, this two-period consumer chooses a portfolio on the *conditional* mean variance frontier, which is not on the *unconditional* frontier. The same is true of the multiperiod CAPM. Of course, if returns are not i.i.d. over time, the multi-period derivation is invalid anyway.

The log utility CAPM expressed with the inverse market return is a beautiful model, since it holds both conditionally and unconditionally. There are no free parameters that can change with conditioning information:

$$1 = E_t\left(\frac{1}{R^W_{t+1}}R_{t+1}\right) \Leftrightarrow 1 = E\left(\frac{1}{R^W_{t+1}}R_{t+1}\right).$$

In fact there are no free parameters at all! Furthermore, the model makes no distributional assumptions, so it can apply to any asset, and the model requires no specification of the investment opportunity set, or (macro language) no specification of technology.

Linearizing the log utility CAPM comes at enormous price. The expectations in the linearized log utility CAPM (83) are *conditional*. Thus, the apparent simplification of linearity destroys the nice unconditional feature of the log utility CAPM. In addition, the linearization requires normal returns and so vastly lowers the applicability of the model.

Should the CAPM price options?

As I have derived them, the quadratic utility CAPM and the nonlinear log utility CAPM should apply to *all* payoffs: stocks, bonds, options, contingent claims, etc. However, if we assume normal return distributions to obtain a linear CAPM from log utility, we can no longer hope to price options, since option returns are non-normally distributed (that's the point of options!) Even the normal distribution for regular returns is a questionable assumption. You may hear the statement "the CAPM is not designed to price derivative securities"; the statement refers to the log utility plus normal-distribution derivation of the linear CAPM.

Why bother linearizing a model? Why take the log utility model $m = 1/R^W$ which should price *any* asset, and turn it into $m_{t+1} = a_t + b_t R^W_{t+1}$ that loses the clean conditioning-down property and cannot price non-normally distributed payoffs? These tricks were developed before the $p = E(mx)$ expression of asset pricing models, when (linear) expected return-beta models were the only thing around. You need a linear model of m to get an expected return - beta model. More importantly, the tricks were developed when it was hard to estimate nonlinear models. It's clear how to estimate a β and a λ by regressions, but estimating nonlinear models used to be a big headache. Now, GMM has made it easy to estimate nonlinear models. Thus, in my opinion, linearization is mostly intellectual baggage.

The desire for linear representations and this normality trick is one of the central reasons why many asset pricing models are written in continuous time. In most continuous time models, everything is locally normal. Unfortunately for empiricists, this approach adds time-aggregation and another layer of unobservable conditioning information into the predictions of the model. For this reason, most empirical work is still based on discrete-time models. However, the local normal distributions in continuous time, even for option returns, is a good reminder that normal approximations probably aren't that bad, so long as the time interval is kept short.

What about the wealth portfolio?

The log utility derivation makes clear just how expansive is the concept of the wealth portfolio. To own a (share of) the *consumption* stream, you have to own not only all stocks, but all bonds, real estate, privately held capital, publicly held capital (roads, parks, etc.), and human capital – a nice word for "people". Clearly, the CAPM is a poor defense of common proxies such as the value-weighted NYSE portfolio. And keep in mind that given ex-post mean-variance efficient portfolios of any subset of assets (like stocks) out there, taking the theory seriously is our only guard against fishing.

Ex-post returns.

The log utility model also allows us for the first time to look at what moves returns *ex-post* as well as ex-ante. (Below, we will look at this issue in more depth). Recall that, in the log utility model, we have

$$R_{t+1}^{W} = \frac{1}{\beta}\frac{c_{t+1}}{c_t}. \tag{84}$$

Thus, the wealth portfolio return is high, ex-post, when consumption is high. This holds at every frequency: If stocks go up between 12:00 and 1:00, it must be because (on average) we all decided to have a big lunch. This seems silly. Aggregate consumption and asset returns are likely to be de-linked at high frequencies, but *how* high (quarterly?) and by what mechanism are important questions to be answered.

Implicit consumption-based models

Many users of alternative models clearly are motivated by a belief that the consumption-based model doesn't work, no matter how well measured consumption might be. This view is not totally unreasonable; as above, perhaps transactions costs de-link consumption and asset returns at high frequencies, and some diagnostic evidence suggests that the consumption behavior necessary to save the consumption model is too wild to be believed. However, the derivations make clear that the CAPM and ICAPM are not *alternatives* to the consumption-based model, they are *special cases* of that model In each case $m_{t+1} = \beta u'(c_{t+1})/u'(c_t)$ still operates. We just added assumptions that allowed us to substitute c_t in favor of other variables. One cannot adopt the CAPM on the belief that the consumption based model is *wrong.* If you think the consumption-based model is wrong, the economic justification for the alternative factor models evaporates.

The only plausible excuse for factor models is a belief that consumption *data* are unsatisfactory. However, while asset return data are well measured, it is not obvious that the S&P500 or other portfolio returns are terrific measures of the return to total wealth. "Macro factors" used by Chen, Roll and Ross (1986) and others are distant proxies for the quantities they want to measure, and macro factors based on other NIPA aggregates (investment, output, etc.) suffer from the same measurement problems as aggregate consumption.

In large part, the "better performance" of the CAPM and ICAPM comes from throwing away content. Again $m_{t+1} = \delta u'(c_{t+1})/u'(c_t)$ is there in any CAPM or ICAPM. The CAPM and ICAPM make predictions concerning consumption data that are wildly implausible, not only of admittedly poorly measured consumption data but any imaginable perfectly measured consumption data as well. For example, equation (84) says that the standard deviation of the wealth portfolio return equals the standard deviation of consumption growth. The latter is about 1% per year. All the miserable failures of the log-utility consumption-based model apply equally to the log utility CAPM. Finally, many "free parameters" of the models are not free parameters at all.

In sum, the poor performance of the consumption-based model is an important nut to chew on, not just a blind alley or failed attempt that we can safely disregard and go on about our business.

Identity of state variables

The ICAPM does not tell us the *identity* of the state variables $\mathbf{z}_t$, and many authors use the ICAPM as an obligatory citation to theory on the way to using factors composed of ad-hoc portfolios, leading Fama (1991) to characterize the ICAPM as a "fishing license." It really isn't: one could do a lot to insist that the factor-mimicking portfolios actually are the projections of some identifiable state variables on to the space of returns, and one could do a lot to make sure the candidate state variables really are plausible state variables for an explicitly stated optimization problem. For example, one could check that they actually do forecast something. The fishing license comes as much from habits of applying the theory as from the theory itself.

General equilibrium models

The CAPM and other models are really *general* equilibrium models. Looking at the derivation through general-equilibrium glasses, we have specified a set of linear technologies with returns R^i that do not depend on the amount invested. Some derivations make further assumptions, such as an initial capital stock, and no labor or labor income.

8.4 Arbitrage Pricing Theory (APT)

The APT: If a set of asset returns are generated by a linear factor model

$$R^i = E(R^i) + \sum_{j=1}^{N} \beta_{ij} \tilde{f}_j + \varepsilon^i$$

$$E(\varepsilon^i) = E(\varepsilon^i \tilde{f}_j) = 0.$$

Then (with additional assumptions) there is a discount factor m linear in the factors $m = a + \mathbf{b}'\mathbf{f}$ that prices the returns.

The APT starts from a statistical characterization. There is a big common component to stock returns: when the market goes up, most individual stocks also go up. Beyond the market, groups of stocks such as computer stocks, utilities, etc. move together. Finally, each stock's return has some completely idiosyncratic movement. This is a characterization of *realized* returns, *outcomes* or *payoffs*. The point of the APT is to start with this statistical characterization of *outcomes*, and derive something about *expected* returns or *prices*.

The intuition behind the APT is that the completely idiosyncratic movements in asset returns should not carry any risk prices, since investors can diversify them away by holding portfolios. Therefore, risk prices or expected returns on a security should be related to the security's covariance with the common components or "factors" only.

The job of this section is then 1) to describe a mathematical model of the tendency for stocks to move together, and thus to define the "factors" and residual idiosyncratic components, and 2) to think carefully about what it takes for the idiosyncratic components to have zero (or small) risk prices, so that only the common components matter to asset pricing.

There are two lines of attack for the second item. 1) If there were no residual, then we could price securities from the factors by *arbitrage* (really, by the law of one price, but the current distinction between law of one price and arbitrage came after the APT was named.) Perhaps we can extend this logic and show that if the residuals are *small*, they must have small risk prices. 2) If investors all hold well-diversified portfolios, then only variations in the factors drive consumption and hence marginal utility.

Much of the original appeal and marketing of the APT came from the first line of attack, the attempt to derive pricing implications *without* the economic structure required of the CAPM, ICAPM, or any other model derived as a specialization of the consumption-based model. In this section, I will first try to see how far we can in fact get with purely law of one price arguments. I will conclude that the answer is, "not very far," and that the most satisfactory argument for the APT is in fact just another specialization of the consumption-based model.

8.4.1 Factor structure in covariance matrices

I define and examine the factor decomposition

$$x^i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{f} + \varepsilon^i; \quad E(\varepsilon^i) = 0,\ E(\mathbf{f}\varepsilon^i) = 0$$

The factor decomposition is equivalent to a restriction on the payoff covariance matrix.

The APT models the tendency of asset payoffs (returns) to move together via a statistical *factor decomposition*

$$x^i = \alpha_i + \sum_{j=1}^{M} \beta_{ij} f_j + \varepsilon^i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{f} + \varepsilon^i. \tag{85}$$

The f's are the *factors,* the β are the *betas* or *factor loadings* and the ε are *residuals.* The terminology is unfortunate. A discount *factor* m, pricing *factors* $\mathbf{f}$ in $m = \mathbf{b}'\mathbf{f}$ and this *factor decomposition* (or *factor structure*) for returns are totally unrelated uses of the word "factor." Don't blame me, I didn't invent the terminology! The APT is conventionally written with $x^i =$ returns, but it ends up being much less confusing to use prices and payoffs.

It is a convenient and conventional simplification to fold the factor means into the constant, and write the factor decomposition with zero-mean factors $\tilde{f} \equiv f - E(f)$.

$$x^i = E(x^i) + \sum_{j=1}^{M} \beta_{ij} \tilde{f}_j + \varepsilon^i. \tag{86}$$

Remember that $E(x^i)$ is still just a statistical characterization, not yet the prediction of a model.

We can construct the factor decomposition as a regression equation. Define the β_{ij} as regression coefficients, and then the ε_i are uncorrelated with the factors by construction,

$$E(\varepsilon_i \tilde{f}_j) = 0.$$

The content — the assumption that keeps (86) from describing any arbitrary set of returns — is an assumption that the ε_i are *uncorrelated with each other*.

$$E(\varepsilon^i \varepsilon^j) = 0.$$

(More general versions of the model allow some limited correlation across the residuals but the basic story is the same.)

The factor structure is thus a restriction on the covariance matrix of payoffs. For example, if there is only one factor, then

$$cov(x^i, x^j) = E[(\beta_i \tilde{f} + \varepsilon^i)(\beta_j \tilde{f} + \varepsilon^j)] = \beta_i \beta_j \sigma^2(f) + \left\{ \begin{array}{ll} \sigma^2_{\varepsilon^i} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{array} \right. .$$

Thus, with $N =$ number of securities, the $N(N-1)/2$ elements of a variance-covariance matrix are described by N betas, and $N+1$ variances. A vector version of the same thing is

$$cov(\mathbf{x},\mathbf{x}') = \boldsymbol{\beta}\boldsymbol{\beta}'\sigma^2(f) + \left[\begin{array}{ccc} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \ddots \end{array}\right].$$

With more (orthogonalized) factors, one obtains

$$cov(\mathbf{x},\mathbf{x}') = \boldsymbol{\beta}_1\boldsymbol{\beta}_1'\sigma^2(f_1) + \boldsymbol{\beta}_2\boldsymbol{\beta}_2'\sigma^2(f_2) + \ldots + \text{(diagonal matrix)}$$

In all these cases, we describe the covariance matrix a singular matrix $\boldsymbol{\beta}\boldsymbol{\beta}'$ (or a sum of a few such singular matrices) plus a diagonal matrix.

If we know the factors we want to use ahead of time, say the market (value-weighted portfolio) and industry portfolios, we can estimate a factor structure by running regressions. Often, however, we don't know the identities of the factor portfolios ahead of time. In this case we have to use one of several statistical techniques under the broad heading of *factor analysis* **(**that's where the word "factor" came from in this context) to estimate the factor model. One can estimate a factor structure quickly by simply taking an eigenvalue decomposition of the covariance matrix, and then setting small eigenvalues to zero. More formal estimates can come from maximum likelihood.

8.4.2 Exact factor pricing

With no error term,

$$x^i = E(x^i)1 + \boldsymbol{\beta}_i'\tilde{\mathbf{f}}.$$

implies

$$p(x^i) = E(x^i)p(1) + \boldsymbol{\beta}_i'p(\tilde{\mathbf{f}})$$

and thus

$$m = \mathbf{b}'\mathbf{f};\ p(x^i) = E(mx^i)$$

$$E(R^i) = R^f + \boldsymbol{\beta}_i'\boldsymbol{\lambda}.$$

using only the law of one price.

Suppose that there are no idiosyncratic terms ε_i. This is called an *exact factor model.* Now look again at the factor decomposition,

$$x^i = E(x^i)1 + \boldsymbol{\beta}_i'\tilde{\mathbf{f}}.$$

This initially statistical decomposition expresses the payoff in question as a *portfolio* of the factors and a constant (risk-free payoff). Thus, the price can only depend on the prices of the factors f,

$$p(x^i) = E(x^i)p(1) + \boldsymbol{\beta}_i' p(\tilde{\mathbf{f}}). \tag{87}$$

The *law of one price* assumption lets you take prices of right and left sides.

If the factors are returns, their prices are 1. If the factors are not returns, their prices are free parameters which can be picked to make the model fit as well as possible. Since there are fewer factors than payoffs, this procedure is not vacuous. (Recall that the prices of the factors are related to the λ in expected return beta representations. λ is determined by the expected return of a return factor, and is a free parameter for non-return factor models.)

We are really done, but the APT is usually stated as "there is a *discount factor* linear in $\mathbf{f}$ that prices returns R^i," or "there is an expected return-beta representation with $\mathbf{f}$ as factors." Therefore, we should take a minute to show that the rather obvious relationship (87) between prices is equivalent to discount factor and expected return statements.

Assuming only the law of one price, we know there is a discount factor m linear in factors that price the factors. We usually call it x^*, but call it f^* here to remind us that it prices the factors f. As with x^*, $f^* = p(\mathbf{f})' E(\mathbf{ff}')^{-1}\mathbf{f}$ satisfies $p(\mathbf{f}) = E(f^*\mathbf{f})$. If it prices the factors, it must price any portfolio of the factors; hence $f^* = \mathbf{b}'\mathbf{f}$ prices all payoffs x^i that follow the factor structure.

We could now go from m linear in the factors to an expected return-beta model using the above theorems that connect the two representations. But there is a more direct and very slick connection. Start with (87), specialized to returns $x^i = R^i$ and of course $p(R^i) = 1$. Use $p(1) = 1/R^f$ and solve for expected return as

$$E(R^i) = R^f + \boldsymbol{\beta}_i' \left[-R^f p(\tilde{\mathbf{f}}) \right] = R^f + \boldsymbol{\beta}_i' \boldsymbol{\lambda}.$$

The last equality defines $\boldsymbol{\lambda}$. Expected returns are linear in the betas, and the constants (λ) are related to the prices of the factors. In fact, this is the same definition of $\boldsymbol{\lambda}$ that we arrived at above connecting $m = \mathbf{b}'\mathbf{f}$ to expected return-beta models.

8.4.3 Approximate APTs using the law of one price.

Attempts to extend the exact factor model to an approximate factor pricing model when errors are "small," or markets are "large," still only using law of one price.

For fixed m, the APT gets better and better as R^2 or the number of assets increases.

However, for any fixed R^2 or size of market, the APT can be arbitrarily bad.

These observations say that we must go beyond the law of one price to derive factor pricing models.

Actual returns do not display an exact factor structure. There is some idiosyncratic or residual risk; we cannot exactly replicate the return of a given stock with a portfolio of a few large factor portfolios. However, the idiosyncratic risks are often "small." For example, factor model regressions of the form (85) often have very high R^2, especially when portfolios rather than individual securities are on the left hand side. And the residual risks are still idiosyncratic: Even if they are a large price of an individual security's variance, they should be a small contributor to the variance of well diversified portfolios. Thus, there is reason to hope that the APT holds approximately. Surely, if the residuals are "small" and/or "idiosyncratic," the price of an asset can't be "too different" from the price predicted from its factor content?

To think about these issues, start again from a factor structure, but this time put in a residual,

$$x^i = E(x^i)1 + \boldsymbol{\beta}_i'\tilde{\mathbf{f}} + \varepsilon^i$$

Again take prices of both sides,

$$p(x^i) = E(x^i)p(1) + \boldsymbol{\beta}_i' p(\tilde{\mathbf{f}}) + E(m\varepsilon^i)$$

Now, what can we say about the price of the residual $p(\varepsilon^i) = E(m\varepsilon^i)$?

Figure 17 illustrates the situation. Portfolios of the factors span a payoff space, the line connecting f^* and $\beta_i' f$ in the figure. The payoff we want to price, x^i is not in that space, since the residual ε^i is not zero. A discount factor f^* prices the factors, and the space of all discount factors that price the factors is the line m orthogonal to f^*. The residual is orthogonal to the factor space, since it is a regression residual, and to f^* in particular, $E(f^*\varepsilon^i) = 0$. This means that f^* assigns zero price to the residual. But the other discount factors on the m line are *not* orthogonal to ε^i, so generate non-zero price for the residual ε^i. As we sweep along the line of discount factors m that price the f, in fact, we generate every price from $-\infty$ to ∞ for the residual. Thus, the law of one price does not nail down the price of the residual ε^i and hence the price or expected return of x^i.

Limiting arguments

We would like to show that the price of x^i has to be "close to" the price of $\boldsymbol{\beta}_i'\mathbf{f}$. One notion of "close to" is that in some appropriate limit the price of x^i converges to the price of $\boldsymbol{\beta}_i'\mathbf{f}$. "Limit" means, of course, that you can get arbitrarily good accuracy by going far enough in the direction of the limit (for every $\varepsilon > 0$ there is a $\delta....$). Thus, establishing a limit result is a way to argue for an approximation.

Here is one theorem that seems to imply that the APT should be a good approximation for portfolios that have high R^2on the factors. I state the argument for the case that there is a constant factor, so the constant is in the f space and $E(\varepsilon^i) = 0$. The same ideas work in the

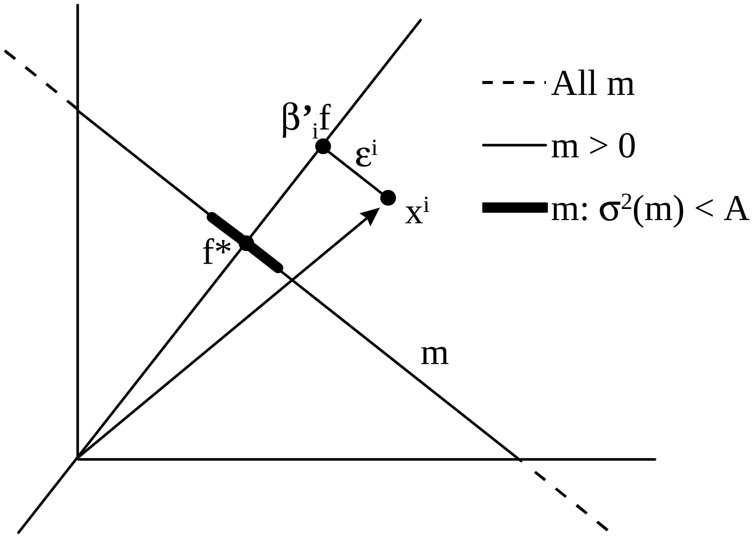


Figure 17. Approximate arbitrage pricing.

less usual case that there is no constant factor, using second moments in place of variance.

Theorem 3 *Fix a discount factor m that prices the factors. Then, as $var(\varepsilon^i) \rightarrow 0$, $p(x^i) \rightarrow p(\boldsymbol{\beta}_i'\mathbf{f})$.*

Graphical argument: $E(\varepsilon^i) = 0$ so $var(\varepsilon^i) = E(\varepsilon^{i2}) = ||\varepsilon^i||$. Thus, as the size of the ε^i vector in Figure 17 gets smaller, x^i gets closer and closer to $\boldsymbol{\beta}_i'\mathbf{f}$. For any fixed m, the induced pricing function (lines perpendicular to the chosen m) is continuous. Thus, as x^i gets closer and closer to $\boldsymbol{\beta}_i'\mathbf{f}$, its price gets closer and closer to $\boldsymbol{\beta}_i'\mathbf{f}$.

Regression interpretation. Remember, the factor model is defined as a regression, so

$$var(x^i) = var(\boldsymbol{\beta}_i'\mathbf{f}) + var(\varepsilon^i)$$

Thus, the variance of the residual is related to the regression R^2.

$$\frac{var(\varepsilon^i)}{var(x^i)} = 1 - R^2$$

The theorem says that as $R^2 \rightarrow 1$, the price of the residual goes to zero.

We were hoping for some connection between the fact that the risks are *idiosyncratic* and factor pricing. Even if the idiosyncratic risks are a large part of the payoff at hand, they are a small part of a well-diversified portfolio. The next theorem shows that portfolios with high R^2 don't have to happen by chance; well-diversified portfolios will always have this characteristic.

Theorem 4 *As the number of primitive assets increases, the R^2 of well-diversified portfolios increases to 1.*

Proof: Start with an equally weighted portfolio

$$x^p = \frac{1}{N}\sum_{i=1}^{N} x^i.$$

Going back to the factor decomposition (85) for each individual asset x^i, the factor decomposition of x^p is

$$x^p = \frac{1}{N}\sum_{i=1}^{N}\left(a_i + \boldsymbol{\beta}_i'\mathbf{f} + \varepsilon^i\right) = \frac{1}{N}\sum_{i=1}^{N} a_i + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_i'\mathbf{f} + \frac{1}{N}\sum_{i=1}^{N}\varepsilon^i = a^p + \boldsymbol{\beta}_p^{'}\mathbf{f} + \varepsilon^p.$$

The last equality defines notation $\alpha^p, \boldsymbol{\beta}_p, \varepsilon^p$. But

$$var(\varepsilon^p) = var\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon^i\right)$$

So long as the variance of ε^i are bounded, and given the factor assumption $E(\varepsilon^i\varepsilon^j) = 0$,

$$\lim_{N\to\infty} var(\varepsilon^p) = 0.$$

Obviously, the same idea goes through so long as the portfolio spreads some weight on all the new assets, i.e. so long as it is "well-diversified." ■

These two theorems can be interpreted to say that the APT holds approximately (in the usual limiting sense) for portfolios that either naturally have high R^2, or well-diversified portfolios in large enough markets. We have only used the law of one price.

Law of one price arguments fail

Now, let me pour some cold water on these results. I *fixed* m and then let other things take limits. The flip side is that for any nonzero residual ε^i, no matter how small, we can pick a discount factor m that prices the factors and assigns *any* price to x^i!

Theorem 5 *For any nonzero residual ε^i there is a discount factor that prices the factors* $\mathbf{f}$ *(consistent with the law of one price) and that assigns* any *desired price in $(-\infty, \infty)$ to the return R^i.*

So long as $||\varepsilon^i|| > 0$, as we sweep the choice of m along the dashed line, the inner product of m with ε^i and hence x^i varies from $-\infty$ to ∞.

Thus, for a given size $R^2 < 1$, or a given finite market, the law of one price says absolutely nothing about the prices of payoffs that do not exactly follow the factor structure. The law of one price says that two ways of constructing the same portfolio must give the same price. If the residual is not exactly zero, there is no way of replicating the payoff x^i from the factors and no way to infer anything about the price of x^i from the price of the factors.

I think the contrast between this theorem and those of the last subsection accounts for most of the argument over the APT. If you fix m and take limits of N or ε, the APT gets arbitrarily good. But if you fix N or ε, as one does in any application, the APT can get arbitrarily bad as you search over possible m.

The lesson I learn is that the effort to *extend* prices from an original set of securities ($\mathbf{f}$ in this case) to new payoffs that are not exactly spanned by the original set of securities, using only the law of one price, is fundamentally doomed. To extend a pricing function, you need to add some restrictions beyond the law of one price.

Beyond the law of one price: arbitrage and Sharpe ratios.

So far, we have used only the law of one price restriction that there is an m. Perhaps we can do better by imposing the no-arbitrage restriction that m must be positive. Graphically, we are now restricted to the solid m line in Figure 17. Since that line only extends a finite amount, restricting us to strictly positive $m's$ gives rise to finite upper and lower *arbitrage* *bounds* on the price of ε^i and hence x^i. (The word *arbitrage bounds* comes from option pricing, and we will see these ideas again in that context. If this idea worked, it would restore the APT to "arbitrage pricing" rather than "law of one-pricing".)

Alas, in applications of the APT (as often in option pricing), the arbitrage bounds are too wide to be of much use. The positive discount factor restriction is equivalent to saying "if portfolio A gives a higher payoff than portfolio B in *every state of nature,* then the price of A must be higher than the price of B." Since stock returns and factors are continuously distributed, not two-state distributions as I have graphed for figure 17, there typically are no strictly dominating portfolios, so adding $m > 0$ does not help.

I think it is possible to continue in this line and derive an approximate APT that is useful in finite markets with $R^2 < 1$. The issue is, can we rule out the wild discount factors—way out on the edges of the discount factor line—that one must invoke to justify a price of x^i "far" from the price of $\boldsymbol{\beta}'\mathbf{f}$. We have found that the law of one price and no-arbitrage do not rule out such wild prices. But surely we can rule out such prices without taking the opposite extreme of completely specifying the discount factor model, i.e. start with the consumption-based model?

One obvious possibility is to restrict the *variance* and hence the size ($||m|| = E(m^2) = \sigma^2(m) + E(m)^2 = \sigma^2(m) + 1/R^{f2}$) of the discount factor. Figure 17 includes a plot of the discount factors with limited variance, size, or length in the geometry of that figure. The restricted range of discount factors produces a restricted range of prices for x^i. We obtain upper and lower price *bounds* for the price of x^i in terms of the factor prices, and the bounds shrink to $\boldsymbol{\beta}'p(\mathbf{f})$ as the allowed variance of the discount factor shrinks. Precisely, then, we solve the problem

$$\min_{\{m\}} \text{ (or } \max_{\{m\}}\text{)} \;\; p(x^i) = E(mx^i) \;\; s.t. \; E(m\mathbf{f}) = p(\mathbf{f}), \; m \geq 0, \; \sigma^2(m) \leq A$$

Limiting the variance of the discount factor is of course the same as limiting the maximum *Sharpe ratio* **(**mean / standard deviation of excess return) available from portfolios of the factors and x^i. Recall that

$$\frac{E\left(R^e\right)}{\sigma(R^e)} \leq \frac{\sigma(m)}{E(m)}.$$

Thus, Saá-Requejo and I (1996) dub this idea "good-deal" pricing, as an extension of "arbitrage pricing." Limiting $\sigma(m)$ rules out "good deals" as well as pure arbitrage opportunities. Though a bound on Sharpe ratios or discount factor volatility is not a totally preference-free concept, it clearly imposes a great deal less structure than the CAPM or ICAPM which are essentially full general equilibrium models. Ross (1976) included this suggestion in his original APT paper, though it seems to have disappeared from the literature since then in the failed effort to derive an APT from the law of one price alone. Ross pointed out that deviations from factor pricing could provide very high Sharpe ratio opportunities, which seem implausible though not violations of the law of one price.

If we impose a good-deal bound, we obtain well-behaved limits, that do not depend on the order of "for all" and "there exists." For given R^2, all discount factors satisfying the good-deal bound produce price bounds, and the price bounds shrink as the R^2 shrinks or as the good-deal bound shrinks. I describe good-deal pricing in more detail below in an option-pricing context.

8.5 APT vs. ICAPM

A factor structure in the covariance of returns or high R^2 in regressions of returns on factors are sufficient (APT) but not necessary (ICAPM) for factor pricing.

Differing inspiration for factors.

The disappearance of absolute pricing.

The APT and ICAPM stories are often confused. Factor structure can employ factor pricing (APT), but factor pricing does not require a factor structure. In the ICAPM there is no presumption that factors $\mathbf{f}$ in a pricing model $m = \mathbf{b}'\mathbf{f}$ describe the covariance matrix of returns. The factors don't have to be orthogonal or i.i.d. either. High R^2 in time-series regressions of the returns on the factors may imply factor pricing (APT), but again are not necessary. The regressions of returns on factors can have as low an R^2 as one wishes in the ICAPM.

The biggest difference between APT and ICAPM for empirical work is in the inspiration for factors. The APT suggests that one start with a statistical analysis of the covariance matrix of returns and find portfolios that characterize common movement. The ICAPM suggests that one start by thinking about good proxies for marginal utility growth, or state variables that describe the conditional distribution of future asset returns and non-asset income.

The difference between the derivations of factor pricing models, and in particular an approximate law-of-one-price basis vs. a proxy for marginal utility basis seems not to have had much impact on practice. In practice, we just test models $m = \mathbf{b}'\mathbf{f}$ and rarely worry about derivations. The best evidence for this view is the introductions of famous papers. Chen, Roll and Ross (1986) describe one of the earliest popular multifactor models, using industrial production and inflation as some of the main factors. They do not even present a factor decomposition of test asset returns, or the time-series regressions. A reader might well categorize the paper as much closer to an ICAPM. Fama and French (199x) describe the currently most popular multifactor model, and their introduction describes it as an ICAPM in which the factors are state variables. But the factors are sorted on size and book/market just like the test assets, the time-series R^2 are all above 90%, and much of the explanation involves "common movement" in test assets captured by the factors. A a reader might well categorize the model as much closer to an APT.

Chapter 8 Factor pricing models

In the first chapter, I made a distinction between *relative* pricing and *absolute* pricing. In the former, we price one security given the prices of others, while in the latter, we price each security by reference to fundamental sources of risk. The factor pricing stories are interesting in that they start with a nice absolute pricing model, the consumption-based model, and throw out enough information to end up with relative models. The CAPM prices R^i *given* the market, but throws out the consumption-based model's description of where the market return came from.

PART II
Estimating and evaluating asset pricing models

Our first task in bringing an asset pricing model to data is to *estimate* the free parameters. Examples of such parameters are β and γ in $m = \beta(c_{t+1}/c_t)^{-\gamma}$, or $\mathbf{b}$ in $m = \mathbf{b}'\mathbf{f}$. Then we want to evaluate the model. Is it a good model or not? Is another model better?

Statistical analysis helps in model evaluation by providing a *distribution theory* for numbers we create from the data. A distribution theory answers the question, if we generate artificial data over and over again from a statistical model, generating a number from the data each time, what is the resulting probability distribution of that number? In particular, we are interested in a distribution theory for the estimated parameters, and for the pricing errors, which helps us to judge whether pricing errors are just bad luck or if they indicate a failure of the model. We also will want to generate distributions for statistics that compare one model to another, or provide other interesting evidence, to judge how much sample luck affects those calculations.

All of the statistical methods I discuss in this part achieve exactly these ends. They give methods for estimating free parameters; they provide a distribution theory for those parameters, and they provide statistics for model evaluation, in particular a quadratic form of pricing errors in the form $\hat{\boldsymbol{\alpha}}'V^{-1}\hat{\boldsymbol{\alpha}}$.

I start by focusing on the GMM approach. Then I consider traditional regression tests and their maximum likelihood formalization. I emphasize the fundamental similarities between these three methods, as I emphasized the similarity between $p = E(mx)$, expected return-beta models, and mean-variance frontiers. A concluding essay highlights the differences between the methods and argues that the GMM approach will be most useful for most empirical work in the future.

I use the word *evaluation* rather than *test* deliberately. Statistical hypothesis testing is one very small part of the process by which we evaluate and refine asset pricing models, or discard them in favor of new ones. Statistical tools exist only to answer the sampling distribution questions in this process. Many models are kept that give economically small but statistically significant pricing errors, and many more models are quickly forgotten that have statistically insignificant but economically large pricing errors, or just do not tell as clean a story.

Chapter 9. GMM estimation and testing of asset pricing models

The basic idea in the GMM approach is very straightforward. The asset pricing model predicts

$$E(p_t) = E\left[m(\text{data}_{t+1}, \text{parameters})\; x_{t+1}\right]. \tag{88}$$

The most natural way to check this prediction is to examine sample averages, i.e. to calculate

$$\frac{1}{T}\sum_{t=1}^{T} p_t \text{ and } \frac{1}{T}\sum_{t=1}^{T}\left[m(\text{data}_{t+1}, \text{parameters})\; x_{t+1}\right]. \tag{89}$$

GMM *estimates* the parameters by making these sample averages as close to each other as possible. It works out a distribution theory for those estimates. This distribution theory is a generalization of the simplest exercise in statistics: the distribution of the sample mean. Then, it suggests that we *evaluate* the model by looking at how close the sample averages are to each other, or equivalently by looking at the pricing errors. It gives a statistical *test* of the hypothesis that the underlying population means are in fact zero.

9.1 GMM in explicit discount factor models.

It's easiest to start our discussion of GMM in the context of an explicit discount factor model, such as the consumption-based model. I treat the special structure of linear factor models later. I start with the basic classic recipe as given by Hansen and Singleton (1982) and then explore the intuition behind it and useful variants.

9.1.1 Recipe

Definitions

$$\begin{aligned}\mathbf{u}_{t+1}(\mathbf{b}) &\equiv m_{t+1}(\mathbf{b})x_{t+1} - p_t \\ \mathbf{g}_T(\mathbf{b}) &\equiv E_T\left[\mathbf{u}_t(\mathbf{b})\right] \\ S &\equiv \sum_{j=-\infty}^{\infty} E\left[\mathbf{u}_t(\mathbf{b})\ \mathbf{u}_{t-j}(\mathbf{b})'\right]\end{aligned}$$

GMM estimate

$$\hat{\mathbf{b}}_2 = argmin_{\mathbf{b}}\ \mathbf{g}_T(\mathbf{b})'\hat{S}^{-1}\mathbf{g}_T(\mathbf{b}).$$

Standard errors

$$var(\hat{\mathbf{b}}_2) = \frac{1}{T}(D'S^{-1}D)^{-1};\ \ D \equiv \frac{\partial g_T(\mathbf{b})}{\partial \mathbf{b}}$$

Test of the model (“overidentifying restrictions”)

$$TJ_T = T\min\left[\mathbf{g}_T(\mathbf{b})'S^{-1}\mathbf{g}_T(\mathbf{b})\right] \sim \chi^2(\#\text{moments} - \#\text{parameters}).$$

Discount factor models involve some unknown parameters as well as data, so I write $m_{t+1}(\mathbf{b})$ to remind ourselves of the dependence on parameters. For example, if $m_{t+1} = \beta(c_{t+1}/c_t)^{-\gamma}$, then $\mathbf{b} \equiv [\beta\ \gamma]'$. I write $\hat{\mathbf{b}}$ to denote estimates when it is important to distinguish estimated from other values.

Again, any asset pricing model implies

$$E(\mathbf{p}_t) = E\left[m_{t+1}(\mathbf{b})\mathbf{x}_{t+1}\right]. \tag{90}$$

It’s easiest to write this equation in the form $E(\cdot) = 0$

$$E\left[m_{t+1}(\mathbf{b})\mathbf{x}_{t+1} - \mathbf{p}_t\right] = 0. \tag{91}$$

I use boldface for $\mathbf{x}$ and $\mathbf{p}$ because these objects are typically vectors; we typically check whether a model for m can price a number of assets simultaneously. Equations (91) are often called the *moment conditions*.

It’s convenient to define the *errors* $\mathbf{u}_t(\mathbf{b})$ as the object whose mean should be zero,

$$\mathbf{u}_{t+1}(\mathbf{b}) = m_{t+1}(\mathbf{b})x_{t+1} - p_t$$

Given values for the parameters $\mathbf{b}$, we could construct a time series on $\mathbf{u}_t$ and look at its mean.

Define $\mathbf{g}_T(\mathbf{b})$ as the sample mean of the $\mathbf{u}_t$ errors, when the parameter vector is $\mathbf{b}$ in a sample of size T:

$$\mathbf{g}_T(\mathbf{b}) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t(\mathbf{b}) = E_T\left[\mathbf{u}_t(\mathbf{b})\right].$$

The last equality introduces the handy notation E_T for sample means,

$$E_T(\cdot) = \frac{1}{T}\sum_{t=1}^{T}(\cdot).$$

(It might make more sense to denote these quantities $\hat{E}$ and $\hat{g}$ to denote estimates, as I do elsewhere. However, Hansen's T subscript notation is so widespread that doing so would cause more confusion than it solves.)

The *first stage estimate* of $\mathbf{b}$ minimizes a quadratic form of the sample mean of the errors,

$$\hat{\mathbf{b}}_1 = argmin_{\{\hat{b}\}} \ \mathbf{g}_T(\hat{\mathbf{b}})'W\mathbf{g}_T(\hat{\mathbf{b}})$$

for some arbitrary matrix W (usually, $W = I$). This estimate is consistent, asymptotically normal, and you can and often should stop here, as I explain below.

Using $\hat{\mathbf{b}}_1$, form an estimate $\hat{S}$ of

$$S \equiv \sum_{j=-\infty}^{\infty} E\left[\mathbf{u}_t(\mathbf{b})\ \mathbf{u}_{t-j}(\mathbf{b})'\right].$$

(Below I discuss various interpretations of and ways to construct this estimate.) Form a *second stage* estimate $\hat{\mathbf{b}}_2$ using the matrix $\hat{S}$ in the quadratic form,

$$\hat{\mathbf{b}}_2 = argmin_{\mathbf{b}}\ \mathbf{g}_T(\mathbf{b})'\hat{S}^{-1}\mathbf{g}_T(\mathbf{b}).$$

$\hat{\mathbf{b}}_2$ is a consistent, asymptotically normal, and asymptotically efficient estimate of the parameter vector $\mathbf{b}$. "Efficient" means that it has the smallest variance-covariance matrix among all estimators that set different linear combinations of $\mathbf{g}_T(\mathbf{b})$ to zero.

The variance-covariance matrix of $\hat{\mathbf{b}}_2$ is

$$var(\hat{\mathbf{b}}_2) = \frac{1}{T}(D'S^{-1}D)^{-1}$$

where

$$D \equiv \frac{\partial g_T(\mathbf{b})}{\partial \mathbf{b}}$$

or, more explicitly,

$$D = E_T\left(\frac{\partial \mathbf{u}_{t+1}(\mathbf{b})}{\partial \mathbf{b}}\right)\Bigg|_{\mathbf{b}=\hat{\mathbf{b}}} = E_T\left(\frac{\partial}{\partial \mathbf{b}}\left[(m_{t+1}(\mathbf{b})\mathbf{x}_{t+1} - p_t)\right]\right)\Bigg|_{\mathbf{b}=\hat{\mathbf{b}}}$$

This variance-covariance matrix can be used to test whether a parameter or group of parameters are equal to zero, via

$$\frac{\hat{b}_i}{\sqrt{var(\hat{\mathbf{b}})_{ii}}} \sim N(0,1)$$

and

$$\hat{\mathbf{b}}_j \left[var(\hat{\mathbf{b}})_{jj} \right]^{-1} \hat{\mathbf{b}}_j \sim \chi^2(\#\text{included } b's)$$

where $\mathbf{b}_j$ =subvector, $var(\mathbf{b})_{jj}$ =submatrix.

Finally, the *test of overidentifying restrictions* is a test of the overall fit of the model. It states that T times the minimized value of the second-stage objective is distributed χ^2 with degrees of freedom equal to the number of moments less the number of estimated parameters.

$$TJ_T = T\min \left[\mathbf{g}_T(\mathbf{b})' S^{-1} \mathbf{g}_T(\mathbf{b}) \right] \sim \chi^2(\#\text{moments} - \#\text{parameters}).$$

See Hansen (1982) or Ogaki (1993) for many important statistical assumptions. The most important is that $m,\ p,$ and x must be *stationary* random variables. so that time-series averages converge to population means.

9.2 Interpreting GMM

Notation.

Stationarity and choice of units.

Forecast errors and instruments.

$g_T(\mathbf{b})$ is a pricing error.

GMM picks parameters to minimize pricing errors and evaluates the model by the size of pricing errors.

The optimal weighting matrix tells you to pay attention to the assets with best-measured pricing errors.

Notation; instruments and returns

Most of the effort involved with GMM is simply mapping a given problem into the very general notation. The equation

$$E\left[m_{t+1}(\mathbf{b})\mathbf{x}_{t+1} - \mathbf{p}_t\right] = 0$$

can capture a lot. Here, I translate it for the most common case.

We often test asset pricing models using returns, in which case the moment conditions are

$$E\left[m_{t+1}(\mathbf{b})\mathbf{R}_{t+1} - 1\right] = 0.$$

It is common to add *instruments* as well. Mechanically, you can multiply both sides of

$$1 = E_t\left[m_{t+1}(\mathbf{b})\mathbf{R}_{t+1}\right]$$

by any variable z_t observed at time t before taking unconditional expectations, resulting in

$$E(z_t) = E\left[m_{t+1}(\mathbf{b})\mathbf{R}_{t+1}z_t\right]$$

or

$$0 = E\left\{\left[m_{t+1}(\mathbf{b})\mathbf{R}_{t+1} - 1\right]z_t\right\}. \tag{92}$$

If payoffs are generated by a vector of two returns $\mathbf{R} = [R^a \;\; R^b]'$ and one instrument z, equation (92) might look like

$$E\left\{\left[\begin{array}{c} m_{t+1}(\mathbf{b})\;R^a_{t+1} \\ m_{t+1}(\mathbf{b})\;R^b_{t+1} \\ m_{t+1}(\mathbf{b})\;R^a_{t+1}z_t \\ m_{t+1}(\mathbf{b})\;R^b_{t+1}z_t \end{array}\right] - \left[\begin{array}{c} 1 \\ 1 \\ z_t \\ z_t \end{array}\right]\right\} = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \end{array}\right].$$

Using the Kronecker product $\otimes$ meaning "multiply every element by every other element" we can denote the same relation compactly by

$$E\left\{\left[m_{t+1}(\mathbf{b})\;\mathbf{R}_{t+1} - 1\right]\otimes\mathbf{z}_t\right\} = 0, \tag{93}$$

or, emphasizing the managed-portfolio interpretation and $p = E(mx)$ notation,

$$E\left[m_{t+1}(\mathbf{b})(\mathbf{R}_{t+1}\otimes\mathbf{z}_t) - (\mathbf{1}\otimes\mathbf{z}_t)\right] = 0.$$

Stationarity

Stationarity is the most important statistical requirement for consistency and the GMM distribution theory. ("Stationary" of often misused to mean constant, or i.i.d.. The statistical definition of stationarity is that the joint distribution of x_t, x_{t-j} depends only on j and not on t.) Sample averages must converge to population means as the sample size grows, and stationarity implies this result.

This step usually amounts to a choice of sensible units. For example, though we could express the pricing of a stock as

$$p_t = E_t\left[m_{t+1}(d_{t+1} + p_{t+1})\right]$$

it would not be wise to do so. For stocks, p and d rise over time and so are typically not stationary; their unconditional means are not defined. It is better to divide by p_t and express the model as

$$1 = E_t \left[m_{t+1} \frac{d_{t+1} + p_{t+1}}{p_t} \right] = E_t \left(m_{t+1} R_{t+1} \right)$$

The stock *return* is plausibly stationary.

Dividing by dividends is an alternative and I think underutilized way to achieve stationarity:

$$\frac{p_t}{d_t} = E_t \left[m_{t+1} \left(1 + \frac{p_{t+1}}{d_{t+1}} \right) \frac{d_{t+1}}{d_t} \right].$$

Now we map $\left(1 + \frac{p_{t+1}}{d_{t+1}}\right) \frac{d_{t+1}}{d_t}$ into x_{t+1} and $\frac{p_t}{d_t}$ into p_t. This formulation allows us to focus on *prices* rather than one-period returns.

Bonds are a claim to a dollar, so bond prices do not grow over time. Hence, it might be all right to examine

$$p_t^b = E(m_{t+1} \ 1)$$

with no transformations.

Stationarity is not always a black and white question in practice. As variables become "less stationary", as they experience longer and longer swings in a sample, the asymptotic distribution can becomes a less reliable guide to a finite-sample distribution. For example, the level of interest rates is surely a stationary variable in a fundamental sense: it was 6% in ancient Babylon, about 6% in 14th century Italy, and about 6% again today. Yet it takes very long swings away from this unconditional mean, moving slowly up or down for even 20 years at a time. The asymptotic distribution theory of some estimators will be particularly bad approximation to the correct finite sample distribution theory in such a case.

It is also important to choose *test assets* in a way that is stationary. For example, individual stocks change character over time, increasing or decreasing size, exposure to risk factors, leverage, and even nature of the business. For this reason, it is common to sort stocks into portfolios based on characteristics such as betas, size, book/market ratios, industry and so forth. The statistical characteristics of the *portfolio* returns may be much more stationary than the characteristics of individual securities, which float in and out of the various portfolios.

Forecast errors and instruments

The asset pricing model says that, although expected *returns* can vary across time and assets, expected *discounted* returns should always be the same, 1. The error $u_{t+1} = m_{t+1} R_{t+1} - 1$ is the ex-post discounted return. $u_{t+1} = m_{t+1} R_{t+1} - 1$ represents a *forecast error*. Like any forecast error, u_{t+1} should be conditionally and unconditionally mean zero.

In an econometric context, z is an *instrument* because it is uncorrelated with the error $\mathbf{u}_{t+1}$. $E(z_t\mathbf{u}_{t+1})$ is the numerator of a regression coefficient of $\mathbf{u}_{t+1}$ on z_t; thus adding instruments basically checks that the error or ex-post discounted return is unforecastable by linear regressions.

If an asset's return is higher than predicted when z_t is unusually high, but not on average, scaling by z_t will pick up this feature of the data. Then, the moment condition checks that the discount rate is unusually low at such times, or that the conditional covariance of the discount rate and asset return moves sufficiently to justify the high conditionally expected return.

As I explained in Chapters 2 and 7, adding instruments can also be interpreted as including the returns of managed portfolios, strategies that put more or less money into assets as linear functions of the information variable z.

So far I have been careful to say that $E(p) = E(mx)$ is *an* implication of the model. As chapter 7 emphasizes, adding instruments is in principle able to capture *all* of the model's predictions.

Pricing errors

The moment conditions are

$$g(\mathbf{b}) = E\left[m_{t+1}(\mathbf{b})x_{t+1} - p_t\right] = E\left[m_{t+1}(\mathbf{b})x_{t+1}\right] - E\left[p_t\right].$$

Thus, each moment is the difference between actual ($E(p)$) and predicted ($E(mx)$) price, or *pricing error*.

In the language of expected returns, recall that $1 = E(mR)$ can be translated to a predicted expected return,

$$E(R) = \frac{1}{E(m)} - \frac{cov(m,R)}{E(m)}.$$

Therefore, we can write the pricing error as

$$\begin{aligned} g(\mathbf{b}) &= E(mR) - 1 = E(m)\left(E(R) - \frac{1}{E(m)} + \frac{cov(m,R)}{E(m)}\right) \\ g(\mathbf{b}) &= \frac{1}{R^f}(\text{actual mean return - predicted mean return.}) \end{aligned}$$

Similarly, if we express the model in expected return-beta language,

$$E(R^i) = \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{\lambda}$$

then the GMM objective is proportional to the Jensen's alpha measure of mis-pricing,

$$g(\mathbf{b}) = \frac{1}{R^f}\alpha_i.$$

Thus, GMM picks parameters to make *pricing errors* as small as possible, and tests the model by the size of its pricing errors.

First-stage estimates

If we could, we'd pick $\mathbf{b}$ to make every element of $g_T(\mathbf{b}) = \mathbf{0}$ — to have the model price assets perfectly in sample. However, there are usually more moment conditions (returns times instruments) than there are parameters. There should be, because theories with as many free parameters as facts (moments) are pretty vacuous. Thus, we choose $\mathbf{b}$ to make $g_T(\mathbf{b})$ as small as possible. The easiest way to make a vector such as $g_T(\mathbf{b})$ "small" is to minimize a quadratic form,

$$\min_{\{\mathbf{b}\}} \mathbf{g}_T(\mathbf{b})' W \mathbf{g}_T(\mathbf{b}). \tag{94}$$

W is a *weighting matrix* that tells us how much attention to pay to each moment, or how to trade off doing well in pricing one asset or linear combination of assets vs. doing well in pricing anther. For example, $W = I$ says to treat all assets symmetrically. In this case, the objective is the sum of squared pricing errors.

The sample pricing error $\mathbf{g}_T(\mathbf{b})$ may be a *nonlinear* function of $\mathbf{b}$. Thus, you may have to use a numerical search to find the value of $\mathbf{b}$ that minimizes the objective in (94). However, since the objective is locally quadratic, the search is usually straightforward.

Second-stage estimates

What weighting matrix should you use? You might start with $W = I$, i.e., "try to price all assets equally well". This is an example of an *economically interesting* metric. You might start with different elements on the diagonal of W if you think some assets are more interesting or informative than others. In particular, a first-stage W that is not the identity matrix can be used to offset differences in units between the moments.

However, some asset returns may have much more variance than other assets. For those assets, $g_T = E_T(m_t R_t - 1)$ will be a much less accurate measurement of $E(mR-1)$, since it will vary more from sample to sample. Hence, one might think of paying less attention to pricing errors from assets with high return variance. One could implement this idea by using a W matrix composed of inverse variances of $E_T(m_t R_t - 1)$ on the diagonal. More generally, since asset returns are correlated, one might think of using the covariance matrix of $E_T(m_t R_t - 1)$. This weighting matrix pays most attention to linear combinations of moments about which the data set at hand has the most information. Hence it is a *statistical metric* for judging how "small" the moments g_T are. This idea is exactly the same as heteroskedasticity and cross-correlation corrections that lead you from OLS to GLS in linear regressions.

The covariance matrix of $g_T = E_T(\mathbf{u}_{t+1})$ is the variance of a sample mean. Exploiting the fact that $E(\mathbf{u}_t) = 0$, and that $\mathbf{u}_t$ is stationary so $E(u_1 u_2) = E(u_t u_{t+1})$ depends only on the distance between the two u's and not on time itself, we have

$$\begin{aligned} var(g_T) &= var\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_{t+1}\right) = E\left[\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_{t+1}\right)^2\right] \\ &= \frac{1}{T^2}\left[TE(\mathbf{u}_t\mathbf{u}_t') + (T-1)\left(E(\mathbf{u}_t\mathbf{u}_{t-1}') + E(\mathbf{u}_t\mathbf{u}_{t+1}')\right)\right) + ...\right] \end{aligned} \tag{95}$$

As $T \to \infty,\ (T-j)/T \to 1$, so

$$var(g_T) \to \frac{1}{T}\sum_{j=-\infty}^{\infty} E(\mathbf{u}_t\mathbf{u}_{t-j}') = \frac{1}{T}S.$$

The last equality denotes S, known for other reasons as the *spectral density matrix at frequency zero* of $\mathbf{u}_t$. (Precisely, S so defined is the variance-covariance matrix of the g_T for fixed $\mathbf{b}$. The actual variance-covariance matrix of g_T must take into account the fact that we chose $\mathbf{b}$ to minimize g_T. I give that formula below. The point here is heuristic.)

This fact suggests that a good weighting matrix might be the inverse of S. In fact, Hansen (1982) shows formally that the choice

$$W^* = S^{-1},\ \ S \equiv \sum_{j=-\infty}^{\infty} E(\mathbf{u}_t\mathbf{u}_{t-j}')$$

is the statistically *optimal weighing matrix*, in the sense that it produces estimates with lowest asymptotic variance.

You may be more used to the formula $\sigma(u)/\sqrt{T}$ for the standard deviation of a sample mean. This formula is a special case that holds when the $u_t's$ are i.i.d. In that case $E_t(\mathbf{u}_t\mathbf{u}_{t-j}') = 0,\ j \neq 0$, so the previous equation reduces to

$$var\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_{t+1}\right) = \frac{1}{T}E(\mathbf{u}\mathbf{u}') = \frac{var(\mathbf{u})}{T}.$$

This is probably the first statistical formula you ever saw – the variance of the sample mean. In GMM, it is the last statistical formula you'll ever see as well. GMM amounts to just generalizing the simple ideas behind the distribution of the sample mean to parameter estimation and general statistical contexts.

As you can see, the variance formulas used in GMM *do not* include the usual assumptions that variables are i.i.d., homoskedastic, etc. You can put such assumptions in if you want to – we'll see how below, and adding such assumptions simplifies the formulas and can improve the small-sample performance when the assumptions are justified – but you don't *have* to add these assumptions. That's why the formulas look a little different.

Testing

Once you've estimated the parameters that make a model "fit best", the natural question is, how well does it fit? It's natural to look at the pricing errors and see if they are "big". A natural measure of "big" is, are the pricing errors statistically big? That's exactly the question answered by the J_T test. Recall,

$$TJ_T = T\left[\mathbf{g}_T(\hat{\mathbf{b}})'S^{-1}\mathbf{g}_T(\hat{\mathbf{b}})\right] \sim \chi^2(\#\text{moments} - \#\text{parameters}).$$

J_T looks like the minimized pricing errors divided by their variance-covariance matrix. The distribution theory just says that sample means converge to a normal, so sample means squared divided by variance converges to the square of a normal, or χ^2. Thus, the J_T test tells you whether the pricing errors are "big" relative to their sampling variation under the null that the model is true. (If $\mathbf{b}$ were fixed, S would in fact be the asymptotic variance-covariance matrix of the $\mathbf{g}_T$, and the result would be χ^2 with $\#$ moments degrees of freedom. The reduction in degrees of freedom corrects for the fact that we chose the parameters to make $\mathbf{g}_T$ small. More details below.)

The first and second stage estimates should remind you of procedures with standard linear regression models: if the errors are not i.i.d., then you run an OLS regression, which is consistent, but not efficient. You can then use the OLS estimate to obtain a series of residuals, estimate a variance-covariance matrix of residuals, and then do GLS, which is also consistent and more efficient, meaning that the sampling variation in the estimated parameters is lower.

9.3 Estimating the spectral density matrix

Hints on estimating the spectral density or long run covariance matrix. 1) Remove means 2) How many covariance terms to include 3) Bartlett/Newey West and other covariance weighting schemes 4) If you use S as a weighting matrix, don't let the number of moments get large relative to sample size, or impose parametric restrictions 5) Iteration and simultaneous b, S estimation.

The optimal weighting matrix S depends on *population* moments, and depends on the parameters $\mathbf{b}$. Work back through the definitions,

$$S = \sum_{j=-\infty}^{\infty} E(\mathbf{u}_t\mathbf{u}'_{t-j}) = \sum_{j=-\infty}^{\infty} E\left[\left(m_{t+1}(\mathbf{b})\mathbf{x}_{t+1} - \mathbf{p}_t\right)\mathbf{u}'_{t-j}\right].$$

How do we construct this matrix? Following the usual philosophy, we estimate population moments by their sample counterparts. Thus, use the first stage $\mathbf{b}$ estimates and the data to construct sample versions of the definition of S. This produces a consistent estimate of the true spectral density matrix, which is all the asymptotic distribution theory requires.

In asymptotic theory, you can use consistent first stage $\mathbf{b}$ estimates formed by *any* nontrivial weighting matrix. In practice, of course, you should use a sensible weighting matrix like $W = I$ so that the first stage estimates are not ridiculously inefficient. There are several additional considerations to be aware of in estimating spectral density matrices

1) Removing means. Under the null, $E(\mathbf{u}_t) = 0,$ so it shouldn't matter whether one estimates the covariance matrix by removing means, using

$$\frac{1}{T}\sum_{t=1}^{T}\left[(\mathbf{u}_t - \bar{\mathbf{u}})(\mathbf{u}_t - \bar{\mathbf{u}})'\right]; \quad \bar{\mathbf{u}} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t$$

or whether one estimates the second moment matrix by not removing means. However, Hansen and Singleton (1982) advocate removing the means in sample, and this is generally a good idea. Under *alternatives* in which $E(\mathbf{u})$ is zero, removing means should give more reliable performance.

In addition, the major obstacle to second-stage estimation is that estimated S matrices (and even simple variance-covariance matrices) are nearly singular. Second moment matrices $E(\mathbf{u}\mathbf{u}') = cov(\mathbf{u}, \mathbf{u}') + E(\mathbf{u})E(\mathbf{u}')$ are even worse.

2) Correlations under the null or alternative? Under some null hypotheses, $E_t(\mathbf{u}_{t+1}) = 0,$ so $E(\mathbf{u}_t\mathbf{u}_{t-j}) = 0$ for $j \neq 0.$ For example, this is true in the canonical case, $0 = E_t(m_{t+1}R_{t+1} - 1) = E_t(u_{t+1}).$ The discounted return should be unforecastable, using past discounted returns as well as any other variable. Thus, one could exploit the null to only include *one* term, and estimate

$$\hat{S} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{u}_t'.$$

Again, however, the null might not be correct, and the errors might be correlated. If so, you might make a big mistake by leaving them out. If the null is correct, the extra terms will converge to zero and you will only have lost a few degrees of freedom needlessly estimating them. With this in mind, one might want to include at least a few extra autocorrelations, even when the null says they don't belong.

Monte Carlo evidence (Hodrick 199x, Campbell 1994) suggests that imposing the null hypothesis to simplify the spectral density matrix helps to get the *size* of test statistics right – the probability of rejection given the null is true. Using more general spectral density matrices that can accommodate alternatives can help with the *power* of test statistics – the probability of rejection given that the alternative is true.

This trade-off requires some thought. For *measurement* rather than pure *testing*, using a spectral density matrix that can accommodate alternatives may make for more robust test statistics. For example, if you are running regressions to see if a variable such as dividend-price ratio forecasts returns, and calculating an S matrix to develop standard errors for the

OLS regression coefficients, it may make sense to use more lags than required. While the null hypothesis that nothing forecasts returns is interesting and implies the number of lags, the spirit really is more measurement than testing.

If you are testing an asset pricing model that predicts u should not be autocorrelated, and there is a lot of correlation – if this issue makes a big difference – then this is an indication that something is wrong with the model; that including u as one of your instruments z would result in a rejection.

3) Downweight higher order correlations. Why not include all available autocorrelations? The problem with this approach is that the last autocorrelation $E(u_t u_{t-T+1})$ is estimated from *one* data point. Hence it will be a pretty unreliable estimate. For this reason, the estimator using all possible autocorrelations is *inconsistent*. (Consistency means that as the sample grows, the probability distribution of the estimator converges to the true value.) To get a consistent estimate, you have to promise to let the number of included correlations increase more slowly than sample size. Even in a finite sample, higher autocorrelations are more and more badly measured, so you want to think about leaving them out.

Furthermore, even S estimates that use few autocorrelations are not always positive definite in sample. This is embarrassing when one tries to invert the estimated spectral density matrix, as called for in the formulas. Therefore, it is a good idea to construct consistent estimates that are automatically positive definite in every sample. One such estimate is the Bartlett estimate, used in this application by Newey and West (1987). It is

$$\hat{S} = \sum_{j=-k}^{k} \frac{k-|j|}{k} \frac{1}{T} \sum_{t=1}^{T} (\mathbf{u}_t \mathbf{u}'_{t-k}). \tag{96}$$

As you can see, only autocorrelations up to kth $(k < T)$ order are included, and higher order autocorrelations are downweighted. A variety of other weighting schemes have been advocated with the same effect. See Andrews (19xx).

The Newey-West estimator is basically the variance of kth sums,

$$\begin{aligned} Var\left(\sum_{j=1}^{k} \mathbf{u}_{t-j}\right) &= kE(\mathbf{u}_t\mathbf{u}'_t) + (k-1)[E(\mathbf{u}_t\mathbf{u}'_{t-1}) + E(\mathbf{u}_{t-1}\mathbf{u}'_t)] + \cdots \\ +[E(\mathbf{u}_t\mathbf{u}'_{t-k}) + E(\mathbf{u}_{t-k}\mathbf{u}'_t)] &= k\sum_{j=-k}^{k} \frac{k-|j|}{k} E(\mathbf{u}_t\mathbf{u}'_{t-k}). \end{aligned}$$

This logic also gives some intuition for the S matrix. Recall that we're looking for the variance *across samples* of the sample mean $var(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t)$. We only have one sample mean to look at, so we estimate the variance of the sample mean by looking at the variance *in a single sample* of shorter sums, $var\left(\frac{1}{k}\sum_{j=1}^{k}\mathbf{u}_j\right)$. The S matrix is sometimes called the *long-run covariance* matrix for this reason.

In fact, one could estimate S directly as a variance of kth sums and obtain almost the same estimator, that would also be positive definite in any sample,

$$\begin{aligned} \mathbf{v}_t &= \sum_{j=1}^{k} \mathbf{u}_{t-j}; \bar{\mathbf{v}} = \frac{1}{T-k} \sum_{t=k+1}^{T} \mathbf{v}_t \\ \hat{S} &= \frac{1}{k} \frac{1}{T-k} \sum_{t=k+1}^{T} \left(\mathbf{v}_t - \bar{\mathbf{v}}\right) \left(\mathbf{v}_t - \bar{\mathbf{v}}\right)'. \end{aligned}$$

This estimator has been used when measurement of S is directly interesting (Cochrane 1998, Lo and MacKinlay 1988).

What value of $k,$ or how wide a window if of another shape, should you use? Here again, you have to use some judgment. Too short values of k, and you don't correct for correlation that might be there in the errors. Too long a value of k, and the performance of the estimate and test deteriorates. If $k = T/2$ for example, you are really using only two data points to estimate the variance of the mean. The optimum value then depends on how much persistence or low-frequency movement there is in a particular application, vs. accuracy of the estimate.

There is an extensive statistical literature about optimal window width, or size of k. Alas, this literature mostly characterizes the *rate* at which k should increase with sample size. You must promise to increase k as sample size increases, but not as quickly, $\lim_{T\rightarrow\infty} k = \infty,\ \lim_{T\rightarrow\infty} k/T = 0,$ in order to obtain consistent estimates. In practice, promises about what you'd do with more data are pretty meaningless. (And usually broken once more data arrives.)

4) Consider parametric structures for autocorrelation, cross-correlation, and heteroskedasticity.

Monte Carlo evidence seems to suggest that if there is a lot of autocorrelation (or heteroskedasticity) in the data, "nonparametric" corrections such as (96) don't perform very well. The asymptotic distribution theory that ignores sampling variation in covariance matrix estimates is a poor approximation to the finite-sample distribution, so one should use a Monte-Carlo or other method to get at the finite-sample distribution of such a test statistic.

One alternative is to impose a parametric structure on the correlation pattern. For example, if you model a scalar u as an AR(1) with parameter $\rho,$ then you can estimate two numbers ρ and σ_u^2 rather than a whole list of autocorrelations, and calculate

$$S = \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}) = \sigma_u^2 \sum_{j=-\infty}^{\infty} \rho^{|j|} = \sigma_u^2 \frac{1+\rho}{1-\rho}$$

If this structure is correct, imposing it can result in much more efficient test statistics since one has to estimate many fewer coefficients. Similar parametric structures could be used to model the cross-sectional correlation of large number of moments, or the heteroskedasticity structure. Of course, there is the danger that the parametric structure is wrong.

Alternatively one could transform the data in such a way that there is less correlation to correct for in the first place.

5) Size problems.

If you try to estimate a covariance matrix that is larger than the number of data points (say 2000 NYSE stocks and 800 monthly observations), the estimate of S, like any other covariance matrix, is singular by construction. This fact leads to obvious numerical problems when you try to invert S! More generally, when the number of moments is much more than around 1/10 the number of data points, S estimates tend to become unstable and near-singular. Used as a weighting matrix, such an S matrix tells you to pay lots of attention to strange and probably spurious linear combinations of the moments. For this reason, most second-stage GMM estimations are limited to a few assets and a few instruments.

A good, but as yet untried alternative might be to impose a factor structure or other well-behaved structure on the covariance matrix. The universal practice of grouping assets into portfolios before analysis implies an assumption that the true S has a factor structure. It might be better to estimate an S imposing a factor structure on all the primitive assets.

Another response to the difficulty of estimating S is to stop at first stage estimates, and only use S for standard errors. One might also use a highly structured estimate of S as weighting matrix, while using a less constrained estimate for the standard errors.

This problem is of course not unique to GMM. Any estimation technique requires us to calculate a covariance matrix. Many traditional estimates simply assume that errors are cross-sectionally independent. This leads to understatements of the standard errors far worse than the small sample performance of any GMM estimate.

6) Alternatives to the two-stage procedure.

Hansen and Singleton (1982) describe the above two-step procedure, and it has become popular for that reason. Two alternative procedures may perform better in practice, i.e. may result in asymptotically equivalent estimates with better small-sample properties.

a) *Iterate*. The second stage estimate $\hat{\mathbf{b}}_2$ will not imply the same spectral density as the first stage. It might seem appropriate that the estimate of $\mathbf{b}$ and of the spectral density should be consistent, i.e. to find a fixed point of $\hat{\mathbf{b}} = \min_{\{\mathbf{b}\}}[g_T(\mathbf{b})'S(\hat{\mathbf{b}})^{-1}g_T(\mathbf{b})]$. One way to search for such a fixed point is to iterate: find $\mathbf{b}_2$ from

$$\hat{\mathbf{b}}_2 = \min_{\{\mathbf{b}\}} g_T(\mathbf{b})'S^{-1}(\mathbf{b}_1)g_T(\mathbf{b}) \tag{97}$$

where $\mathbf{b}_1$ is a first stage estimate, held fixed in the minimization over $\mathbf{b}_2$. Then use $\hat{\mathbf{b}}_2$ to find $S(\hat{\mathbf{b}}_2)$, find

$$\hat{\mathbf{b}}_3 = \min_{\{\mathbf{b}\}}[g_T(\mathbf{b})'S(\hat{\mathbf{b}}_2)^{-1}g_T(\mathbf{b})],$$

and so on. There is no fixed point theorem that such iterations will converge, but they often do, especially with a little massaging. (I once used $S\left[(\mathbf{b}_j + \mathbf{b}_{j-1})/2\right]$ in the beginning part of an iteration to keep it from oscillating between two values of $\mathbf{b}$). Ferson and Foerster (199x) find that iteration gives better small sample performance than two-stage GMM in Monte Carlo experiments.

b) *Pick* $\mathbf{b}$ *and* S *simultaneously.* It is *not* true that S must be held fixed as one searches for $\mathbf{b}$. Instead, one can use a new $S(\mathbf{b})$ for each value of $\mathbf{b}$. Explicitly, one can estimate $\mathbf{b}$ by

$$\min_{\{\mathbf{b}\}} \left[g_T(\mathbf{b})' S^{-1}(\mathbf{b}) g_T(\mathbf{b}) \right] \tag{98}$$

The estimates produced by this simultaneous search will not be numerically the same in a finite sample as the two-step or iterated estimates. The first order conditions to (97) are

$$\left(\frac{\partial g_T(\mathbf{b})}{\partial \mathbf{b}} \right)' S^{-1}(\mathbf{b}_1) g_T(\mathbf{b}) = 0 \tag{99}$$

while the first order conditions in (98) add a term involving the derivatives of $S(\mathbf{b})$ with respect to $\mathbf{b}$. However, the latter terms vanish asymptotically, so the asymptotic distribution theory is not affected. Hansen, Heaton and Luttmer (19xx) conduct some Monte Carlo experiments and find that this estimate may have small-sample advantages. On the other hand, one might worry that the one-step minimization will find regions of the parameter space that blow up the spectral density matrix $S(\mathbf{b})$ rather than lower the pricing errors g_T.

Often, one choice will be much more convenient than another. For linear models, one can find the minimizing value of $\mathbf{b}$ from the first order conditions (99) analytically. This fact eliminates the need to search so even an iterated estimate is much faster. For nonlinear models, each step involves a numerical search over $g_T(\mathbf{b})' S g_T(\mathbf{b})$. Rather than perform this search many times, it may be much quicker to minimize once over $g_T(\mathbf{b})' S(\mathbf{b}) g_T(\mathbf{b})$. On the other hand, the latter is not a locally quadratic form, so the search may run into greater numerical difficulties.

Chapter 10. General formulas, other uses of GMM

GMM procedures can be used to implement a host of estimation and testing exercises. Just about anything you might want to estimate can be written as a special case of GMM. To do so, you just have to remember (or look up) a few very general formulas, and then map them into your case. I start with the general formulas and then give a few examples of interesting but hard-looking questions that can be mapped into the formulas.

10.1 General GMM formulas

The general GMM estimate $a_T g_T(\hat{b}) = 0$

Distribution of $\hat{b}$: $Tcov(\hat{b}) = (ad)^{-1}aSa'(ad)^{-1\prime}$

Distribution of $g_T(\hat{b})$: $Tcov\left[g_T(\hat{b})\right] = \left(I - d(ad)^{-1}a\right) S \left(I - d(ad)^{-1}a\right)'$

The "optimal" estimate uses $a = d'S^{-1}$

With $a = d'S^{-1}, Tcov(\hat{b}) = (d'S^{-1}d)^{-1}, Tcov\left[g_T(\hat{b})\right] = S - d(d'S^{-1}d)^{-1}d'$, and a χ^2 test that $g_T(b) = 0$ simplifies to the famous J_T test, $Tg_T(\hat{b})'S^{-1}g_T(\hat{b}) \rightarrow \chi^2(\#\text{moments}-\#\text{parameters})$.

Express a model as

$$E[f(x_t, b)] = 0$$

Everything is a vector: f can represent a vector of L sample moments, x_t can be M data series, b can be N parameters.

Definition of the GMM estimate. We estimate parameters $\hat{b}$ to set some linear combination of sample means of f to zero,

$$\hat{b}: \text{ set } a_T g_T(\hat{b}) = 0$$

where

$$g_T(b) \equiv \frac{1}{T}\sum_{t=1}^{T} f(x_t, b).$$

This defines the GMM estimate.

Any statistical procedure divides into "how to produce the number" and "what is the distribution theory of that number". The point then is then distribution theory for the estimate $\hat{b}$ and for the minimized moment conditions $g_T(\hat{b})$.

Standard errors of the estimate. Hansen (1982), Theorem 3.1 tells us that the asymptotic distribution of the GMM estimate is

$$\sqrt{T}(\hat{b}-b) \rightarrow \mathcal{N}\left[0,\ (ad)^{-1}aSa'(ad)^{-1\prime}\right] \tag{100}$$

where

$$d \equiv E\left[\frac{\partial f}{\partial b'}(x_t,\ b)\right] = \frac{\partial g_T(b)}{\partial b'}$$

(i.e., d is defined as the population moment in the first equality, which we estimate in sample by the second equality),

$$a \equiv \text{plim } a_T$$

$$S \equiv \sum_{j=-\infty}^{\infty} E\left[f(x_t,b),\ f(x_{t-j}b)'\right].$$

Don't forget the $\sqrt{T}$! In practical terms, this means to use

$$var(\hat{b}) = \frac{1}{T}(ad)^{-1}aSa'(ad)^{-1\prime} \tag{101}$$

as the covariance matrix for standard errors and tests.

The "optimal" choice of weighting matrix is

$$a = d'S^{-1}, \tag{102}$$

This choice, $d'S^{-1}g_T(\hat{b}) = 0$ is the first order condition to $\min_{\{b\}} g_T(b)'S^{-1}g_T(b)$. With this weighting matrix, the standard error formula reduces to

$$\sqrt{T}(\hat{b}-b) \rightarrow \mathcal{N}\left[0,\ (d'S^{-1}d)^{-1}\right]. \tag{103}$$

This is Hansen's Theorem 3.2.

Distribution of the moments. Hansen's Lemma 4.1 gives the sampling distribution of the $g_T(b)$:

$$\sqrt{T}g_T(\hat{b}) \rightarrow \mathcal{N}\left[0, \left(I - d(ad)^{-1}a\right) S \left(I - d(ad)^{-1}a\right)'\right]. \tag{104}$$

As we have seen, S would be the asymptotic variance-covariance matrix of sample means, if we did not estimate any parameters, which sets some linear combinations of the g_T to zero. The $I - d(ad)^{-1}a$ terms account for the fact that in each sample some linear combinations of g_T are set to zero. Thus, this variance-covariance matrix is singular. With the optimal weighting matrix (102), we get the simplified formula

$$cov(\hat{b}) = S - d(d'S^{-1}d)^{-1}d'$$

J_T and χ^2tests. A sum of squared standard normals is distributed χ^2. Therefore, it is natural to use the distribution theory for g_T to see if all the g_T are "too big". Equation 104 suggests that we form the statistic

$$Tg_T(\hat{b})' \left[\left(I - d(ad)^{-1}a \right) S \left(I - d(ad)^{-1}a \right)' \right]^{-1} g_T(\hat{b}) \tag{105}$$

and that it should have a χ^2 distribution. It does, but with a hitch: The variance-covariance matrix is singular, so you have to pseudo-invert it. For example, you can perform an eigenvalue decomposition $\sum = Q\Lambda Q'$ and then invert only the non-zero eigenvalues. Also, the χ^2 distribution has degrees of freedom given by the number non-zero linear combinations of g_T, the number of moments less number of estimated parameters.

If we use the optimal set of moments (102), then Hansen's Lemma 4.2 tells us that

$$Tg_T(\hat{b})'S^{-1}g_T(\hat{b}) \rightarrow \chi^2(\#\text{moments} - \#\text{parameters}). \tag{106}$$

While one can obtain an equivalent statistic by plugging the optimal matrix (102) into the formula (104) or (105), this result is nice since we get to use the already-calculated and non-singular S^{-1}.

To derive (106) from (104), Hansen factors $S = CC'$ and then finds the asymptotic covariance matrix of $C^{-1}g_T(\hat{b})$ using (104). The result is

$$var\left[\sqrt{T}C^{-1}g_T(\hat{b}) \right] = I - C^{-1}d(d'S^{-1}d)^{-1}d'C^{-1\prime}.$$

This is an idempotent matrix of rank #moments-#parameters, so (106) follows.

10.2 Standard errors of anything by delta method

Often, we can write the estimate we want as a function of sample means,

$$b = \phi\left[E(x_t)\right] = \phi(\mu).$$

In this case, the formula (100) reduces to

$$var(b_T) = \frac{1}{T}\left[\frac{d\phi}{d\mu}\right]' \sum_{j=-\infty}^{\infty} cov(x_t, x'_{t-j}) \left[\frac{d\phi}{d\mu}\right]. \tag{107}$$

The formula is very intuitive. The variance of the sample mean is the covariance term inside. The derivatives just linearize the function ϕ near the true b.

10.3 Using GMM for regressions

By mapping OLS regressions in to the GMM framework, we derive formulas for OLS standard errors that correct for autocorrelation and conditional heteroskedasticity of the errors. The general formula is

$$var(\hat{\boldsymbol{\beta}}) = \frac{1}{T} E(\mathbf{x}_t \mathbf{x}_t')^{-1} \left[\sum_{j=-\infty}^{\infty} E(u_t \mathbf{x}_t \mathbf{x}_{t-j}' u_{t-j}) \right] E(\mathbf{x}_t \mathbf{x}_t')^{-1}.$$

and it simplifies in special cases.

Mapping any statistical procedure into GMM makes it easy to develop an asymptotic distribution that corrects for, or is insensitive to, statistical problems such as non-normality, serial correlation and conditional heteroskedasticity. To illustrate, as well as to develop the very useful formulas, I map OLS regressions into GMM.

Correcting OLS standard errors for econometric problems is *not* the same thing as GLS. When errors do not obey the OLS assumptions, OLS is consistent, and often more robust than GLS, but its standard errors need to be corrected.

OLS picks parameters β to minimize the variance of the residual:

$$\min_{\{\beta\}} E_T \left[(y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2 \right].$$

We find $\hat{\beta}$ from the first order condition, which states that the residual is orthogonal to the right hand variable:

$$g_T(\hat{\boldsymbol{\beta}}) = E_T \left[\mathbf{x}_t (y_t - \mathbf{x}_t' \hat{\boldsymbol{\beta}}) \right] = 0 \qquad (108)$$

This condition is exactly identified–the number of moments equals the number of parameters. Thus, we set the sample moments exactly to zero and there is no weighting matrix ($a = I$). We can solve for the estimate analytically,

$$\hat{\boldsymbol{\beta}} = \left[E_T(\mathbf{x}_t \mathbf{x}_t') \right]^{-1} E_T(\mathbf{x}_t \mathbf{y}_t).$$

This is the familiar OLS formula. The rest of the ingredients to equation (100) are

$$d = E(\mathbf{x}_t \mathbf{x}_t')$$

$$f(\mathbf{x}_t, \boldsymbol{\beta}) = \mathbf{x}_t (y_t - \mathbf{x}_t' \boldsymbol{\beta}) = \mathbf{x}_t u_t$$

where u_t is the regression residual. Then, equation (100) gives

$$var(\hat{\boldsymbol{\beta}}) = \frac{1}{T}E(\mathbf{x}_t\mathbf{x}_t')^{-1}\left[\sum_{j=-\infty}^{\infty} E(u_t\mathbf{x}_t\mathbf{x}_{t-j}'u_{t-j})\right]E(\mathbf{x}_t\mathbf{x}_t')^{-1}. \tag{109}$$

This is our general formula for OLS standard errors. Let's look at some special cases:

Serially uncorrelated, homoskedastic errors:

These are the usual OLS assumptions. It's good to see the usual standard errors emerge. Formally, the assumptions are

$$E(u_t \mid \mathbf{x}_t, \mathbf{x}_{t-1} \ldots u_{t-1},\ u_{t-2} \ldots) = 0 \tag{110}$$

$$E(u_t^2 \mid \mathbf{x}_t, \mathbf{x}_{t-1} \ldots u_t,\ u_{t-1} \ldots) = \text{constant} = \ \sigma_u^2. \tag{111}$$

The first assumption means that only the $j = 0$ term enters the sum

$$\sum_{j=-\infty}^{\infty} E(u_t\mathbf{x}_t\mathbf{x}_{t-j}'u_{t-j}) = \ E(u_t^2\mathbf{x}_t\mathbf{x}_t').$$

The second assumption means that

$$E(u_t^2\mathbf{x}_t\mathbf{x}_t') = E(u_t^2)E(\mathbf{x}_t\mathbf{x}_t') = \sigma_u^2E(\mathbf{x}_t\mathbf{x}_t').$$

Hence equation (109) reduces to our old friend,

$$var(\hat{\boldsymbol{\beta}}) = \frac{1}{T}\sigma_u^2E(\mathbf{x}_t\mathbf{x}_t')^{-1} = \sigma_u^2\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1} = \sigma_u^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

The last notation is typical of econometrics texts, in which $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_T \end{bmatrix}'$ represents the data matrix.

2) Heteroskedastic errors.

If we delete the conditional homoskedasticity assumption (111), we can't pull the u out of the expectation, so the standard errors are

$$var(\hat{\boldsymbol{\beta}}) = \frac{1}{T}E(\mathbf{x}_t\mathbf{x}_t')^{-1}E(u_t^2\mathbf{x}_t\mathbf{x}_t')E(\mathbf{x}_t\mathbf{x}_t')^{-1}.$$

These are known as "Heteroskedasticity corrected standard errors" or "White standard errors" after White (19xx).

3) Hansen-Hodrick errors

Hansen and Hodrick (1982) consider regressions in which the forecasting interval is longer than the sampling interval, i.e.

$$y_{t+k} = \beta' \mathbf{x}_t + \varepsilon_{t+k} \quad t = 1, 2, ...T.$$

Fama and French (1988) also use regressions of overlapping long horizon returns on variables such as dividend/price ratio and term premium. Such regressions are an important part of the evidence for predictability in asset returns.

Under the null that returns are unforecastable, we will still see correlation in the ε_t due to overlapping data. Formally, unforecastable returns y implies

$$E(\varepsilon_t \varepsilon_{t-j}) = 0 \text{ for } |j| \geq k$$

but not for $|j| < k$. Therefore, we can only rule out terms in S lower than k. Since we might as well correct for potential heteroskedasticity while we're at it, the standard errors are

$$var(b_T) = \frac{1}{T} E(\mathbf{x}_t \mathbf{x}_t')^{-1} \left[\sum_{j=-k}^{k} E(u_t \mathbf{x}_t \mathbf{x}_{t-j}' u_{t-j}) \right] E(\mathbf{x}_t \mathbf{x}_t')^{-1}.$$

If the sum in the middle is not positive definite, you could add a weighting to the sum, possibly increasing the number of lags so that the lags near k are not unusually underweighted. Again, estimating extra lags that should be zero under the null only loses a little bit of power.

10.4 Problems

1. Use the delta method to derive the sampling variance of an autocorrelation coefficient.

Chapter 11. GMM variations

Lots of calculations beyond formal parameter estimation and overall model testing are useful in the process of evaluating a model and comparing it to others models. But one also wants to understand sampling variation in such calculations, and mapping the questions into the GMM framework allows us to do this easily. In addition, alternative estimation and evaluation procedures may be more intuitive or robust to model misspecification than the two (or multi) stage procedure described above. As we did with OLS regressions, one can map a wide variety of alternative methods into the general GMM framework to derive a distribution theory.

In this chapter I discuss four such variations on the GMM method. 1) I show how to compare two models, to see if one model drives out another in pricing a set of assets. 2) I show how to use the GMM approach to mean-variance frontier questions 3) I argue that it is often wise and interesting to use *prespecified* weighting matrices rather than the S^{-1} weighting matrix, and I show how to do this. 4) I show how to use the distribution theory for the g_T beyond just forming the J_T test in order to evaluate the importance of individual pricing errors.

All of these calculations are nothing more than creative applications of the general GMM formulas for variance covariance matrix of the estimated coefficients, equation (101) and variance covariance matrix of the moments g_T, equation (104).

11.1 Horse Races

How to test whether one set of factors drives out another. Test $\mathbf{b}_2 = 0$ in $m = \mathbf{b}_1'\mathbf{f}_1 + \mathbf{b}_2'\mathbf{f}_2$, and an equivalent chi-squared difference test.

An interesting exercise for linear factor models is to test whether one set of factors drives out another. For example, Chen Roll and Ross (1986) test whether their five "macroeconomic factors" price assets so well that one can ignore even the market return. Given the large number of ad-hoc factors that have been proposed, a statistical procedure for testing which factors survive in the presence of the others is desirable. As I showed above, when the factors are correlated, it is most interesting to test this proposition by testing whether the b in $m = b'f$ are zero rather than testing factor risk premia, since the b tell us when a factor helps to price assets and the λ tell us whether a factor is priced.

Start by estimating a general model

$$m = \mathbf{b}_1'\mathbf{f}_1 + \mathbf{b}_2'\mathbf{f}_2. \tag{112}$$

We want to know, given factors $\mathbf{f}_1$, do we need the $\mathbf{f}_2$ to price assets – i.e. is $\mathbf{b}_2 = 0$? There are two ways to do this.

First and most obviously, we have an asymptotic covariance matrix for $[\mathbf{b}_1\mathbf{b}_2]$, so we can form a t test (if b_2 is scalar) or χ^2 test for $\mathbf{b}_2 = 0$ by forming the statistic

$$\hat{\mathbf{b}}_2' var(\hat{\mathbf{b}}_2)^{-1}\hat{\mathbf{b}}_2 \sim \chi^2_{\#\mathbf{b}_2}$$

where $\#\mathbf{b}_2$ is the number of elements in the $\mathbf{b}_2$ vector. This is a Wald test..

Second, estimate a restricted system $m = \mathbf{b}_1'\mathbf{f}_1$. Since there are fewer free parameters and the same number of moments than in (112), we expect the criterion J_T to rise. If we use the same weighting matrix, (usually the one estimated from the unrestricted model (112)) then the J_T cannot in fact decline. But if $\mathbf{b}_2$ really is zero, it shouldn't rise "much". How much?

$$TJ_T(\text{restricted}) - TJ_T(\text{unrestricted}) \sim \chi^2(\#\text{of restrictions})$$

This is a "χ^2 difference" test, due to Newey and West (19xx) . It works very much like a *likelihood ratio* test.

11.1.1 Mean-variance frontier and performance evaluation

A GMM, $p = E(mx)$ approach to testing whether a return expands the mean-variance frontier. Just test whether $m = a + bR$ prices all returns. If there is no risk free rate, use two values of a.

It is common to summarize asset data by mean-variance frontiers. For example, a large literature has examined the desirability of international diversification in a mean-variance context. Stock returns from many countries are not perfectly correlated, so it looks like one can reduce portfolio variance a great deal for the same mean return by holding an internationally diversified portfolio. But is this real, or just sampling error? Even if the value-weighted portfolio were ex-ante mean-variance efficient, an ex-post mean-variance frontier constructed from historical returns on the roughly 2000 NYSE stocks would leave the value-weighted portfolio well inside the ex-post frontier. So is "I should have bought Japanese stocks in 1960" (and sold them in 1990!) a signal that broad-based international diversification a good idea now, or is it simply 20/20 hindsight regret like "I should have bought Microsoft in 1982?" Similarly, we often want to know "can a portfolio manager exploit superior information to form a portfolio that is better than one can form by a passive mean-variance construction, or is a better performance in sample just due to luck?"

DeSantis (1992) and Chen and Knez (1992,1993) showed how to examine such questions in a $p = E(mx)$, GMM framework, by applying the above horse race. We exploit the connection between mean-variance efficiency and linear discount factor models, and the GMM distribution theory. Let $\mathbf{R}^d$ be a vector of domestic asset returns and $\mathbf{R}^f$ a vector of foreign

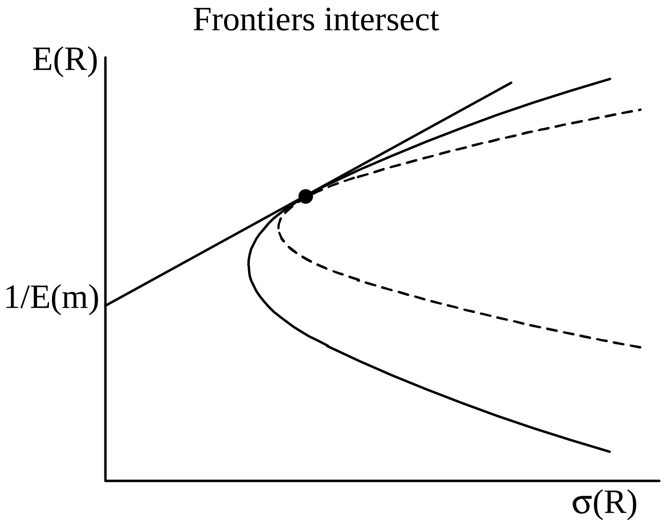


Figure 18. Mean variance frontiers might intersect rather than coincide.

asset returns. If a discount factor $m = a + \mathbf{b}_d'\mathbf{R}^d$ prices both $\mathbf{R}^d$ and $\mathbf{R}^f$, then $\mathbf{R}^d$ is on the mean variance frontier generated by $\mathbf{R}^d$ and $\mathbf{R}^f$. We *know* $m = a + \mathbf{b}_d'\mathbf{R}^d + \mathbf{b}_f'\mathbf{R}^f$ prices both sets of returns by construction, (this is x^*). Thus, we can use a Wald test on $\mathbf{b}_f = 0$, or a χ^2 difference test.

To test a portfolio manager's skill, let $m = a + bR^w + b_pR^p$ where R^p is the return on a portfolio managed by a portfolio manager. We want to know whether the portfolio manager can exploit superior knowledge, skill, or information to get outside the mean-variance frontier. Thus, we test for $b_p = 0$ in a system that includes (at least) R^w and R^p as moments.

There is a slight subtlety in this test. There are two ways in which a return R^{mv} might be on the mean-variance frontier of a larger collection of securities: the frontiers could just *intersect* at R^{mv}, as shown in Figure 18, or the frontiers could *coincide* globally.

For *intersection*, $m = a + \mathbf{b}_d'\mathbf{R}^d$ will price both $\mathbf{R}^d$ and $\mathbf{R}^f$ only for one value of a, or equivalently $E(m)$ or choice of the intercept, as shown. If the frontiers coincide, then $m = a + \mathbf{b}_d'\ \mathbf{R}^d$ prices both $\mathbf{R}^d$ and $\mathbf{R}^f$ for *any* value of a. Equivalently, the d portfolio is on the (d, f) frontier for *any* intercept, where this is true for only *one* value of the intercept in the case of intersection. Thus, to test for coincident frontiers, one must test whether $m = a + \mathbf{b}_d'$ $\mathbf{R}^d$ prices both $\mathbf{R}^d$ and $\mathbf{R}^f$ for two prespecified values of a simultaneously.

11.2 Prespecified weighting matrices

Prespecified rather than "optimal" weighting matrices can emphasize economically interesting results, they can avoid the trap of blowing up standard errors rather than improving pricing errors, they can lead to estimates that are more robust to small model misspecifications, as OLS is often preferable to GLS in a regression context, and they allow you to force GMM to use one set of moments for estimation and another for testing. The GMM formulas for this case are

$$var(\hat{\mathbf{b}}) = \frac{1}{T}(D'WD)^{-1}D'WSWD(D'WD)^{-1}$$

$$var(g_T) = \frac{1}{T}(I - D(D'WD)^{-1}D'W)S(I - WD(D'WD)^{-1}D').$$

So far, we have assumed that at some stage a matrix S will be used as the weighting matrix, so the final minimization will have the form $\min_b g_T'(b)S^{-1}g_T(b)$. As we have seen, this objective maximizes the *statistical* information in the sample about a model. However, there are several reasons why one may want to use a prespecified weighting matrix instead, or as a diagnostic accompanying more formal statistical tests.

Keep in mind that "using a prespecified weighting matrix" and the identity matrix in particular, is *not* the same thing as ignoring cross-correlation in the distribution theory. The S matrix will still show up in all the standard errors and test statistics.

11.2.1 How to use prespecified weighting matrices

The general theory is expressed in terms of the linear combination of the moments that is set to zero $a_T g_T(\mathbf{b}) = 0$. With weighting matrix W, the first order conditions to $\min_{\{\mathbf{b}\}} g_T'(\mathbf{b})Wg_T(\mathbf{b})$ are

$$(\partial g_T(\mathbf{b})/\partial \mathbf{b}')'\, W g_T(\mathbf{b}) = D'Wg_T(\mathbf{b}) = 0,$$

so we map into the general case with $a = D'W$. Plugging this value into (101), we obtain the variance-covariance matrix of the estimated coefficients from a prespecified W estimate,

$$var(\hat{\mathbf{b}}) = \frac{1}{T}(D'WD)^{-1}D'WSWD(D'WD)^{-1}. \qquad (113)$$

Check that with $W = S^{-1}$, this formula reduces to $1/T\ (D'S^{-1}D)^{-1}$.

Plugging $a = D'W$ into equation (104), we find the variance-covariance matrix of the moments g_T

$$var(g_T) = \frac{1}{T}(I - D(D'WD)^{-1}D'W)S(I - WD(D'WD)^{-1}D') \qquad (114)$$

One might think that the sampling error of g_T is just S; however this thought ignores the fact that degrees of freedom are lost in estimation, so a linear combination of rows of g_T to 0. The singular (rank #moments - #parameters) variance covariance matrices given above correct for this fact.

Equation (114) can be the basis of χ^2 tests for the overidentifying restrictions. If we interpret $()^{-1}$ to be a generalized inverse, then

$$g_T' var(g_T)^{-1} g_T \sim \chi^2(\#moments - \#parameters).$$

This procedure would work for the "optimal" weighting matrix S^{-1} as well; Hansen (1982) shows that $g_T' S^{-1} g_T$ yields (numerically) the same result. One way to compute a generalized inverse is to start with an eigenvalue decomposition $S = Q\Lambda Q'$; then the generalized inverse is $S^{-1} = Q\Lambda^+ Q'$, where Λ^+ uses the inverse of the nonzero eigenvalues but leaves the zero eigenvalues alone.

11.2.2 Motivations for prespecified weighting matrices

Level playing field. The S matrix changes as the model and as its parameters change. (See the definition). As a result, comparing models by their J_T values is dangerous, since models may "improve" because they simply blow up the estimates of S, rather than make any progress on lowering the pricing errors g_T. By using a weighting matrix that does not vary from model to model, or across parameter values for a given model, one imposes a level playing field and avoids this problem when the J_T test is used as a model comparison statistic.

(No one would formally use a comparison of J_T tests across models to compare them. The minute you think carefully about it, you realize that you must use the same weighting matrix as well as the same moments, and the χ^2 difference test does both. But it has proved nearly irresistible for authors to claim success for a new model over previous ones by noting improved J_T statistics in introductions and conclusions, despite different weighting matrices, different moments, and sometimes much larger pricing errors..)

Robustness, as with OLS vs. GLS. When errors are autocorrelated or heteroskedastic, every econometrics textbook shows you how to "improve" on OLS by making appropriate GLS corrections. If you correctly model the error covariance matrix and if the regression is perfectly specified, this procedure can improve efficiency, i.e. give estimates with lower asymptotic standard errors. However, GLS is much less robust. If you model the error covariance matrix incorrectly, the estimates can be much worse than OLS. Also, the GLS transformations can zero in on slightly misspecified areas of the model producing garbage.

GLS is “best,” but OLS is “pretty darn good” and is usually much more robust than GLS. Furthermore, one often has enough data that wringing every last ounce of statistical precision (low standard errors) from the data is less important than producing estimates that do not depend on questionable statistical assumptions, and that transparently focus on the interesting features of the data. Thus, it is often a good practice to use OLS *estimates*, but correct the *standard errors* of the OLS estimates for these features of the error covariance matrices, for example using the formulas we developed above.

For example, the GLS transformation for highly serially correlated errors essentially turns a regression in levels into a regression in first differences. But relationships that are quite robust in levels often disappear in first differences, especially of high frequency data, because of small measurement errors. Lucas (198x) followed a generation of money demand estimates $M_t = a + bY_t + \varepsilon_t$ that had been run in quasi-first differences following GLS advice, and that had found small and unstable income elasticities, since day-to-day variation in measured money demand has little to do with day-to-day variation in income. Lucas ran the regression by OLS in levels, only correcting standard errors for serial correlation and found the pattern evident in any graph that the level of money and income track very well over years and decades. .

GMM works the same way. First-stage or otherwise fixed weighting matrix estimates may give up something in (asymptotic) efficiency if the statistical and economic models are precisely right, but may be much more robust to statistical and economic problems. You still want to use the S matrix in computing standard errors, though, as you want to correct OLS standard errors, and the following formulas show you how to do this.

Even if in the end one wants to produce “efficient” estimates and tests, it is a good idea to calculate standard errors and model fit tests for the first-stage estimates. Ideally, the parameter estimates should not change by much, and the second stage standard errors should be tighter. If the “efficient” parameter estimates do change a great deal, it is a good idea to diagnose why this is so – which moments the efficient parameter estimates are paying attention to – and then decide whether the difference in results is truly due to efficiency gain or not.

Near-singular S. The spectral density matrix is often near-singular, since asset returns are highly correlated with each other, and since we often include many assets relative to the number of data points. As a result, second stage GMM (and, as we will see below, maximum likelihood or any other efficient technique) tries to minimize differences and differences of differences of asset returns in order to extract statistically orthogonal components. One may feel that this feature leads GMM to place a lot of weight on poorly estimated, economically uninteresting, or otherwise non-robust aspects of the data. In particular, portfolios of the form $100R_1 - 99R_2$ assume that investors can in fact purchase such heavily leveraged portfolios. Short-sale costs often rule out such portfolios or significantly alter their returns, so one may not want to emphasize pricing them correctly in the estimation and evaluation.

For example, suppose that S is given by

$$S = \left[\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right].$$

so

$$S^{-1} = \frac{1}{1-\rho^2} \left[\begin{array}{cc} 1 & -\rho \\ -\rho & 1 \end{array} \right].$$

We can factor S^{-1} into a "square root" by the Choleski decomposition. This produces a triangular matrix C such that $C'C = S^{-1}$. You can check that the matrix

$$C = \left[\begin{array}{cc} \frac{1}{\sqrt{1-\rho^2}} & \frac{-\rho}{\sqrt{1-\rho^2}} \\ 0 & 1 \end{array} \right] \tag{115}$$

works. Then, the GMM criterion

$$\min g_T' S^{-1} g_T$$

is equivalent to

$$\min (g_T' C')(C g_T).$$

Cg_T gives the linear combination of moments that efficient GMM is trying to minimize. Looking at (115), as $\rho \rightarrow 1,$ the (2,2) element stays at 1, but the (1,1) and (1,2) elements get very large and of opposite signs. For example, if $\rho = 0.95,$ then

$$C = \left[\begin{array}{cc} 3.20 & -3.04 \\ 0 & 1 \end{array} \right].$$

In this example, GMM pays a little attention to the second moment, but places *three* times as much weight on the *difference* between the first and second moments. Larger matrices produce even more extreme weights.

Economically interesting moments. As explained above, and as we see in the example of the last section, the optimal weighting matrix makes GMM pay close attention to apparently *well-measured* linear combinations of moments in both estimation and evaluation. One may want to force the estimation and evaluation to pay attention to *economically* interesting moments instead. The initial portfolios are usually formed on an economically interesting characteristic such as size, beta, book/market or industry. One typically wants in the end to see how well the model prices these initial portfolios, not how well the model prices potentially strange portfolios of those portfolios. If a model fails, one may want to characterize that failure as "the model doesn't price small firms" not "the model doesn't price a portfolio of $900\times$ small firms $-600\times$ large firms $-299\times$ medium firms."

11.2.3 Some prespecified weighting matrices

Two examples of economically interesting weighting matrices are the second-moment matrix of returns, advocated by Hansen and Jagannathan (1992) and the simple identity matrix, which is used implicitly in much empirical asset pricing.

Second moment matrix. Hansen and Jagannathan (1992) advocate the use of the second moment matrix of payoffs $W = E(\mathbf{x}\mathbf{x}')^{-1}$ in place of S. They motivate this weighting matrix as an interesting distance measure between a model for m, say y, and the space of true m's. Precisely, the minimum distance (second moment) between a candidate discount factor y and the space of true discount factors is the same as the minimum value of the GMM criterion with $W = E(\mathbf{x}\mathbf{x}')^{-1}$ as weighting matrix.

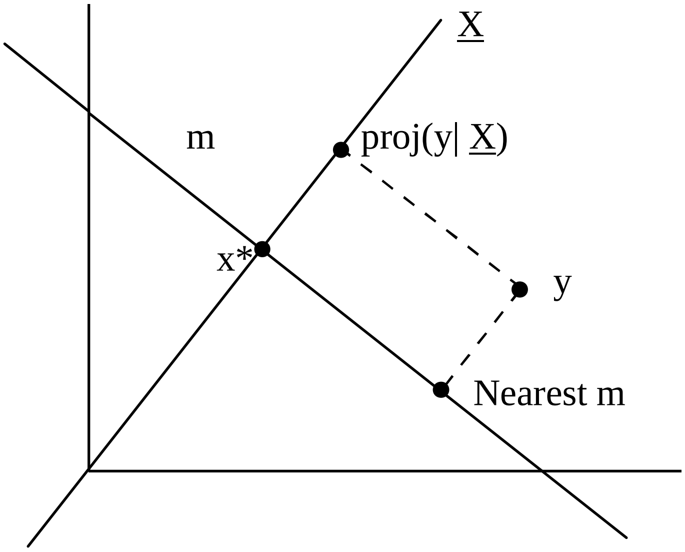


Figure 19. Distance between y and nearest m = distance between $proj(y|X)$ and x^*.

To see why this is true, refer to figure 19. The distance between y and the nearest valid m is the same as the distance between $proj(y \mid \underline{X})$ and x^*. As usual, consider the case that $\underline{X}$ is generated from a vector of payoffs $\mathbf{x}$ with price $\mathbf{p}$. From the OLS formula,

$$proj(y \mid \underline{X}) = E(y\mathbf{x}')E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}$$

x^* is the portfolio of $\mathbf{x}$ that prices $\mathbf{x}$ by construction,

$$x^* = \mathbf{p}'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}$$

Then, the distance between y and the nearest valid m is:

$$\begin{aligned}\|proj(y|\underline{X}) - x^*\| &= \left\|E(y\mathbf{x}')E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x} - \mathbf{p}'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}\right\| \\ &= \left\|\left(E(y\mathbf{x}') - \mathbf{p}'\right)E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}\right\| \\ &= [E(y\mathbf{x}) - \mathbf{p}]'E(\mathbf{x}\mathbf{x}')^{-1}[E(y\mathbf{x}) - \mathbf{p}] \\ &= g_T'E(\mathbf{x}\mathbf{x}')^{-1}g_T\end{aligned}$$

You might want to choose parameters of the model to minimize this "economic" measure of model fit, or this economically motivated linear combination of pricing errors, rather than the statistical measure of fit S^{-1}. You might also use the minimized value of this criterion to compare two models. In that way, you are sure the better model is better because it improves on the pricing errors rather than just blowing up the weighting matrix.

Identity matrix. Using the identity matrix weights the initial choice of assets equally in estimation and evaluation. This choice has a particular advantage with large systems in which S is nearly singular, as it avoids most of the problems associated with inverting a near-singular S matrix. It also ensures that the GMM estimation pays equal attention to the initial choice of portfolios, which were usually selected with some care. Many empirical asset pricing studies use OLS cross-sectional regressions, which are the same thing as a first stage GMM estimate with an identity weighting matrix.

Comparing the second moment and identity matrices.

The second moment matrix gives an objective that is invariant to the initial choice of assets. If we form a portfolio $\mathbf{Ax}$ of the initial payoffs $\mathbf{x}$, with nonsingular $\mathbf{A}$ (don't throw away information) then

$$[E(y\mathbf{Ax}) - \mathbf{Ap}]'E(\mathbf{Axx}'\mathbf{A}')^{-1}[E(y\mathbf{Ax}) - \mathbf{Ap}] = [E(y\mathbf{x}) - \mathbf{p}]'E(\mathbf{xx}')^{-1}[E(y\mathbf{x}) - \mathbf{p}].$$

The optimal weighting matrix S shares this property. It is not true of the identity matrix: the results will depend on the initial choice of portfolios.

Kandel and Stambaugh (19xx) have suggested that the results of several important asset pricing model tests are highly sensitive to the choice of portfolio; i.e. that authors inadvertently selected a set of portfolios on which the CAPM does unusually badly in a particular sample. Insisting that weighting matrices have this invariance to portfolio selection might be a good discipline against this kind of fishing.

On the other hand, if you want to focus on the model's predictions for economically interesting portfolios, then it wouldn't make much sense for the weighting matrix to undo the specification of economically interesting portfolios! For example, many studies want to focus on the ability of a model to describe expected returns that seem to depend on a characteristic such as size, book/market, industry, momentum, etc.

The second moment matrix is often even more nearly singular than the spectral density matrix, since $E(\mathbf{xx}') = cov(\mathbf{x}) + E(\mathbf{x})E(\mathbf{x})'$. Therefore, it often emphasizes portfolios with even more extreme short and long positions, and is no help on overcoming the near singularity of the S matrix. If the number of moments (test assets times instruments) is much above 1/20 -

11.2.4 Estimating on one group of moments, testing on another.

You may also want to force the system to use one set of moments for *estimation* and another for *testing*. The real business cycle literature in macroeconomics does this extensively, typically using "first moments" for estimation ("calibration") and "second moments" (i.e. first moments of squares) for evaluation. A statistically minded macroeconomist might like to know whether the departures of model from data "second moments" are large compared to sampling variation, and would like to include sampling uncertainty about the parameter estimates in this evaluation. You might similarly want to choose parameters using one set of asset returns (stocks; domestic assets; size portfolios, first 9 size deciles) and then see how the model does "out of sample" on another set of assets (bonds; foreign assets; book/market portfolios, small firm portfolio). However, you want the distribution theory for evaluation on the second set of moments to incorporate sampling uncertainty about the parameters in their estimation on the first set of moments.

You can do all this very simply by using an appropriate prespecified weighting matrix. Construct a weighting matrix W which is zero in the columns and rows corresponding to the test moments. Then, those moments will not be used in estimation. (You could start with S for some efficiency, but the identity weighting matrix is more consistent with the philosophy of the exercise.) Consider this a fixed-weighting matrix estimate, and then use formula (114) to construct a χ^2 test of the moments you want to test.

11.3 Testing moments

How to test one or a group of pricing errors, such as small firm returns. 1) Use the formula for $var(g_T)$ 2) A chi-squared difference test. How to estimate with one group of moments and test on another.

You may want to see how well a model does on particular moments or particular pricing errors. For example, the celebrated "small firm effect" states that an unconditional CAPM $(m = a+bR^W$, no scaled factors) does badly in pricing the returns on a portfolio that always holds the smallest 1/10th or 1/20th of firms in the NYSE. You might want to see whether a new model prices the small returns well.

It is a nice diagnostic for any asset pricing model to plot predicted excess returns vs. actual excess returns in the data. Such plots generalize traditional plots of average return vs.

estimated beta for the CAPM and allow a visual sense of how well the model explains the cross sectional variation of average returns. Of course, one would like standard errors for such plots, a method of computing vertical error bars or the uncertainty about the difference between predicted moment and actual moment.

We have already seen that individual elements of g_T measure the pricing errors or expected return errors. Thus, all we need is the sampling error of g_T to measure the accuracy of a pricing error and to test whether an individual moment is "too far off" or not.

One possibility is to use the sampling distribution of g_T, (114) to construct a t-test (for a single g_T, such as small firms) or χ^2 test (for groups of g_T, such as small firms $\otimes$ instruments). As usual this is the *Wald* test.

Alternatively, you can construct a χ^2 difference or likelihood ratio-like test. Start with a general model that includes all the moments, and form an estimate of the spectral density matrix S. Now *set* to zero the moments you want to test, and denote $g_{sT}(\mathbf{b})$ the vector of moments, including the zeros (s for "smaller"). Consider choosing $\mathbf{b}_s$ to minimize $g_{sT}(\mathbf{b})'S^{-1}g_{sT}(\mathbf{b})$ using the same weighting matrix S. The criterion will be *lower* than the original criterion $g_T(\mathbf{b})'S^{-1}g_T(\mathbf{b})$, since there are the same number of parameters and fewer moments. But, if the moments we want to test truly are zero, the criterion shouldn't be *that much* lower. Thus, form a χ^2 difference test

$$Tg_T(\hat{\mathbf{b}})'S^{-1}g_T(\hat{\mathbf{b}}) - Tg_{sT}(\hat{\mathbf{b}}_s)S^{-1}g_{sT}(\hat{\mathbf{b}}_s) \sim \chi^2(\#\text{eliminated moments}).$$

Of course, don't fall into the obvious trap of picking the largest of 10 pricing errors and noting it's more than two standard deviations from zero. The distribution of the *largest* of 10 pricing errors is much wider than the distribution of a single one. To use this distribution, you have to pick which pricing error you're going to test *before* you look at the data.

11.4 Applying GMM to linear factor models

When $m_{t+1} = a + \mathbf{b}'\mathbf{f}_{t+1}$ and the test assets are excess returns, the GMM estimate is

$$\begin{aligned} \hat{\mathbf{b}} &= -(C'WC)^{-1}C'WE_T(\mathbf{R}^e) \\ C &\equiv E_T(\mathbf{R}^e\mathbf{f}') \end{aligned}$$

This is a GLS *cross-sectional* regression of average returns on the covariance of returns with factors. The overidentifying restrictions test is a quadratic form in the pricing errors,

$$\hat{\boldsymbol{\alpha}}'V^{-1}\tilde{\boldsymbol{\alpha}} \sim \chi^2(\#\text{assets} - \#\text{factors})$$

Linear factor models are discount factor models of the form

$$m_{t+1} = a + \mathbf{b}'\mathbf{f}_{t+1},$$

where $\mathbf{f}_{t+1}$ is a vector of time series such as portfolio returns (CAPM, APT), or "macro factors" (ICAPM, CRR), and can include factors scaled by instruments to accommodate conditioning information. The linearity simplifies the formulas, and the GMM procedure becomes similar to traditional two pass regression procedures.

Linear factor models are most often applied to excess returns. If we only use excess returns, a is not identified $(0 = E(m\mathbf{R}^e) \Leftrightarrow 0 = E(2m\mathbf{R}^e))$, so we can normalize to $a = 1$ and $E(\mathbf{f}) = \mathbf{0}$. Then, the GMM estimate of $\mathbf{b}$ is

$$\hat{\mathbf{b}} = -(C'WC)^{-1}C'WE_T(\mathbf{R}^e) \tag{116}$$

where

$$C \equiv E_T(\mathbf{R}^e\mathbf{f}')$$

is a matrix of covariances of returns with the factors.

To see this, proceed through the recipe as given above. The vector of sample moments or pricing errors is

$$g_T(\mathbf{b}) = E_T(m\mathbf{R}^e) = E_T\left(\mathbf{R}^e\mathbf{f}'\mathbf{b} + 1\right)\mathbf{b} = E_T\left(\mathbf{R}^e\right) + C\mathbf{b}$$

The GMM minimization is

$$\min g_T(\mathbf{b})'Wg_T(\mathbf{b}).$$

The first order condition is

$$\left(\frac{\partial g_T(\mathbf{b})}{\partial \mathbf{b}'}\right)' Wg_T(\mathbf{b}) = 0.$$

Note

$$D = \frac{\partial g_T(\mathbf{b})}{\partial \mathbf{b}'} = E_T(\mathbf{R}^e\mathbf{f}') = C.$$

Writing out the first order condition,

$$C'W\left[C\mathbf{b}+E_T\left(\mathbf{R}^e\right)\right] = 0.$$

and hence (116).

This GMM estimate has a natural interpretation. As we have seen many times before, $E(mR^e) = 0$ implies that expected returns should be linear in the covariance of returns with factors. Thus, the model predicts

$$E(\mathbf{R}^e) = -C\mathbf{b}. \tag{117}$$

To estimate $\mathbf{b}$ one might naively have started by tacking on an error term representing sample variation,

$$E_T(\mathbf{R}^e) = C\mathbf{b} + \mathbf{u}$$

and then estimated by OLS. The first-stage estimate is, from (116) exactly this OLS estimate,

$$\hat{\mathbf{b}}_1 = -(C'C)^{-1}C'E_T(\mathbf{R}^e).$$

This is a *cross-sectional* regression. The "data points" in the regression are sample average returns (y) and covariances of returns with factors (x) *across* test assets. We are picking the parameter $\mathbf{b}$ to make the model fit explain the cross-section of asset prices as well as possible.

The errors $\mathbf{u}$ in this cross-sectional regression are correlated across equations. Thus, we at least have to correct first-stage standard errors for this correlation, and we might think about a GLS cross-sectional regression to improve efficiency. Since S is proportional to the covariance matrix of $\mathbf{u}$, The second-stage GMM estimate

$$\hat{\mathbf{b}}_2 = -(C'S^{-1}C)^{-1}C'S^{-1}E_T(\mathbf{R}^e)$$

is exactly a *GLS* cross-sectional regression of sample mean returns on sample covariances.

The first-stage or OLS cross-sectional regression standard errors are, from (113) and $C = D$, exactly what we expect for an OLS regression with correlated errors,

$$var(\hat{\mathbf{b}}_1) = \frac{1}{T}(C'C)^{-1}C'SC(C'C)^{-1}.$$

while the second stage or GLS cross-sectional regression standard errors specialize to

$$var(\hat{\mathbf{b}}_2) = \frac{1}{T}(C'S^{-1}C)^{-1}.$$

Finally, the overidentifying restrictions test is a quadratic form in the pricing errors or Jensen's alphas,

$$g_T(\hat{\mathbf{b}}) = E_T(\mathbf{R}^e) + C\hat{\mathbf{b}} = \text{average return - predicted average return} = \hat{\boldsymbol{\alpha}}$$

From (114), this test based on the first stage estimates is

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_1' var(g_T)^{+}\hat{\boldsymbol{\alpha}}_1 &\sim \chi^2(\#\text{assets} - \#\text{factors}) \\
\hat{\boldsymbol{\alpha}}_1 &= g_T(\hat{\mathbf{b}}_1) = E_T(\mathbf{R}^e) + C\hat{\mathbf{b}}_1 \\
var(g_T) &= \frac{1}{T}(I - C(C'C)^{-1}C')S(I - C(C'C)^{-1}C'),
\end{aligned}$$

while the test based on the second stage estimates is most conveniently expressed as

$$\begin{aligned}
T\hat{\boldsymbol{\alpha}}_2' S^{-1}\hat{\boldsymbol{\alpha}}_2 &\sim \chi^2(\#\text{assets} - \#\text{factors}) \qquad (118)\\
\hat{\boldsymbol{\alpha}}_2 &= g_T(\hat{\mathbf{b}}_2) = E_T(\mathbf{R}^e) + C\hat{\mathbf{b}}_1
\end{aligned}$$

General Case

To evaluate a linear factor model using levels rather than just excess returns, and perhaps using instruments as well, we go through the same mechanics with $\mathbf{p}$ and $\mathbf{x}$, and without renormalizing to $a = 1, E(\mathbf{f}) = 0$. This is a slightly more refined way to implement any test, since it recognizes that setting the *sample* $E_T(\mathbf{f}) = 0$ in applying the above formulas introduces another estimate, whose sampling error should be accounted for in the distribution theory. Treating the constant $a \times 1$ as a constant factor, the model is

$$m_{t+1} = \mathbf{b}'\mathbf{f}_{t+1}.$$

The GMM estimate is

$$\hat{\mathbf{b}} = \left[E_T(\mathbf{f}\mathbf{x}')WE_T(\mathbf{x}\mathbf{f}')\right]^{-1} E_T(\mathbf{f}\mathbf{x}')WE_T(\mathbf{p}). \tag{119}$$

This is still a (potentially GLS) cross-sectional regression of average prices $E_T(\mathbf{p})$ on second moments$E_T(\mathbf{x}\mathbf{f}')$ of payoffs with factors. The model

$$\mathbf{p} = E(m\mathbf{x}) = E(\mathbf{x}\mathbf{f}')\mathbf{b}$$

says that prices should be proportional to second moments, so again this is a natural regression to run.

Chapter 12. Regression-based tests

Again, our basic objective in a statistical analysis is a method of producing estimates of free parameters, a distribution theory for those parameters, and model evaluation statistics such as $\hat{\boldsymbol{\alpha}}' V^{-1} \hat{\boldsymbol{\alpha}}$. In this chapter, I cover the classic regression tests of linear factor models, expressed in expected return-beta form. As you will see, they are closely related to the $p = E(mx)$,GMM tests of linear factor models we investigated in the last chapter. In the next chapter, I cover the formalization of these regression tests via maximum likelihood, and we will see they can also be formalized as an instance of GMM.

12.1 Time-series regressions

When the factor is also a return, we can evaluate the model

$$E(R^{ei}) = \beta_i E(f)$$

by running OLS *time series regressions*

$$R_t^{ei} = \alpha_i + \beta_i f_t + \varepsilon_t^i; \; t = 1, 2, ...T$$

for each asset, as suggested by Black, Jensen and Scholes. The OLS distribution formulas (with corrected standard errors) provide standard errors of α and β. With errors that are i.i.d. over time, the asymptotic joint distribution of the intercepts gives

$$T \left[1 + \left(\frac{E_T(f)}{\hat{\sigma}(f)} \right)^2 \right]^{-1} \hat{\boldsymbol{\alpha}}' \hat{\Sigma}^{-1} \hat{\boldsymbol{\alpha}} \sim \chi_N^2$$

The Gibbons-Ross-Shanken test is a multivariate, fixed-f counterpart,

$$\frac{T - N - K}{N} \left(1 + E_T(\mathbf{f})' \hat{\Omega}^{-1} E_T(\mathbf{f}) \right)^{-1} \hat{\boldsymbol{\alpha}}' \hat{\Sigma}^{-1} \hat{\boldsymbol{\alpha}} \; \sim F_{N, T-N-K}$$

I show how to construct the same test statistics with heteroskedastic and autocorrelated errors, as suggested by MacKinlay and Richardson, via GMM.

I start with the simplest case. We have a factor pricing model with a single factor which is an excess return (for example, the CAPM, with $R^{em} = R^m - R^f$), and the test assets are all excess returns. We express the model in expected return - beta form. The betas are defined by regression coefficients

$$R_t^{ei} = \alpha_i + \beta_i f_t + \varepsilon_t^i \qquad (120)$$

and the model states that expected returns are linear in the betas:

$$E(R^{ei}) = \beta_i E(f). \tag{121}$$

Since the factor is also an excess return, the model applies to the factor as well, so $E(f) = 1 \times \lambda$.

Comparing the model (121) and the expectation of the time series regression (120) we see that the model has one and only one implication for the data: *all the regression intercepts* α_i *should be zero.* The regression intercepts are equal to the pricing errors. This prediction is only true when the factors are themselves excess returns. With factors that are not priced by the model, the factor risk premium λ is not equal to the expected value of the factor, so the regression intercepts do not have to be zero. There is a restriction relating means of factors and intercepts, but it is more complicated and does not lead to such an easy regression based test.

Given this fact, Black Jensen and Scholes (19xx) suggested a natural strategy for estimation and evaluation: Run time-series regressions (120) for each test asset. If you assume that the errors are uncorrelated over time and homoskedastic, you can use standard OLS formulas for a distribution theory of the parameters, and in particular you can use t-tests to check whether the pricing errors α are in fact zero. The standard approach to OLS standard errors can also give us a test whether *all* the pricing errors are *jointly* equal to zero. Dividing the $\hat{\boldsymbol{\alpha}}$ regression coefficients by their variance-covariance matrix leads to a χ^2 test,

$$T\left[1+\left(\frac{E_T(f)}{\hat{\sigma}(f)}\right)^2\right]^{-1}\hat{\boldsymbol{\alpha}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\alpha}} \sim\chi^2_N \tag{122}$$

where $E_T(f)$ denotes sample mean, $\hat{\sigma}^2(f)$ denotes sample variance, $\hat{\boldsymbol{\alpha}}$ is a vector of the estimated intercepts,

$$\hat{\boldsymbol{\alpha}} = \left[\begin{array}{cccc} \hat{\alpha}_1 & \hat{\alpha}_2 & ... & \hat{\alpha}_N \end{array}\right]'$$

$\hat{\Sigma}$ is the residual covariance matrix, i.e. the sample estimate of $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \Sigma$, where

$$\boldsymbol{\varepsilon}_t = \left[\begin{array}{cccc} \varepsilon_t^1 & \varepsilon_t^2 & \cdots & \varepsilon_t^N \end{array}\right]'.$$

As usual when testing hypotheses about regression coefficients, this test is valid asymptotically. The asymptotic distribution theory assumes that $\sigma^2(f)$ (i.e. $X'X$) and Σ have converged to their probability limits; therefore it is asymptotically valid even though the factor is stochastic and Σ is estimated, but it ignores those sources of variation in a finite sample. It does not require that the errors are normal, relying on the central limit theorem so that $\hat{\alpha}$ is normal, but it does assume that the errors are homoskedastic (constant Σ) and not autocorrelated. I derive (122) below.

Also as usual in a regression context, we can derive a finite-sample F distribution for the hypothesis that a set of parameters are jointly zero, for fixed values of the right hand variable f_t,.

$$\frac{T-N-1}{N}\left[1+\left(\frac{E_T(f)}{\hat{\sigma}(f)}\right)^2\right]^{-1}\hat{\boldsymbol{\alpha}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\alpha}}\ \sim F_{N,T-N-1} \tag{123}$$

This is the Gibbons Ross and Shanken (198x) or GRS test statistic. The F distribution recognizes sampling variation in $\hat{\Sigma}$, which is not included in (122). This distribution requires that the errors ε are normal as well as i.i.d. and homoskedastic. With normal errors, the $\hat{\alpha}$ are normal and $\hat{\Sigma}$ is an independent Wishart (the multivariate version of a χ^2), so the ratio is F. This distribution is exact in a finite sample; however it assumes fixed values of the right hand variable f. Thus, it only answers the sampling question "what if we redraw the ε shocks, with the same time series of f?" not, "what if we redraw the entire data set?"

Tests (122) and (123) have a very intuitive form. The basic part of the test is a quadratic form in the pricing errors, $\hat{\boldsymbol{\alpha}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\alpha}}$. If there were no βf in the model, then the $\hat{\boldsymbol{\alpha}}$ would simply be the sample mean of the regression errors ε_t. Assuming i.i.d. ε_t, the variance of their sample mean is just $1/T\Sigma$. Thus, if we knew Σ then $T\hat{\boldsymbol{\alpha}}'\Sigma^{-1}\hat{\boldsymbol{\alpha}}$ would be a sum of squared sample means divided by their variance-covariance matrix, which would have an asymptotic χ^2_N distribution, or a finite sample χ^2_N distribution if the ε_t are normal. But we have to estimate Σ, which is why the finite-sample distribution is F rather than χ^2. We also estimate the β, and the second term in (122) and (123) accounts for that fact.

Recall that a single beta representation exists if and only if the reference return is on the mean-variance frontier. Thus, the test can also be interpreted as a test whether f is ex-ante mean-variance efficient, after accounting for sampling error. Even if f is on the true or ex-ante mean-variance frontier, other returns will outperform it in sample due to luck, so the return f will usually be inside the ex-post mean-variance frontier. Still, it should not be too far inside that frontier, and Gibbons Ross and Shanken show that the test statistic can be expressed in terms of how far inside the ex-post frontier the return f is,

$$\frac{T-N-1}{N}\frac{\left(\frac{\mu_q}{\sigma_q}\right)^2-\left(\frac{E_T(f)}{\hat{\sigma}(f)}\right)^2}{1+\left(\frac{E_T(f)}{\hat{\sigma}(f)}\right)^2}.$$

$\left(\frac{\mu_q}{\sigma_q}\right)^2$ is the Sharpe ratio of the *ex-post* tangency portfolio (maximum ex-post Sharpe ratio) formed from the test assets plus the factor f.

If there are many factors that are excess returns, the same ideas work, with some cost of algebraic complexity. The regression equation is

$$R^{ei}=\alpha_i+\boldsymbol{\beta}_i'\mathbf{f}_t+\varepsilon_t^i.$$

The asset pricing model

$$E(R^{ei}) = \boldsymbol{\beta}_i' E(\mathbf{f})$$

again predicts that the intercepts should be zero. We can estimate α and β with OLS time-series regressions. Assuming normal i.i.d. errors, the quadratic form $\hat{\boldsymbol{\alpha}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\alpha}}$ has the distribution,

$$\frac{T-N-K}{N}\left(1+E_T(\mathbf{f})'\hat{\Omega}^{-1}E_T(\mathbf{f})\right)^{-1}\ \hat{\boldsymbol{\alpha}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\alpha}}\ \sim F_{N,T-N-K} \qquad (124)$$

where

$$\begin{aligned} N &= \text{Number of assets} \\ K &= \text{Number of factors} \\ \hat{\Omega} &= \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{f}_t - E_T(\mathbf{f})\right]\left[\mathbf{f}_t - E_T(\mathbf{f})\right]' \end{aligned}$$

The main difference is that the Sharpe ratio of the single factor is replaced by the natural generalization $E_T(\mathbf{f})'\hat{\Omega}^{-1}E_T(\mathbf{f})$.

12.1.1 Derivation of (122).

You can easily derive (122) by following the standard OLS approach to the covariance matrix of estimated parameters. However, it is simpler and more elegant to derive (122) as an instance of GMM, and this approach allows us to generate straightforwardly the required corrections for autocorrelated and heteroskedastic disturbances. (MacKinlay and Richardson (1991) advocate GMM approaches to regression tests in this way.) The mechanics are only slightly different than what we did to generate distributions for OLS regression coefficients in section xx, since we keep track of N OLS regressions simultaneously.

Write the equations for all N assets together in vector form,

$$\mathbf{R}_t^e = \boldsymbol{\alpha} + \boldsymbol{\beta} f_t + \boldsymbol{\varepsilon}_t.$$

We use the usual OLS moments to estimate the coefficients,

$$g_T(\mathbf{b}) = \left[\begin{array}{c} E_T\left(\mathbf{R}_t^e - \boldsymbol{\alpha} - \boldsymbol{\beta} f_t\right) \\ E_T\left[\left(\mathbf{R}_t^e - \boldsymbol{\alpha} - \boldsymbol{\beta} f_t\right) f_t\right] \end{array}\right] = E_T\left(\left[\begin{array}{c} \boldsymbol{\varepsilon}_t \\ f_t\boldsymbol{\varepsilon}_t \end{array}\right]\right) = 0$$

These moments exactly identify the parameters, so the a matrix in $ag_T(\hat{\mathbf{b}}) = 0$ is the identity matrix. Solving, the GMM estimates are of course the OLS estimates,

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= E_T\left(\mathbf{R}_t^e\right) - \hat{\boldsymbol{\beta}} E_T(f_t) \\ \hat{\boldsymbol{\beta}} &= \frac{E_T\left[\left(\mathbf{R}_t^e - E_T\left(\mathbf{R}_t^e\right)\right) f_t\right]}{E_T\left[\left(f_t - E_T(f_t)\right) f_t\right]} = \frac{cov_T(\mathbf{R}_t^e, f_t)}{var_T(f_t)}. \end{aligned}$$

The d matrix in the general GMM formula is

$$d \equiv \frac{\partial g_T(b)}{\partial b'} = -\left[\begin{array}{cc} I_N & I_N E_T(f_t) \\ I_N E_T(f_t) & I_N E_T(f_t^2) \end{array}\right] = -\left[\begin{array}{cc} 1 & E(f_t) \\ E(f_t) & E(f_t^2) \end{array}\right] \otimes I_N$$

where I_N is an $N \times N$ identity matrix. The S matrix is

$$S \equiv \sum_{j=-\infty}^{\infty} E\left[f(x_t, b),\ f(x_{t-j}b)'\right] = \sum_{j=-\infty}^{\infty} \left[\begin{array}{cc} E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-j}') & E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-j}' f_{t-j}) \\ E(f_t \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-j}') & E(f_t \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-j}' f_{t-j}) \end{array}\right]$$

If we assume that f and ε are independent,

$$S = \sum_{j=-\infty}^{\infty} \left[\begin{array}{cc} 1 & E(f_t) \\ E(f_t) & E(f_t f_{t-j}) \end{array}\right] \otimes E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-j}'). \qquad (125)$$

If we assume that the errors and factors are not correlated over time,

$$S \equiv \left[\begin{array}{cc} E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') & E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' f_t) \\ E(f_t \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') & E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-j}') f_t^2 \end{array}\right] \qquad (126)$$

And if we assume that f and ε are both independent and uncorrelated over time,

$$S = \left[\begin{array}{cc} 1 & E(f_t) \\ E(f_t) & E(f_t^2) \end{array}\right] \otimes \Sigma \qquad (127)$$

Now we can plug into the general variance-covariance matrix formula (101),

$$var(\hat{b}) = \frac{1}{T}(ad)^{-1} a S a' (ad)^{-1\prime}.$$

Using the case (127), we obtain[8]

$$var\left(\begin{array}{c} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{array}\right) = \frac{1}{T}\left(\left[\begin{array}{cc} 1 & E(f_t) \\ E(f_t) & E(f_t^2) \end{array}\right]^{-1} \otimes \Sigma\right) = \frac{1}{T}\left(\frac{1}{var(f)}\left[\begin{array}{cc} E(f_t^2) & -E(f_t) \\ -E(f_t) & 1 \end{array}\right] \otimes \Sigma\right)$$

We're interested in the top left corner. Using $E(f^2) = E(f)^2 + var(f)$,

$$var\left(\hat{\boldsymbol{\alpha}}\right) = \frac{1}{T}\left(1 + \frac{E(f)^2}{var(f)}\right)\Sigma.$$

This is the traditional formula, but there is now no real reason to assume that the errors are i.i.d. By simply calculating a sample version of (125), (126), (127), we can easily construct standard errors and test statistics that do not require these assumptions.

[8] You need $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ if you keep the simplifying $\otimes$ notation.

12.2 Cross-sectional regressions

We can fit

$$E(R^{ei}) = \boldsymbol{\beta}_i'\boldsymbol{\lambda} + \boldsymbol{\alpha}_i$$

by running a *cross-sectional* regression of average returns on the betas. This technique can be used whether the factor is a return or not.

I discuss OLS and GLS cross-sectional regressions, I find formulas for the standard errors of $\lambda,$ and a χ^2 test whether the α are jointly zero. OLS formulas for the for the standard errors of λ and α ignore the fact that β is also random. I derive Shanken's correction for this fact as an instance of GMM, and show how to implement the same approach for autocorrelated and heteroskedastic errors. I show how the results are almost identical to the GMM, $m = a + \mathbf{b}'\mathbf{f}$ formulation of linear factor models derived in section 4.

Start again with the a K factor model, written as

$$E(R^{ei}) = \boldsymbol{\beta}_i'\boldsymbol{\lambda};\, i = 1, 2, ...N$$

The central economic idea is that the model should explain why average returns vary across assets; expected returns of an asset should be high if that asset has high betas or risk exposure to factors that carry high risk premia.

Figure 20 graphs the case of a single factor such as the CAPM. Each dot represents one asset i. The model says that average returns should be proportional to betas, so plot the sample average returns against the betas. Even if the model is true, this plot will not work out perfectly in each sample, so there will be some spread as shown.

Given these ideas, a natural idea is to run a *cross-sectional regression* to fit a line through the scatterplot of Figure 20. First find estimates of the betas from a time series regression,

$$R_t^{ei} = a_i + \boldsymbol{\beta}_i'\mathbf{f}_t + \varepsilon_t^i,\ \ t = 1, 2, ...T \text{ for each } i,$$

and then estimate the factor risk premia $\boldsymbol{\lambda}$ from a regression across assets of average returns on the betas,

$$E_T(R^{ei}) = \boldsymbol{\beta}_i'\boldsymbol{\lambda} + \alpha_i,\ i = 1, 2....N. \tag{128}$$

As in the figure, $\boldsymbol{\beta}$ are the right hand variables, $\boldsymbol{\lambda}$ are the regression coefficients, and the cross-sectional regression residuals α_i are the pricing errors. One can run the cross-sectional regression with or without a constant. The theory says that the constant or zero-beta excess return should be zero. One can impose this restriction or estimate with a constant and then

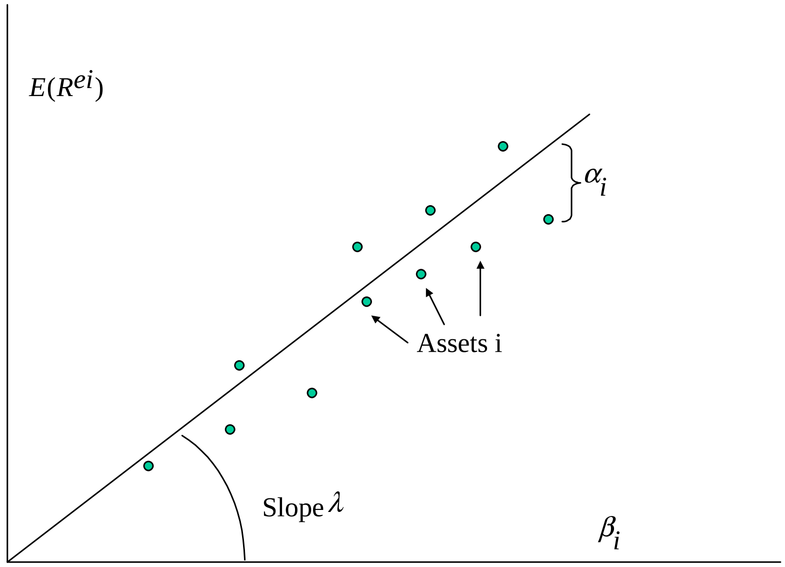


Figure 20. Cross-sectional regression

see if it comes out sufficiently small. Importantly, one can run the cross-sectional regression when the factor is not a return.

12.2.1 OLS cross-sectional regression

It will simplify notation to consider a single factor; the case of multiple factors looks the same with vectors in place of scalars. Denote vectors from 1 to N with boldface, i.e. $\boldsymbol{\varepsilon}_t = \left[\begin{array}{cccc} \varepsilon_t^1 & \varepsilon_t^2 & \cdots & \varepsilon_t^N \end{array}\right]'$, $\boldsymbol{\beta} = \left[\begin{array}{cccc} \beta_1 & \beta_2 & \cdots & \beta_N \end{array}\right]'$, and similarly for $\mathbf{R}_t^e$ and $\boldsymbol{\alpha}$. For simplicity take the case of no intercept. With this notation OLS cross-sectional estimates are

$$\begin{aligned} \hat{\lambda} &= \left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}' E_T(\mathbf{R}^e) \\ \hat{\boldsymbol{\alpha}} &= E_T(\mathbf{R}^e) - \hat{\lambda}\boldsymbol{\beta}. \end{aligned} \tag{129}$$

Next, we need a distribution theory for the estimated parameters. The most natural thing to do is to apply the standard OLS distribution formulas. I start with the traditional assumption that the errors are i.i.d. over time, and independent of the factors. Denote $\Sigma = E\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right)$. Since the α_i are just time series averages of the ε_t^i shocks, the errors in the cross-sectional regression have correlation matrix $E\left(\boldsymbol{\alpha}\boldsymbol{\alpha}'\right) = \frac{1}{T}\Sigma$. Thus the conventional OLS formulas for the covariance matrix of OLS estimates and residual with correlated errors give

$$\sigma^2\left(\hat{\lambda}\right) = \frac{1}{T}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1} \tag{130}$$

$$cov(\hat{\boldsymbol{\alpha}}) = \frac{1}{T}\left(I - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right)\Sigma\left(I - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right) \tag{131}$$

To rederive these formulas, substitute the regression (20) into the formulas for parameter estimates (129) and take expectations. See (137) below before you use them..

We could test whether all pricing errors are zero with the statistic

$$\hat{\boldsymbol{\alpha}}' cov(\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}} \sim \chi^2_{N-1}. \tag{132}$$

The distribution is χ^2_{N-1} not χ^2_N because the covariance matrix is singular. The singularity and the extra terms in (131) result from the fact that the λ coefficient was estimated along the way, and means that we have to use a generalized inverse. (If there are K factors, we obviously end up with χ^2_{N-K}.)

12.2.2 GLS cross-sectional regression

Since the residuals in the cross-sectional regression (20) are correlated with each other, standard textbook advice is to run a GLS cross-sectional regression rather than OLS, using

$E(\boldsymbol{\alpha}\boldsymbol{\alpha}') = \frac{1}{T}\Sigma$ as the error covariance matrix:

$$\begin{aligned} \hat{\lambda} &= \left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\Sigma^{-1}E_T(\mathbf{R}^e) \qquad (133) \\ \hat{\boldsymbol{\alpha}} &= E_T(\mathbf{R}^e) - \hat{\lambda}\boldsymbol{\beta}. \end{aligned}$$

The standard regression formulas give the variance of these estimates as

$$\sigma^2\left(\hat{\lambda}\right) = \frac{1}{T}\left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1} \qquad (134)$$

$$cov(\hat{\boldsymbol{\alpha}}) = \frac{1}{T}\left(\Sigma - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right) \qquad (135)$$

The comments of section 2 warning that OLS is sometimes much more robust than GLS apply equally in this case. The GLS regression should improve efficiency, i.e. give more precise estimates. However, Σ may be hard to estimate and to invert, especially if the cross-section N is large. One may well choose the robustness of OLS over the asymptotic statistical advantages of GLS.

A GLS regression can be understood as a transformation of the space of returns, to focus attention on the statistically most informative portfolios. Finding (say, by Choleski decomposition) a matrix C such that $CC' = \Sigma^{-1}$, the GLS regression is the same as an OLS regression of $CE_T(\mathbf{R}^e)$ on $C\boldsymbol{\beta}$, i.e. of testing the model on the portfolios $C\mathbf{R}^e$. The statistically most informative portfolios are those with the lowest residual variance Σ, therefore GLS pays most attention to nearly riskfree portfolios formed by extreme long and short positions. The statistical theory assumes that the covariance matrix has converged to its true value. However, in most samples, the ex-post mean-variance frontier still seems to indicate lots of luck, and this is especially true if the cross section is large, anything more than 1/10 of the time series. If this is true, the GLS regression is paying lots of attention to nearly riskless portfolios that only seem so due to luck in a specific sample.

Again, we could test the hypothesis that all the α are equal to zero with (132). (Though the appearance of the statistic is the same, the covariance matrix is smaller, reflecting the greater power of the GLS test.) As with the J_T test, (106) we can develop an equivalent test that does not require a generalized inverse;

$$T\hat{\boldsymbol{\alpha}}'\Sigma^{-1}\hat{\boldsymbol{\alpha}} \sim \chi^2_{N-1}. \qquad (136)$$

To derive (136), I proceed exactly as in the derivation of the J_T test (106). Define, say by Choleski decomposition, a matrix C such that $CC' = \Sigma$. Now, find the covariance matrix of $\sqrt{T}C^{-1}\hat{\boldsymbol{\alpha}}$:

$$cov(\sqrt{T}C^{-1}\alpha) = C^{-1}\left(CC' - \boldsymbol{\beta}\left(\boldsymbol{\beta}'C^{-1\prime}C^{-1}\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right)C^{-1\prime} = I - \boldsymbol{\delta}\left(\boldsymbol{\delta}'\boldsymbol{\delta}\right)^{-1}\boldsymbol{\delta}\prime$$

where

$$\boldsymbol{\delta} = C^{-1}\boldsymbol{\beta}.$$

In sum, $\hat{\boldsymbol{\alpha}}$ is asymptotically normal so $\sqrt{T}C^{-1}\hat{\boldsymbol{\alpha}}$ is asymptotically normal, $cov(\sqrt{T}C^{-1}\hat{\boldsymbol{\alpha}})$ is an idempotent matrix with rank $N-1$; therefore $T\hat{\boldsymbol{\alpha}}'C^{-1\prime}C^{-1}\hat{\boldsymbol{\alpha}} = T\hat{\boldsymbol{\alpha}}'\Sigma^{-1}\hat{\boldsymbol{\alpha}}$ is χ^2_{N-1}.

12.2.3 Correction for the fact that $\boldsymbol{\beta}$ are estimated, and GMM formulas that don't need i.i.d. errors.

In applying standard OLS formulas to a cross-sectional regression, we assume that the right hand variables $\boldsymbol{\beta}$ are fixed. The $\boldsymbol{\beta}$ in the cross-sectional regression are not fixed, of course, but are estimated in the time series regression. This turns out to matter, even asymptotically.

In this section, I derive the correct asymptotic standard errors. With the simplifying assumption that the errors ε are i.i.d. and independent of the factors, the result is

$$\begin{aligned}
\sigma^2(\hat{\lambda}_{OLS}) &= \frac{1}{T}\left[(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\left(1+\boldsymbol{\lambda}'\Sigma_f^{-1}\boldsymbol{\lambda}\right)+\Sigma_f\right] \qquad (137)\\
\sigma^2(\hat{\lambda}_{GLS}) &= \frac{1}{T}\left[\left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1}\left(1+\boldsymbol{\lambda}'\Sigma_f^{-1}\boldsymbol{\lambda}\right)+\Sigma_f\right]
\end{aligned}$$

where Σ_f is the variance-covariance matrix of the factors. This correction is due to Shanken (1992). Comparing these standard errors to (130) and (134), we see that there is a multiplicative correction $\left(1+\boldsymbol{\lambda}'\Sigma_f^{-1}\boldsymbol{\lambda}\right)$ and an additive correction Σ_f that do not vanish, even asymptotically.

The asymptotic variance-covariance matrix of the pricing errors is

$$\begin{aligned}
cov(\hat{\boldsymbol{\alpha}}_{OLS}) &= \frac{1}{T}\left(I_N-\boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right)\Sigma\left(I_N-\boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\right)\left(1+\boldsymbol{\lambda}'\Sigma_f^{-1}\boldsymbol{\lambda}\right) \qquad (138)\\
cov(\hat{\boldsymbol{\alpha}}_{GLS}) &= \frac{1}{T}\left(\Sigma-\boldsymbol{\beta}\left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right)\left(1+\boldsymbol{\lambda}'\Sigma_f^{-1}\boldsymbol{\lambda}\right) \qquad (139)
\end{aligned}$$

Comparing these results to (131) and (135) we see the same multiplicative correction applies.

We can form the asymptotic χ^2 test of the pricing errors by dividing pricing errors by their variance-covariance matrix, $\hat{\boldsymbol{\alpha}}cov(\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}$. Following (136), we can simplify this result for the GLS pricing errors resulting in

$$T\left(1+\boldsymbol{\lambda}'\Sigma_f^{-1}\boldsymbol{\lambda}\right)\hat{\boldsymbol{\alpha}}'_{GLS}\Sigma^{-1}\hat{\boldsymbol{\alpha}}_{GLS}\sim\chi^2_{N-K}. \qquad (140)$$

Are the corrections important relative to the simple OLS formulas given above? In the CAPM $\lambda = E(R^{em})$ so $\lambda^2/\sigma^2(R^{em}) \approx (0.08/0.16)^2 = 0.25$ in annual data so the multiplicative term is too large to ignore. However, the mean and variance both scale with horizon so for a monthly interval $\lambda^2/\sigma^2(R^{em}) \approx 0.25/12 \approx 0.02$ which is quite small and ignoring the multiplicative term makes little difference.

The additive term can be very important. Consider a one factor model, suppose all the $\boldsymbol{\beta}$ are 1.0, all the residuals are uncorrelated so Σ is diagonal, suppose all assets have the same residual covariance $\sigma^2(\varepsilon)$, and ignore the multiplicative term. Now we can write either covariance matrix in (137) as

$$\sigma^2(\hat{\lambda}) = \frac{1}{T}\left[\frac{1}{N}\sigma^2(\varepsilon) + \sigma^2(f)\right]$$

Even with $N = 1$, most factor models have fairly high R^2, so $\sigma^2(\varepsilon) < \sigma^2(f)$. Typical CAPM values of $R^2 = 1 - \sigma^2(\varepsilon)/\sigma^2\,(f)$ for large portfolios are 0.6-0.7; and multifactor models such as the Fama French 3 factor model have R^2 often over 0.9. Typical numbers of assets $N = 10$ to 50 make the first term vanish compared to the second term.

This example suggests that not only is including Σ_f likely to be an important correction, it may even be the dominant consideration in the sampling error of the $\hat{\lambda}$. Interestingly, and despite the fact that these corrections are easy to make and have been known for almost 20 years, they are very infrequently used.

Comparing (140) to the GRS tests for a time-series regression, (122), (123), (124) we see the same statistic. The only difference is that by estimating $\boldsymbol{\lambda}$ from the cross-section rather than imposing $\lambda = E(f)$, the cross-sectional regression loses degrees of freedom equal to the number of factors. A purely statistical approach will seize on this difference and advocate the GRS test when it can be applied, though we will see later that the cross-sectional regression may be more robust to misspecifications.

Comparing both the standard errors of λ and the covariance matrix of the pricing errors to the GMM results for $p = E(mx),\ m = a + bf$ representation of a linear factor model in section 4, you see that the formulas are almost exactly identical. The $p = E(mx)$ formulation of the model for excess returns was equivalent to $E(R^e) = -Cb$ where C is the *covariance* between returns and factors; thus covariances C enter in place of betas β. The S matrix enters in place of Σ, but that is because the above formulas have assumed i.i.d. errors; when we drop this assumption below we will get formulas that look even more similar. Thus, the GMM, $p = E(mx)$ approach reduces almost exactly to this traditional approach for linear factor models, excess returns, and i.i.d. errors. The only real difference is whether you want to express the covariance between returns and factors in regression coefficient units or just covariances. I have argued above that covariances and hence b is more interesting than β, λ, since b measures whether a factor is useful in pricing assets while λ measures whether a factor is priced.

Derivation and formulas that don't require i.i.d. errors.

The easy and elegant way to account for the effects of "generated regressors" such as the $\boldsymbol{\beta}$ in the cross-sectional regression is to map the whole thing into GMM. Then, we treat the moments that generate the regressors $\boldsymbol{\beta}$ at the same time as the moments that generate the cross-sectional regression coefficient λ, and the covariance matrix S between the two sets of moments captures the effects of generating the regressors on the standard error of the cross-sectional regression coefficients. Comparing this straightforward derivation with the difficulty of Shanken's (1992) paper that originally derived the corrections for $\hat{\lambda}$, and noting that Shanken did not go on to find the formulas (138) that allow a test of the pricing errors is a nice argument for the simplicity and power of the GMM framework.

To keep the algebra manageable, I treat the case of a single factor. The moments are

$$g_T(\mathbf{b}) = \left[\begin{array}{c} E(\mathbf{R}^e_t - \mathbf{a} - \boldsymbol{\beta} f_t) \\ E\left[(\mathbf{R}^e_t - \mathbf{a} - \boldsymbol{\beta} f_t) f_t\right] \\ E\left(\mathbf{R}^e - \boldsymbol{\beta}\lambda\right) \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \end{array}\right] \tag{141}$$

The top two moment conditions exactly identify $\mathbf{a}$ and $\boldsymbol{\beta}$ as the time-series OLS estimates. (Note $\mathbf{a}$ not $\boldsymbol{\alpha}$. The time-series intercept is not necessarily equal to the pricing error in a cross-sectional regression.) The bottom moment condition is the asset pricing model. It is in general overidentified in a sample, since there is only one extra parameter (λ) and N extra moment conditions. If we use a weighting vector $\boldsymbol{\beta}'$ on this condition, we obtain the OLS cross-sectional estimate of λ. If we use a weighting vector $\boldsymbol{\beta}'\Sigma^{-1}$, we obtain the GLS cross-sectional estimate of λ. To accommodate both cases, use a weighting vector $\boldsymbol{\gamma}'$, and then substitute either $\boldsymbol{\gamma}' = \boldsymbol{\beta}'$ or $\boldsymbol{\gamma}' = \boldsymbol{\beta}'\Sigma^{-1}$ at the end to get OLS and GLS results.

The correct standard errors for $\hat{\lambda}$ come straight from the general GMM standard error formula (101). The $\hat{\boldsymbol{\alpha}}$ are not parameters, but are the last N moments. Their covariance matrix is thus given by the GMM formula (104) for the sample variation of the g_T. All we have to do is map the problem into the GMM notation. The parameter vector is

$$\mathbf{b}' = \left[\begin{array}{ccc} \mathbf{a}' & \boldsymbol{\beta}' & \lambda \end{array}\right]$$

The a matrix chooses which moment conditions are set to zero in estimation,

$$a = \left[\begin{array}{cc} I_{2N} & 0 \\ 0 & \boldsymbol{\gamma}' \end{array}\right].$$

The d matrix is the sensitivity of the moment conditions to the parameters,

$$d = \frac{\partial g_T}{\partial \mathbf{b}'} = \left[\begin{array}{ccc} -I_N & -I_N E(f) & 0 \\ -I_N E(f) & -I_N E(f^2) & 0 \\ 0 & -\lambda I_N & -\boldsymbol{\beta} \end{array}\right]$$

The S matrix is the long-run covariance matrix of the moments.

$$S = \sum_{j=-\infty}^{\infty} E\left(\left[\begin{array}{c} \mathbf{R}^e_t - \mathbf{a} - \boldsymbol{\beta} f_t \\ (\mathbf{R}^e_t - \mathbf{a} - \boldsymbol{\beta} f_t) f_t \\ \mathbf{R}^e_t - \boldsymbol{\beta}\lambda \end{array}\right]\left[\begin{array}{c} \mathbf{R}^e_{t-j} - \mathbf{a} - \boldsymbol{\beta} f_{t-j} \\ (\mathbf{R}^e_{t-j} - \mathbf{a} - \boldsymbol{\beta} f_{t-j}) f_{t-j} \\ \mathbf{R}^e_{t-j} - \boldsymbol{\beta}\lambda \end{array}\right]'\right)$$

To evaluate this expression, substitute $\boldsymbol{\varepsilon}_t = \mathbf{R}^e_t - \mathbf{a} - \boldsymbol{\beta} f_t$. Also, write $\mathbf{R}^e_t - \boldsymbol{\beta}\lambda = \mathbf{a} + \boldsymbol{\beta}(f_t - \lambda) + \boldsymbol{\varepsilon}_t$.

The expression simplifies with the assumption of i.i.d. errors independent of the factors. The assumption that the errors are i.i.d. over time means we can ignore the lead and lag terms. Thus, the top left corner is $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t) = \Sigma$. The assumption that the errors are independent from the factors f_t simplifies the terms in which ε_t and f_t are multiplied: $E(\varepsilon_t (\varepsilon_t' f_t)) = E(f)\Sigma$ for example. The result is

$$S = \begin{bmatrix} \Sigma & E(f)\Sigma & \Sigma \\ E(f)\Sigma & E(f^2)\Sigma & E(f)\Sigma \\ \Sigma & E(f)\Sigma & \boldsymbol{\beta}\boldsymbol{\beta}'\sigma^2(f)+\Sigma \end{bmatrix}$$

Multiplying a, d, S together as specified by the GMM formula for the covariance matrix of parameters (101) we obtain the covariance matrix of all the parameters, and its (3,3) element gives the variance of $\hat{\lambda}$. Multiplying the terms together as specified by (104), we obtain the sampling distribution of the $\hat{\boldsymbol{\alpha}}$, (138). The formulas (137) reported above are derived the same way with a vector of factors $\mathbf{f}_t$ rather than a scalar; the second moment condition in (141) then reads $E\left[(\mathbf{R}_t^e - \mathbf{a} - \boldsymbol{\beta}\mathbf{f}_t) \otimes \mathbf{f}_t\right]$.

Once again, there is really no need to make the assumption that the errors are i.i.d. and especially that they are conditionally homoskedastic – that the factor f and errors ε are independent. It is quite easy to estimate an S matrix that does not impose these conditions and calculate standard errors. They will not have the pretty analytic form given above, but they will more closely report the true sampling uncertainty of the estimate.

12.3 Fama-MacBeth Procedure

I introduce the Fama-MacBeth procedure for running cross sectional regression and show that it is numerically equivalent to pooled time-series, cross-section OLS with standard errors corrected for cross-sectional correlation, and also to a single cross-sectional regression on time-series averages with standard errors corrected for cross-sectional correlation.

Fama and MacBeth (1972) suggest an alternative procedure for running cross-sectional regressions, and for producing standard errors and test statistics. This is a historically important procedure, and is still widely used (especially by Fama and coauthors), so it is important to understand it and relate it to other procedures. First, instead of estimating a single cross-sectional regression with the sample averages, they suggest we run a cross-sectional regression *at each time period*, i.e.

$$R_t^{ei} = \beta_i' \lambda_t + \alpha_{it} \;\; i = 1, 2, ...N \text{ for each } t.$$

(I write the case of a single factor for simplicity, but it's easy to extend the model to multiple factors.) Fama and MacBeth use five year rolling regression betas at this stage, but one can also use betas from the full-sample time-series regression. Then, they suggest that we estimate λ and α_i as the average of the cross sectional regression estimates,

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\hat{\lambda}_t;\ \hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\alpha}_{it}.$$

Most importantly, they suggest that we use the standard deviations of the cross-sectional regression estimates to generate the sampling errors for these estimates,

$$\sigma^2(\hat{\lambda}) = \frac{1}{T^2}\sum_{t=1}^{T}\left(\hat{\lambda}_t - \hat{\lambda}\right)^2;\ \sigma^2(\hat{\alpha}_i) = \frac{1}{T^2}\sum_{t=1}^{T}\left(\hat{\alpha}_{it} - \hat{\alpha}_i\right)^2.$$

It's $1/T^2$ because we're finding standard errors of sample means, σ^2/T

This is an intuitively appealing procedure once you stop to think about it. Sampling error is, after all, abut how a statistic would vary from one sample to the next if we repeated the observations. We can't do that with only one sample, but why not cut the sample in half, and deduce how a statistic would vary from one full sample to the next from how it varies from the first half of the sample to the next half? Proceeding, why not cut the sample in fourths, eights and so on? The Fama-MacBeth procedure carries this idea to is logical conclusion, using the variation in the statistic at each point in time to deduce its sampling variation.

We are used to deducing the sampling variance of the sample mean of a series x_t by looking at the variation of x_t through time in the sample, using $\sigma^2(\bar{x}) = \sigma^2(x)/T = \frac{1}{T^2}\sum_t (x_t - \bar{x})^2$. The Fama-MacBeth technique just applies this idea to the slope and pricing error estimates. This procedure assumes that the time series is not autocorrelated, but one could easily extend the idea to estimate the sampling variation of a sample mean using a long run variance matrix, i.e. estimate .

$$\sigma^2(\hat{\lambda}) = \frac{1}{T}\sum_j \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\lambda}_t - \hat{\lambda}\right)\left(\hat{\lambda}_{t-j} - \hat{\lambda}\right)$$

and similarly for $\hat{\alpha}$. Asset return data are usually not highly correlated, but this could have a big effect on the application of the Fama-MacBeth technique to corporate finance data or other regressions in which the cross-sectional estimates are highly correlated over time.

It is natural to use this sampling theory to test whether all the pricing errors are jointly zero as we have before. Denote by $\boldsymbol{\alpha}$ the vector of pricing errors across assets; estimate the covariance matrix of the sample pricing errors by

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{\alpha}}_t \\ cov(\hat{\boldsymbol{\alpha}}) &= \frac{1}{T^2}\sum_{t=1}^{T}\left(\hat{\boldsymbol{\alpha}}_t - \hat{\boldsymbol{\alpha}}\right)\left(\hat{\boldsymbol{\alpha}}_t - \hat{\boldsymbol{\alpha}}\right)'\end{aligned}$$

and then use the test

$$\hat{\boldsymbol{\alpha}}' cov(\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}} \sim \chi^2_{N-1}.$$

12.3.1 Fama MacBeth in depth

The GRS procedure and the formulas given above for a single cross-sectional regression are familiar from any course in regression. The Fama MacBeth procedure seems novel, and it is a useful and simple technique that can be widely used in economics and corporate finance as well as asset pricing. Is it truly different? Is there something different about asset pricing data that requires a fundamentally new technique not taught in standard regression courses? To answer these questions it is worth looking in a little more detail at what it accomplishes and why.

Consider a regression

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} \; i = 1, 2, ...N; \; t = 1, 2, ...T.$$

The data in this regression has a cross-sectional element as well as a time-series element. In corporate finance, for example, one might be interested in the relationship between investment and financial variables, and the data set has many firms (N) as well as time series observations for each firm (T). This expression is the same form as our asset pricing model, with x_{it} standing for the β_i and β standing for λ.

The textbook thing to do in this context is to simply stack the i and t observations together and estimate $\boldsymbol{\beta}$ by OLS. I will call this the *pooled time-series cross-section estimate.* However, the error terms are not likely to be uncorrelated with each other. In particular, the error terms are likely to be cross-sectionally correlated at a given time. If one return is unusually high, another is also likely to be high; if one firm invests an unusually great amount this year, another is also likely to do so. When errors are not uncorrelated, OLS is still consistent, but the OLS distribution theory is wrong, and typically suggests standard errors that are much too small. In the extreme case that the N errors are perfectly correlated at each time period, it is as if there is only one observation for each time period, so one really has T rather than NT observations. Therefore, a real pooled time-series cross-section estimate must include corrected standard errors. People often ignore this fact and report OLS standard errors.

Another thing we could do is first take time series averages and then run a *pure cross-sectional* regression of

$$E_T(y_{it}) = \boldsymbol{\beta}' E_T(\mathbf{x}_{it}) + u_i \; i = 1, 2, ...N$$

This would lose any information due to variation of the $\mathbf{x}_{it}$ over time, but at least it might be easier to figure out a variance-covariance matrix for u_i and correct the standard errors for residual correlation. (You could also average cross-sectionally and than run a single time-series regression. We'll get to that option later.)

In either case, the standard error corrections are just applications of the standard formula: for an OLS regression

$$Y = X\beta + u; \; E(uu') = \Omega$$

the standard errors of the OLS estimate

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

are

$$\sigma^2(\hat{\beta}_{OLS}) = (X'X)^{-1} \; X'\Omega X \; (X'X)^{-1}.$$

Finally, we could run the Fama-MacBeth procedure: run a cross-sectional regression at each point in time; average the cross-sectional $\hat{\boldsymbol{\beta}}_t$ estimates to get an estimate $\hat{\boldsymbol{\beta}}$, and use the time-series standard deviation of $\hat{\boldsymbol{\beta}}_t$ to estimate the standard error of $\hat{\boldsymbol{\beta}}$.

It turns out that the Fama MacBeth procedure is just another way of calculating the standard errors, corrected for cross-sectional correlation.

Proposition 6 *If the $\mathbf{x}_{it}$ variables do not vary over time, and if the errors are cross-sectionally correlated but not correlated over time, then the Fama-MacBeth estimate, the pure cross-sectional OLS estimate and the pooled time-series cross-sectional OLS estimates are identical. Also, the Fama-MacBeth standard errors are identical to the cross-sectional regression or stacked OLS standard errors, corrected for residual correlation. None of these relations hold if the $\mathbf{x}$ vary through time.*

Since they are identical procedures, whether one calculates estimates and standard errors in one way or the other is a matter of taste.

I emphasize one procedure that is incorrect: pooled time series and cross section OLS with no correction of the standard errors. The errors are so highly cross-sectionally correlated in most finance applications that the standard errors so computed are often off by a factor of 10.

The assumption that the errors are not correlated over time is probably not so bad for asset pricing applications, since returns are close to independent. However, when pooled time-series cross-section regressions are used in corporate finance applications, errors are likely to be as severely correlated over time as across firms, if not more so. The "other factors" (ε) that cause, say, company i to invest more at time t than predicted by a set of right hand variables is surely correlated with the other factors that cause company j to invest more. But such factors are especially likely to cause company i to invest more tomorrow as well. In this case, any standard errors must also correct for serial correlation in the errors; the GMM based formulas in section 3 can do this easily.

The Fama-MacBeth standard errors also do not correct for the fact that $\hat{\beta}$ are generated regressors. If one is going to use them, it is a good idea to at least calculate the Shanken correction factors outlined above. Again, the GMM setup used above to derive the Shanken corrections makes this easy.

Proof: We just have to write out the three approaches and compare them. Having assumed that the x variables do not vary over time, the regression is

$$y_{it} = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{it}.$$

We can stack up the cross-sections $i = 1...N$ and write the regression as

$$\mathbf{y}_t = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t.$$

$\mathbf{x}$ is now a matrix with the $\mathbf{x}_i'$ as rows. The error assumptions mean $E(\varepsilon_t\varepsilon_t') = \Sigma$.

Pooled OLS: To run pooled OLS, we stack the time series and cross sections by writing

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{bmatrix};\ \mathbf{X} = \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \\ \vdots \\ \mathbf{x} \end{bmatrix};\ \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{bmatrix}$$

and then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with

$$E(\epsilon\epsilon') = \Omega = \begin{bmatrix} \Sigma & & \\ & \ddots & \\ & & \Sigma \end{bmatrix}$$

The estimate and its standard error is then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\ cov(\hat{\boldsymbol{\beta}}_{OLS}) &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\Omega\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned}$$

Writing this out from the definitions of the stacked matrices, with $\mathbf{X}'\mathbf{X} = T\mathbf{x}'\mathbf{x}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t \\ cov(\hat{\boldsymbol{\beta}}_{OLS}) &= \frac{1}{T}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\left(\mathbf{x}'\Sigma\mathbf{x}\right)\left(\mathbf{x}'\mathbf{x}\right)^{-1}. \end{aligned}$$

We can estimate this sampling variance with

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'; \\ \hat{\boldsymbol{\varepsilon}}_t &\equiv \mathbf{y}_t - \mathbf{x}\hat{\boldsymbol{\beta}}_{OLS} \end{aligned} \tag{142}$$

Pure cross-section: The pure cross-sectional estimator does one cross-sectional regression of the time-series averages. So, take those averages,

$$E_T\left(\mathbf{y}_t\right) = \mathbf{x}\boldsymbol{\beta} + E_T\left(\boldsymbol{\varepsilon}_t\right)$$

where $E_T = \frac{1}{T}\sum_{t=1}^{T}$ and $\mathbf{x} = E_T(\mathbf{x}\ )$ since $\mathbf{x}$ is constant. Having assumed i.i.d. errors over time, the error covariance matrix is

$$E\left(E_T\left(\boldsymbol{\varepsilon}_t\right)E_T\left(\boldsymbol{\varepsilon}_t'\right)\right) = \frac{1}{T}\Sigma.$$

The cross sectional estimate and corrected standard errors are then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{XS} &= \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'E_T(\mathbf{y}_t) \\ \sigma^2(\hat{\boldsymbol{\beta}}_{XS}) &= \frac{1}{T}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\Sigma\mathbf{x}^{-1}\left(\mathbf{x}'\mathbf{x}\right)^{-1} \end{aligned}$$

Thus, the cross-sectional and pooled OLS estimates and standard errors are exactly the same, in each sample.

Fama-MacBeth: The Fama–MacBeth estimator is formed by first running the cross-sectional regression at each moment in time,

$$\hat{\boldsymbol{\beta}}_t = \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{y}_t.$$

Then the estimate is the average of the cross-sectional regression estimates,

$$\hat{\boldsymbol{\beta}}_{FM} = E_T\left(\hat{\boldsymbol{\beta}}_t\right) = \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'E_T\left(\mathbf{y}_t\right).$$

Thus, the Fama-MacBeth estimator is also the same as the OLS estimator, in each sample. The Fama-MacBeth standard error is based on the time-series standard deviation of the $\hat{\boldsymbol{\beta}}_t$. Using cov_T to denote sample covariance,

$$cov\left(\hat{\boldsymbol{\beta}}_{FM}\right) = \frac{1}{T}cov_T\left(\hat{\boldsymbol{\beta}}_t\right) = \frac{1}{T}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'cov_T\left(\mathbf{y}_t\right)\mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}.$$

with

$$\mathbf{y}_t = \mathbf{x}\boldsymbol{\beta}_{FM} + \hat{\boldsymbol{\varepsilon}}_t$$

we have

$$cov_T\left(\mathbf{y}_t\right) = E_T(\hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t') = \hat{\Sigma}$$

and finally

$$cov\left(\hat{\boldsymbol{\beta}}_{FM}\right) = \frac{1}{T}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\hat{\Sigma}\mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}.$$

Thus, the FM estimator of the standard error is also numerically equivalent to the OLS corrected standard error.

Varying x If the x_{it} vary through time, none of the three procedures are equal anymore, since the cross-sectional regressions ignore time-series variation in the x_{it}. As an extreme example, suppose a scalar x_{it} varies over time but not cross-sectionally,

$$y_{it} = \alpha + x_t\beta + \varepsilon_{it};\ i = 1, 2, ...N; t = 1, 2, ...T.$$

The grand OLS regression is

$$\hat{\beta}_{OLS} = \frac{\sum_{it}\tilde{x}_t y_{it}}{\sum_{it}\tilde{x}_t^2} = \frac{\sum_t \tilde{x}_t \frac{1}{N}\sum_i y_{it}}{\sum_t \tilde{x}_t^2}$$

where $\tilde{x} = x - E_T(x)$ denotes the demeaned variables. The estimate is driven by the covariance over time of x_t with the cross-sectional average of the y_{it}, which is sensible because all of the information in the sample lies in time variation. However, you can't even run a cross-sectional estimate, since the right hand variable is constant across i. As a practical example, you might be interested in a CAPM specification in which the betas vary over time (β_t) but not across test assets. This sample still contains information about the CAPM: the time-variation in betas should be matched by time variation in expected returns. But any method based on cross-sectional regressions will completely miss it. ■

Chapter 13. Maximum likelihood

Maximum likelihood is, like GMM, a general organizing principle that is a good place to start when thinking about how to choose parameters and evaluate a model. It comes with a useful asymptotic distribution theory, which, like GMM, is a good place to start when you are unsure about how to treat various problems such as the fact that betas must be estimated in a cross-sectional regression.

As we will see, maximum likelihood is a special case of GMM. It prescribes which moments are statistically most informative. Given those moments ML and GMM are the same. Thus, ML can be used to defend why one picks a certain set of moments, or for advice on which moments to pick if one is unsure. In this sense, maximum likelihood justifies the regression tests above, as it justifies standard regressions. On the other hand, ML does not easily allow you to use other moments, if you suspect that ML's choices are not robust to misspecifications of the economic or statistical model.

13.1 Maximum likelihood

The maximum likelihood principle says to pick the parameters that make the observed data most likely. Maximum likelyhood estimates are asymptotically efficient. The information matrix gives the asymptotic standard errors of ML estimates.

The maximum likelihood principle says to pick that set of parameters that makes the observed data most likely. This is not "the set of parameters that are most likely given the data" – in classical (as opposed to Bayesian) statistics, parameters are numbers not random variables.

To implement this idea, you first have to figure out what the probability of seeing a data set $\{x_t\}$ is, given the free parameters θ of a model. This probability distribution is called the *likelihood function* $f(\{x_t\};\theta)$. Then, the maximum likelihood principle says to pick

$$\hat{\theta} = \arg\max_{\{\theta\}} f(\{x_t\};\theta).$$

For reasons that will soon be obvious, it's much easier to work with the log of this probability distribution

$$\mathcal{L}(\{x_t\};\theta) = \ln f(\{x_t\};\theta),$$

Maximizing the log likelihood is the same things as maximizing the likelihood.

Finding the likelihood function isn't always easy. In a time-series context, the best way to do it is often to first find the log *conditional likelihood function*, the chance of seeing x_{t+1}

given $x_t, x_{t-1}, ...$ and given values for the parameters, $f(x_t|x_{t-1}, x_{t-2}, ...x_0; \theta)$. Since joint probability is the product of conditional probabilities, the log likelihood function is just the sum of the conditional log likelihood functions,

$$\mathcal{L}(\{x_t\}; \theta) = \sum_{t=1}^{T} \ln f(x_t|x_{t-1}, x_{t-2}...x_0; \theta). \tag{143}$$

More concretely, we usually assume normal errors, so the likelihood function is

$$\mathcal{L} = -\frac{T}{2} \ln (2\pi |\Sigma|) - \frac{1}{2} \sum_{t=1}^{T} \boldsymbol{\varepsilon}_t' \Sigma^{-1} \boldsymbol{\varepsilon}_t \tag{144}$$

where $\boldsymbol{\varepsilon}_t$ denotes a vector of shocks; $\boldsymbol{\varepsilon}_t = x_t - E(x_t|x_{t-1}, x_{t-2}...x_0; \theta)$. Then, just invert whatever model you have that produces data $\mathbf{x}_t$ from errors $\boldsymbol{\varepsilon}_t$ to express the likelihood function in terms of data x_t, and maximize.

(There is a small issue about how to start off a model such as (143). Ideally, the first observation should be the unconditional density, i.e.

$$\mathcal{L}(\{x_t\}; \theta) = \ln f(x_1; \theta) + \ln f(x_2|x_1; \theta) + \ln f(x_3|x_2, x_1; \theta)...$$

However, the whole point is that it is usually hard to evaluate the unconditional density or the first terms. Therefore, if as usual the conditional density can be expressed in terms of a finite number k of lags of x_t, one often maximizes the *conditional* likelihood function (conditional on the first k observations), treating the first k observations as fixed rather than random variables.

$$\mathcal{L}(\{x_t\}; \theta) = \ln f(x_{k+1}|x_k, x_{k-1}...x_1; \theta) + \ln f(x_{k+2}|x_k, x_{k-1...}x_2; \theta) + ...$$

Alternatively, one can treat the pre-sample values $\{x_0, x_{-1}, ...x_{-k+1}\}$ as additional parameters over which to maximize the likelihood function.)

Maximum likelihood estimators come with a useful asymptotic (i.e. approximate) distribution theory. First, the distribution of the estimates is

$$\hat{\theta} \sim \mathcal{N} \left(\theta, \left[-\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'} \right]^{-1} \right) \tag{145}$$

If the likelyhood $\mathcal{L}$ has a sharp peak at $\hat{\theta}$, then we know a lot about the parameters, while if the peak is flat, other parameters are just as plausible. The maximum likelihood estimator is *asymptotically efficient* meaning that no other estimator can produce a smaller covariance matrix.

The second derivative in (145) is known as the *information matrix*,

$$\mathcal{I} = -\frac{1}{T}\frac{\partial^2 \mathcal{L}}{\partial\theta\partial\theta'} = -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \ln f(x_{t+1}|x_t, x_{t-1}, ...x_0;\theta)}{\partial\theta\partial\theta'}. \qquad (146)$$

(More precisely, the information matrix is defined as the expected value of the second partial, which is estimated with the sample value.) The information matrix can also be estimated as a product of first derivatives. The expression

$$\mathcal{I} = -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \ln f(x_{t+1}|x_t, x_{t-1}, ...x_0;\theta)}{\partial\theta}\right)\left(\frac{\partial \ln f(x_{t+1}|x_t, x_{t-1}, ...x_0;\theta)}{\partial\theta}\right)'.$$

converges to the same value as (146). (Hamilton 1994 p.429 gives a proof.)

If we estimate a model restricting the parameters, then the maximum value of the likelihood function will necessarily be lower. However, if the restriction is true, it shouldn't be that much lower. This intuition is captured in the *likelihood ratio test*

$$2(\mathcal{L}_{\text{unrestricted}} - \mathcal{L}_{\text{restricted}})\sim\chi^2_{\text{number of restrictions}} \qquad (147)$$

The form and idea of this test is much like the χ^2 difference test for GMM objectives that we met in section xx.

13.2 When factors are returns, ML prescribes a time-series regression.

I add to the economic model $E(\mathbf{R}^e) = \boldsymbol{\beta}E(f)$ a statistical assumption that the regression errors are independent over time and independent of the factors. ML then prescribes a time-series regression with no constant. To prescribe a time series regression with a constant, we drop the model prediction $\alpha = 0$. I show how the information matrix gives the same result as the OLS standard errors.

Given a linear factor model whose factors are also returns, as with the CAPM, ML prescribes a time-series regression test. To keep notation simple, I again treat a single factor f. The economic model is

$$E(\mathbf{R}^e) = \boldsymbol{\beta}E(f) \qquad (148)$$

$\mathbf{R}^e$ is an $N\times 1$ vector of test assets, and $\boldsymbol{\beta}$ is an $N\times 1$ vector of regression coefficients of these assets on the factor (the market return R^{em} in the case of the CAPM).

To apply maximum likelihood, we need to add an explicit statistical model that fully describes the joint distribution of the date. I assume that the market return and regression errors are i.i.d. normal, i.e.

$$\begin{aligned} \mathbf{R}_t^e &= \boldsymbol{\alpha} + \boldsymbol{\beta} f_t + \boldsymbol{\varepsilon}_t \\ f_t &= E(f) + u_t \\ & \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ u_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & 0 \\ 0 & \sigma_u^2 \end{bmatrix} \right) \end{aligned} \tag{149}$$

Equation (149) has no content other than normality. The zero correlation between u_t and $\boldsymbol{\varepsilon}_t$ identifies $\boldsymbol{\beta}$ as a regression coefficient. You can in fact be even more principled and just write $\mathbf{R}^e, R^{em}$ as a general bivariate normal, and a problem asks you to try this approach.

The economic model (148) implies restrictions on this statistical model. Taking expectations of (149), the CAPM implies that the intercepts α should all be zero. Again, this is also the only restriction that the CAPM places on the statistical model (149).

The most principled way to apply maximum likelihood is to impose the null hypothesis throughout. Thus, we write the likelihood function imposing $\boldsymbol{\alpha} = 0$. As above, to construct the likelihood function, we reduce the statistical model to independent error terms, and then add their log probability densities to get the likelihood function.

$$\mathcal{L} = (\text{const}) - \frac{1}{2} \sum_{t=1}^{T} \left(\mathbf{R}_t^e - \boldsymbol{\beta} f_t\right)' \Sigma^{-1} \left(\mathbf{R}_t^e - \boldsymbol{\beta} f_t\right) - \frac{1}{2} \sum_{t=1}^{T} \frac{\left(f_t - E(f)\right)^2}{\sigma_u^2}$$

The estimates follow from the first order conditions,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} &= \Sigma^{-1} \sum_{t=1}^{T} \left(\mathbf{R}_t^e - \boldsymbol{\beta} f_t\right) f_t = 0 \Rightarrow \hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} f_t^2 \right)^{-1} \sum_{t=1}^{T} \mathbf{R}_t^e f_t \\ \frac{\partial \mathcal{L}}{\partial E(f)} &= \frac{1}{\sigma_u^2} \sum_{t=1}^{T} \left(f_t - E(f)\right) = 0 \Rightarrow \widehat{E(f)} = \hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} f_t \end{aligned}$$

($\partial \mathcal{L} / \partial \Sigma$ and $\partial \mathcal{L} / \partial \sigma^2$ also produce ML estimates of the covariance matrices, which turn out to be the standard averages of residuals.)

The ML estimate of $\boldsymbol{\beta}$ is the OLS regression *without* a constant. The null hypothesis says to leave out the constant, and the ML estimator uses that fact to avoid estimating a constant. Since the factor risk premium is equal to the market return, it's not too surprising that the λ estimate is the same as that of the average market return.

The asymptotic standard errors follow from either estimate of the information matrix, for example

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\Sigma^{-1} \sum_{t=1}^{T} f_t^2 = 0$$

Thus,

$$cov(\hat{\boldsymbol{\beta}}) = \frac{1}{T}\frac{1}{E(f^2)}\Sigma = \frac{1}{T}\frac{1}{E(f)^2+\sigma^2(f)}\Sigma. \tag{150}$$

This is the standard OLS formula.

We can also apply maximum likelihood to estimate an unconstrained model, containing intercepts, and then use Wald tests (estimate/standard error) to test the restriction that the intercepts are zero. We also need the unconstrained model to run the likelihood ratio test of the constrained model vs. the unconstrained model. The unconstrainted likelihood function is

$$\mathcal{L} = \text{(const.)} - \frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{R}_t^e - \boldsymbol{\alpha} - \boldsymbol{\beta}f_t\right)' \Sigma^{-1}\left(\mathbf{R}_t^e - \boldsymbol{\alpha} - \boldsymbol{\beta}f_t\right) + ...$$

(I ignore the term in the factor, since it will again just tell us to use the sample mean to estimate the factor risk premium.)

The estimates are now

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\alpha}} &= \Sigma^{-1}\sum_{t=1}^{T}\left(\mathbf{R}_t^e - \boldsymbol{\alpha} - \boldsymbol{\beta}f_t\right) = \mathbf{0} \;\Rightarrow\; \hat{\boldsymbol{\alpha}} = E_T(\mathbf{R}_\mathbf{t}^\mathbf{e}) - \hat{\boldsymbol{\beta}}E_T(f_t) \\
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} &= \Sigma^{-1}\sum_{t=1}^{T}\left(\mathbf{R}_t^e - \boldsymbol{\alpha} - \boldsymbol{\beta}f_t\right)f_t = \mathbf{0} \Rightarrow \hat{\boldsymbol{\beta}} = \frac{cov_T\left(\mathbf{R}_t^e, f_t\right)}{\sigma_T^2\left(f_t\right)}
\end{aligned}$$

Unsurprisingly, the maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the OLS estimates, with a constant.

The inverse of the information matrix gives the asymptotic distribution of these estimates. Since they are just OLS estimates, we're going to get the OLS standard errors, but it's worth seeing it come out of ML.

$$\begin{aligned}
-\left[\frac{\partial^2\mathcal{L}}{\partial\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}\partial\begin{bmatrix}\boldsymbol{\alpha} & \boldsymbol{\beta}\end{bmatrix}}\right]^{-1} &= \begin{bmatrix}\Sigma^{-1} & \Sigma^{-1}E(f)\\ \Sigma^{-1}E(f) & \Sigma^{-1}E(f^2)\end{bmatrix}^{-1} \\
&= \frac{1}{\sigma^2(f)}\begin{bmatrix}E(f^2) & E(f)\\ E(f) & 1\end{bmatrix}\otimes\Sigma
\end{aligned}$$

The covariance matrices of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are thus

$$\begin{aligned} cov(\hat{\boldsymbol{\alpha}}) &= \frac{1}{T}\left[1+\left(\frac{E(f)}{\sigma(f)}\right)^2\right]\Sigma \\ cov(\hat{\boldsymbol{\beta}}) &= \frac{1}{T}\frac{1}{\sigma^2(f)}\Sigma. \end{aligned} \tag{151}$$

These are just the usual OLS standard errors, which we derived above as a special case of GMM standard errors for the OLS time-series regressions when errors are uncorrelated over time and independent of the factors, or by specializing $\sigma^2(X'X)^{-1}$.

You cannot just invert $\partial^2\mathcal{L}/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$ to find the covariance of $\hat{\boldsymbol{\alpha}}$. That would give just Σ as the covariance matrix of $\hat{\boldsymbol{\alpha}}$, which would be wrong. You have to invert the entire information matrix to get the standard error of any parameter. Otherwise, you are ignoring the effect that estimating $\boldsymbol{\beta}$ has on the distribution of $\hat{\boldsymbol{\alpha}}$. In fact, what I presented is really wrong, since we also must estimate Σ. However, it turns out that $\hat{\Sigma}$ is independent of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, so the top left two elements of the true information matrix is the same as I have written here.

The variance of $\hat{\boldsymbol{\beta}}$ in (151) is larger than it is in (150) was when we impose the null of no constant. ML uses all the information it can to produce efficient estimates – estimates with the smallest possible covariance matrix. The ratio of the two formulas is equal to $1 + E(f)^2/\sigma^2(f)$, which we studied above in section xx. In annual data for the CAPM, $\sigma(R^{em}) = 16\%$, $E(R^{em}) = 8\%$, means that unrestricted estimate (151) has a variance 25% larger than the restricted estimate (150), so the gain in efficiency can be important. In monthly data, however the gain is smaller since variance and mean both scale with the horizon. This is also a warning: ML can prescribe silly procedures (running a regression without a constant) in order to get any small improvement in efficiency.

We can use these covariance matrices to construct a Wald (estimate/standard error) test the restriction of the model that the alphas are all zero,

$$T\left(1+\left(\frac{E(R^{em})}{\sigma(R^{em})}\right)^2\right)^{-1}\hat{\boldsymbol{\alpha}}'\Sigma^{-1}\hat{\boldsymbol{\alpha}} \sim \chi^2_N. \tag{152}$$

Again, we already derived this χ^2 test in (122), and its finite sample F counterpart, the GRS F test (123). The other test of the restrictions is the likelihood ratio test (147). Quite generally, likelihood ratio tests are asymptotically equivalent to Wald tests, and so gives the same result. Showing it in this case is not worth the algebra.

13.3 When factors are not excess returns, ML prescribes a cross-sectional regression

If the factors are not returns, we didn't have a choice between time-series and cross-sectional regression, since the intercepts are not zero. As you might suspect, ML prescribes a cross-sectional regression in this case.

The factor model, expressed in expected return beta form, is

$$E(R^{ei}) = \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{\lambda};\ \ i = 1, 2, ..N \tag{153}$$

The betas are defined from time-series regressions

$$R_t^{ei} = a_i + \boldsymbol{\beta}_i'\mathbf{f}_t + \varepsilon_t^i \tag{154}$$

The intercepts a_i in the time-series regressions need not be zero, since the model does not apply to the factors. They are not unrestricted, however. Taking expectations of the time-series regression (154) and comparing it to (153) (as we did to derive the restriction $\alpha = 0$ for the time-series regression), the restriction $\alpha = 0$ implies

$$a_i = \boldsymbol{\beta}_i'\left(\boldsymbol{\lambda} - E(\mathbf{f}_t)\right) \tag{155}$$

Plugging into (154), we can say that the time series regressions must be of the restricted form

$$R_t^{ei} = \boldsymbol{\beta}_i'\boldsymbol{\lambda} + \boldsymbol{\beta}_i'\left[\mathbf{f}_t - E(\mathbf{f}_t)\right] + \varepsilon_t^i. \tag{156}$$

In this form, you can see that $\boldsymbol{\beta}_i\boldsymbol{\lambda}$ determines the mean return. Since there are fewer factors than returns, this is a restriction on the regression (156).

Stack assets $i = 1, 2, ...N$ to a vector; and introduce the auxiliary statistical model that the errors and factors are i.i.d. normal and uncorrelated with each other. Then, the restricted model is

$$\begin{aligned}
\mathbf{R}_t^e &= \mathbf{B}\boldsymbol{\lambda} + \mathbf{B}\left[\mathbf{f}_t - E(\mathbf{f}_t)\right] + \boldsymbol{\varepsilon}_t;\ \ \boldsymbol{\varepsilon}_t \sim \mathcal{N}\left(0, \Sigma\right) \\
\mathbf{f}_t &= E(\mathbf{f}) + \mathbf{u}_t;\ \mathbf{u}_t \sim \mathcal{N}\left(0, V\right) \\
\begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{u}_t \end{bmatrix} &\sim \mathcal{N}\left(0, \begin{matrix} \Sigma & 0 \\ 0 & V \end{matrix}\right)
\end{aligned}$$

where $\mathbf{B}$ denotes a $N \times K$ matrix of regression coefficients of the N assets on the K factors. The likelihood function is

$$\begin{aligned}
\mathcal{L} &= \text{(const.)} - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\Sigma^{-1}\boldsymbol{\varepsilon}_t - \frac{1}{2}\sum_{t=1}^{T}\mathbf{u}_t'V^{-1}\mathbf{u}_t \\
\boldsymbol{\varepsilon}_t &= \mathbf{R}_t^e - \mathbf{B}\left[\boldsymbol{\lambda} + \mathbf{f}_t - E(\mathbf{f})\right];\ \ \mathbf{u}_t = \mathbf{f}_t - E(\mathbf{f})
\end{aligned}$$

Maximizing the likelihood function,

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial E(\mathbf{f})} &: \ 0 = \sum_{t=1}^{T}\mathbf{B}'\Sigma^{-1}\left(\mathbf{R}_t^e - \mathbf{B}\left[\boldsymbol{\lambda} + \mathbf{f}_t - E(\mathbf{f})\right]\right) + \sum_{t=1}^{T}V^{-1}(\mathbf{f}_t - E(\mathbf{f})) \\
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda}} &: \ 0 = \mathbf{B}'\sum_{t=1}^{T}\Sigma^{-1}\left(\mathbf{R}_t^e - \mathbf{B}\left[\boldsymbol{\lambda} + \mathbf{f}_t - E(\mathbf{f})\right]\right)
\end{aligned}$$

The solution to this pair of equations is

$$\widehat{E(\mathbf{f})} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t \tag{157}$$

$$\hat{\boldsymbol{\lambda}} = \left(\mathbf{B}'\Sigma^{-1}\mathbf{B}\right)^{-1}\mathbf{B}'\Sigma^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{R}^e_t. \tag{158}$$

The maximum likelihood estimate of the factor risk premium is a GLS cross-sectional regression of average returns on betas.

As with the CAPM, the maximum likelihood estimates of the regression coefficients $\mathbf{B}$ are slightly altered from the unrestricted OLS values:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{B}} : \sum_{t=1}^{T}\Sigma^{-1}\left(\mathbf{R}^e_t - \mathbf{B}\left[\boldsymbol{\lambda}+\mathbf{f}_t - E(\mathbf{f})\right]\right)\left[\boldsymbol{\lambda}+\mathbf{f}_t - E(\mathbf{f})\right]' = 0 \tag{159}$$

$$\hat{\mathbf{B}} = \sum_{t=1}^{T}\mathbf{R}^e_t\left[\mathbf{f}_t+\boldsymbol{\lambda}-E(\mathbf{f})\right]'\left(\sum_{t=1}^{T}\left[\mathbf{f}_t+\boldsymbol{\lambda}-E(\mathbf{f})\right]\left[\mathbf{f}_t+\boldsymbol{\lambda}-E(\mathbf{f})\right]'\right)^{-1}$$

This is true, even though the $\mathbf{B}$ are defined in the theory as population regression coefficients. The restricted ML uses the restrictions to improve on OLS estimates in a sample. (The matrix notation hides a lot here! If you want to rederive these formulas, it's helpful to start with scalar parameters, e.g. $\mathbf{B}_{ij}$, and to think of it as $\partial\mathcal{L}/\partial\theta = \sum_{t=1}^{T}\left(\partial\mathcal{L}/\partial\varepsilon_t\right)'\partial\varepsilon_t/\partial\theta.$) Therefore, to really implement ML, you have to solve (158) and (159) simultaneously for $\hat{\boldsymbol{\lambda}}$, $\hat{\mathbf{B}}$, along with $\hat{\Sigma}$ whose ML estimate is the usual second moment matrix of the residuals. This can usually be done iteratively: Start with OLS $\hat{\mathbf{B}}$, run an OLS cross-sectional regression for $\hat{\boldsymbol{\lambda}}$, form $\hat{\Sigma}$, and iterate.

13.4 Time series vs. cross-section

I track down why ML prescribes a time-series regression when factors are returns and a cross-sectional regression when factors are not returns. I argue that the cross-sectional regression may be more robust to model misspecification. I show that the time-series / cross-sectional regression issue is the same as the OLS / GLS cross-sectional regression issue, and similar to the issue whether one runs time-series regressions with no intercept, both cases in which one may trade efficiency for robustness.

The issue

When factors are not returns, ML prescribed a cross-sectional regression When the fac-

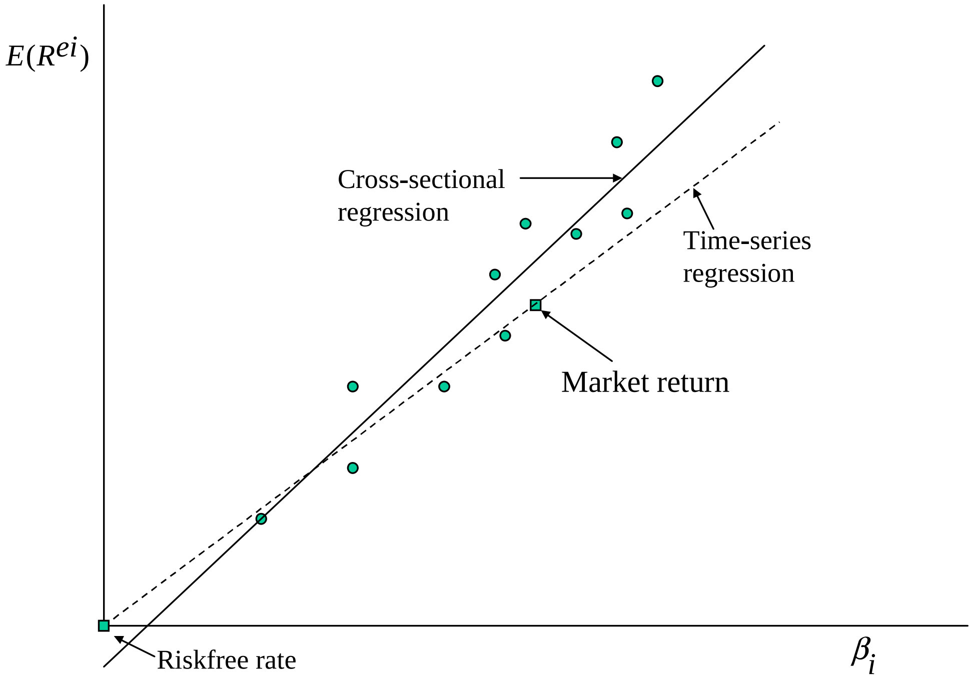


Figure 21. Time-series vs. cross-sectional regression estimates.

tors are returns, ML prescribes a time-series regression. Figure 21 illustrates the difference between the two approaches. The time-series regression estimates the factor risk premium from the average of the factors alone, ignoring any information in the other assets. For example in the CAPM, $\hat{\lambda} = E_T(R^{em})$. Thus, a time-series regression draws the expected return-beta line across assets by making it fit precisely on two points, the market return and the riskfree rate, and the market and riskfree rate have zero estimated pricing error in every sample. The cross-sectional regression draws the expected return-beta line by minimizing the squared pricing error across all assets. It allows some pricing error for the market return and (if the intercept is free) the riskfree rate, if by doing so the pricing errors on other assets can be reduced.

Of course, if the model is correct, the two approaches should converge as the sample gets larger. However, the difference between cross-sectional and time-series approaches can often be large and economically important in samples and models typical of current empirical work. The figure shows a stylized version of CAPM estimates on the size portfolios. High beta assets (the small firm portfolios) have higher average returns than predicted by the CAPM. However, the estimated pricing error for these assets is much smaller if one allows a cross-sectional regression to determine the market price of risk. Classical tests of the CAPM based on beta-sorted portfolios often turn out the other way: the cross-sectional regression market line is flatter than the time-series regression, with an intercept that is higher than the sample riskfree rate. As another example, Fama and French (19xx) report important correlations between betas and pricing errors in a time-series test of a three-factor model on industry portfolios. This correlation cannot happen with an OLS cross-sectional estimate, as the cross-sectional estimate sets the correlation between right hand variables (betas) and error terms (pricing errors) to zero by construction. Thus, such a correlation is an indication that a cross-sectional regression would give quite different results.

(The difference is not always large of course, and there is one special case in which time-series and cross-section agree by construction. If one is testing CAPM and the market return is an equally weighted portfolio of the test assets, then an OLS cross-sectional regression with an estimated intercept passes through the market return, since the average pricing error is set to zero by the cross-sectional regression. In this case, though, time series regression imposes a zero intercept and cross-sectional regression can leave the intercept free.)

When there is a choice – when one is testing a linear factor model in which the factors are also portfolio returns – should one use a time-series or a cross-sectional regression? Since the final evaluation of any model depends on the size of pricing errors, it would seem to makes sense to estimate the model by choosing free parameters to make the pricing errors as small as possible. That is exactly what the cross-sectional regression accomplishes. However, the time-series regression is the maximum likelihood estimator, and thus asymptotically efficient. This seems like a strong argument for the pure time-series approach.

Why ML does what it does

To resolve this issue, we have to understand why ML prescribes such different procedures when factors are and aren't returns. Why does ML ignore all the information in the test asset average returns, and estimate the factor premium from the average factor return only? The answer lies in the structure that we told ML to assume when looking at the data. First, when we write $\mathbf{R}^e = \mathbf{a} + \boldsymbol{\beta} f_t + \varepsilon_t$ and ε independent of f, we tell ML that a sample of returns *already includes the same sample* of the factor, plus extra noise. Thus, the sample of test asset returns cannot possibly tell ML anything more than the sample of the factor alone about the mean of the factor. Second, we tell ML that the factor risk premium equals the mean of the factor, so it may not consider the possibility that the two are different in trying to match the data. When the factor is not also a return, ML still ignores anything but the factor data in estimating the mean of the factor, but now it is allowed to change a different parameter to fit the returns, which it does by cross-sectional regression.

The time-series vs. cross-section issue is essentially the same as the OLS vs. GLS issue. ML prescribes a GLS cross-sectional regression,

$$\hat{\lambda} = \left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\Sigma^{-1}E_T(\mathbf{R}^e).$$

The assets are weighted by inverse of the *residual* covariance matrix Σ, not the return covariance matrix. If we include as a test asset a portfolio that is nearly equal to the factor but with a very small variance, then the elements of Σ in that row and column will be very small and Σ^{-1} will overwhelmingly weight that asset in determining the risk premium. If the limit that we actually include the factor as a test asset, $\left(\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\Sigma^{-1}$ becomes a vector of zeros with a unit element in the position of the factor and we return to $\hat{\lambda} = E_T(f)$.

Model misspecification

As I argued in section 2 and in section 2, it may well be a good idea to use OLS cross-sectional regressions or first-stage GMM rather than the more efficient GLS, because the OLS regression can be more robust to small misspecifications of the economic or statistical model, and OLS may be better behaved in small samples in which Σ^{-1} is hard to estimate.

Similarly, time-series regressions are almost universally run with a constant, though ML prescribes a regression with no constant. The reason must be that researchers feel that omitting some of the information in the null hypothesis, the estimation and test is more robust, though some efficiency is lost if the null economic and statistical models are exactly correct. Since ML prescribes a cross-sectional regression if we drop the restriction $\lambda = E(f)$, running a cross-sectional regression may also be a way to gain robustness at the expense of one degree of freedom.

Here is an example of a common small misspecification that justifies a cross-sectional rather than a time-series approach. Suppose that the test portfolios do follow a single factor model precisely, with an excess return f_t as the single factor. However, we have an imprecise proxy for the true factor,

$$f_t^p = f_t + \delta_t. \tag{160}$$

Most obviously, the market return is an imperfect proxy for the wealth portfolio that the CAPM specifies as the single factor. Multifactor models also use return factors that are imperfect proxies for underlying state variables. For example Fama and French (1993) use portfolios of stocks sorted by book/market value as a return factor, with the explicit idea that it is a proxy for a more rigorously-derivable state variable.

The true model is

$$\mathbf{R}^e = \boldsymbol{\beta} f_t + \boldsymbol{\varepsilon}_t. \tag{161}$$

There is no intercept since f is a return. For the statistical part of the model, I again assume that $\boldsymbol{\varepsilon}_t$ and f_t are jointly normal i.i.d. and independent of each other.

If we had data on f_t, the ML estimate of this model would be, like the CAPM, a pure time series regression. We have to work out what a model using the proxy f^p as reference portfolio looks like. This is an example, and to keep it simple I assume that δ_t is mean zero, and uncorrelated with f_t and $\boldsymbol{\varepsilon}_t$.

$$E(\delta) = 0;\ E(\delta f) = E(\delta\varepsilon) = 0.$$

(This is a poor assumption for the CAPM. Since the market portfolio is a linear combination of the test assets, the error in R^m is the sum of the errors ε and thus unlikely to be uncorrelated with them. This is a more plausible assumption for non-market factors in multifactor models. Similar examples with $E(f\varepsilon) \neq 0$ generate the same sorts of misspecification, but also introduce pricing errors in the test assets.)

Now, if we use the proxy f^p rather than f as the factor, the *test assets* still follow the factor model exactly, but the factor portfolio does not, and the risk premium is no longer equal to the mean of the factor portfolio:

$$\begin{aligned} E(\mathbf{R}^e) &= \boldsymbol{\beta}_p \lambda_p \\ E(f^p) &= \lambda_p - \alpha_p \end{aligned} \tag{162}$$

and $\boldsymbol{\beta}_p$ is the regression coefficient of $\mathbf{R}^e$ on the proxy f^p. Therefore, if you spell out the misspecification, the ML estimate of the factor model is now a cross-sectional regression, not a time-series regression! A similar misspecification occurs when we suspect that the riskfree rate is "too low" and again leads to cross-sectional estimates.

The algebra behind (162) is straightforward,

$$\begin{aligned} cov(\mathbf{R}^e, f^p) &= cov(\mathbf{R}^e, f+\delta) = cov(R^e, f) \\ E(R^e) &= \frac{cov(R^e, f)}{\sigma^2(f)}\lambda = \frac{cov(R^e, f^p)}{\sigma^2(f^p)}\frac{\sigma^2(f^p)}{\sigma^2(f)}\lambda = \beta_p\lambda_p \\ E(f^p) &= E(f) = \lambda = \lambda_p\frac{\sigma^2(f)}{\sigma^2(f^p)} = \lambda_p - \lambda\left(\frac{\sigma^2(\delta)}{\sigma^2(f)}\right) = \lambda_p - \alpha_p \end{aligned}$$

where I have introduced the notation

$$\begin{aligned} \lambda_p &\equiv \left(1 + \frac{\sigma^2\left(\delta\right)}{\sigma^2(f)}\right)\lambda \\ \alpha_p &\equiv \left(\frac{\sigma^2(\delta)}{\sigma^2(f)}\right)\lambda. \end{aligned}$$

Chapter 14. ML, GMM and Regression

As you probably have already noticed, GMM, regression and ML approaches to asset pricing models all look very similar. In each case one estimates a set of parameters, such as the $a, \mathbf{b}$ in $m_t = a - \mathbf{b}\mathbf{f}_t$, the γ, δ in $m_t = \delta\left(c_t/c_{t-1}\right)^{-\gamma}$ or the β_i in $R^{ei} = a + \beta_i f_t + \varepsilon_{it}$. Then, we calculate pricing errors and evaluate the model by a quadratic form in the pricing errors. Here I draw some additional connections and highlight the distinctions between the three approaches. I start with two facts that help to anchor the discussion: 1) ML is a special case of GMM, 2) one can approach either $p = E(mx)$ or expected return/beta expressions of asset pricing models with either ML or GMM approaches to estimation. .

14.1 ML is GMM on the scores

We can regard ML as a special case of GMM. ML uses the information in the auxiliary statistical model to derive statistically most informative moment conditions, moment conditions that fully exhaust the model's implications. To see this fact, start with the first order conditions for maximizing a likelihood function

$$\frac{\partial \mathcal{L}(\{x_t\};\theta)}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \ln f(x_t|x_{t-1}x_{t-2}...;\theta)}{\partial \theta} = 0. \tag{163}$$

This is a GMM estimate. It is the sample counterpart to a population moment condition

$$g(\theta) = E\left(\frac{\partial \ln f(x_t|x_{t-1}x_{t-2}...;\theta)}{\partial \theta}\right) = 0; \tag{164}$$

The term $\partial \ln f(x_t|x_{t-1}x_{t-2}...;\theta)/\partial\theta$ is known as the "score". It is a random variable, formed as a combination of current and past data $(x_t, x_{t-1}...)$. Thus, maximum likelihood is a special case of GMM, a special choice of which moments to examine (163).

For example, suppose that x follows an AR(1) with known variance,

$$x_t = \rho x_{t-1} + \varepsilon_t.$$

Then,

$$\ln f(x_t|x_{t-1}, x_{t-2}...;\rho) = \text{const.} - \frac{\varepsilon_t^2}{2\sigma^2} = \text{const} - \frac{\left(x_t - \rho x_{t-1}\right)^2}{2\sigma^2}$$

and the score is

$$\frac{\partial \ln f(x_t|x_{t-1}x_{t-2}...;\rho)}{\partial \rho} = \frac{\left(x_t - \rho x_{t-1}\right)x_{t-1}}{\sigma^2}.$$

The first order condition for maximizing likelihood, (163), is

$$\frac{1}{T}\sum_{t=1}^{T}\left(x_t-\rho x_{t-1}\right)x_{t-1}=0.$$

This expression as a moment condition, and you'll recognize it as the OLS estimator of ρ, which we have already regarded as a case of GMM.

The example shows another property of scores: *The scores should be unforecastable.* In the example,

$$E_{t-1}\left[\frac{\left(x_t-\rho x_{t-1}\right)x_{t-1}}{\sigma^2}\right]=E_{t-1}\left[\frac{\varepsilon_t x_{t-1}}{\sigma^2}\right]=0. \tag{165}$$

Intuitively, if we used a combination of the x variables $E(g(x_t,x_{t-1},...))=0$ that was predictable, we could form another moment that described the predictability of the g variable and use that moment to get more information about the parameters. To prove this property more generally, start with the fact that $f(x_t|x_{t-1},x_{t-2},...;\theta)$ is a conditional density and therefore must integrate to one,

$$\begin{aligned}
1 &= \int f(x_t|x_{t-1},x_{t-2},...;\theta)dx_t\\
0 &= \int \frac{\partial f(x_t|x_{t-1},x_{t-2},...;\theta)}{\partial\theta}dx_t\\
0 &= \int \frac{\partial \ln f(x_t|x_{t-1},x_{t-2},...;\theta)}{\partial\theta}f(x_t|x_{t-1},x_{t-2},...;\theta)dx_t\\
0 &= E_{t-1}\left[\frac{\partial \ln f(x_t|x_{t-1},x_{t-2},...;\theta)}{\partial\theta}\right]
\end{aligned}$$

Furthermore, as you might expect, *the GMM distribution theory formulas give the same result as the ML distribution,* i.e., the information matrix is the asymptotic variance-covariance matrix. To show this fact, apply the GMM distribution theory (100) to (163). The derivative matrix is

$$d=\frac{\partial g_T(\theta)}{\partial\theta'}=\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2\ln f(x_t|x_{t-1}x_{t-2}...;\theta)}{\partial\theta\partial\theta'}=\mathcal{I}$$

This is the second derivative expression of the information matrix. The S matrix is

$$E\left[\frac{\partial\ln f(x_t|x_{t-1}x_{t-2}...;\theta)}{\partial\theta}\,\frac{\partial\ln f(x_t|x_{t-1}x_{t-2}...;\theta)}{\partial\theta}'\right]=\mathcal{I}$$

The lead and lag terms in S are all zero since we showed above that scores should be unforecastable. This is the outer product definition of the information matrix. There is no a

matrix, since the moments themselves are set to zero. The GMM asymptotic distribution of $\hat{\theta}$ is therefore

$$\sqrt{T}(\hat{\theta}-\theta) \rightarrow \mathcal{N}\left[0,\ d^{-1}Sd^{-1\prime}\right] = \mathcal{N}\left[0,\ \mathcal{I}^{-1}\right].$$

We recover the inverse information matrix, as specified by the ML asymptotic distribution theory.

14.2 ML approach to a consumption-based model

There is nothing that forces us to pair GMM with $p = E(mx)$ type models or ML with regression tests. We have already used the GMM *principle* to construct tests of expected return - beta models. We can also use the ML *principle* to construct estimates and tests of $p = E(mx)$ type models, and many authors do so. For example, we could start with the same statistical assumption that R_t and f_t are jointly normally distributed and i.i.d. over time. $1 = E(mR)$; $m = a + bf$, and, if the factors are returns, $1 = E(mf)$, imply restrictions across the mean and covariance matrix of R^e_t and f_t. We can then write the likelyhood function, and maximize it to find estimates of a, b.

To investigate a less trivial example, here is how we might handle an explicit consumption-based model, taken from Hansen and Singleton (198x). Start with the simplest model with power utility. Using a set of returns R_t the model predicts

$$E_t\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{t+1}\right) = 1$$

To apply maximum likelihood, we need auxiliary statistical assumptions, just as we added to the regression model of the CAPM the assumption that the errors were i.i.d. normal. A natural starting place is to model log consumption growth and log returns as jointly normal and i.i.d. .Then the pricing equation becomes

$$E_t\left(e^{-\delta}e^{-\gamma\Delta c_{t+1}}e^{\mathbf{r}_{t+1}}\right) = \mathbf{1}$$

and, taking logs,

$$\delta - \gamma E\Delta c + E\mathbf{r} + \frac{1}{2}\gamma^2\sigma^2_{\Delta c} + \frac{1}{2}\sigma^2(\mathbf{r}) - \gamma cov(\Delta c, \mathbf{r}) = 0 \tag{166}$$

I use small letters for logs of capital letters, $r = \ln R$, $c = \ln C$, etc., and in the last equality I suppress t subscripts since returns are i.i.d. It will simplify matters to focus on return differences (not really excess returns since we took logs); thus choose one asset r^{N+1} and denote $\mathbf{r}^e = \mathbf{r} - r^{N+1}$. Then we can difference (166) to give

$$E\mathbf{r}^e + \frac{1}{2}\sigma^2(\mathbf{r}^e) - \gamma cov(\Delta c, \mathbf{r}^e) = 0$$

In sum, we have assumed joint normal log returns and consumption growth,

$$\left[\begin{array}{c}\Delta c_{t+1}\\ \mathbf{r}^e_{t+1}\end{array}\right]\sim\mathcal{N}\left(\left[\begin{array}{c}E\Delta c\\ E\mathbf{r}^e\end{array}\right], V=\left[\begin{array}{cc}\sigma^2_{\Delta c} & cov(\Delta c,\mathbf{r}^e)\\ cov(\Delta c,\mathbf{r}^e) & \Sigma\end{array}\right]\right) \tag{167}$$

and in this context, the economic model restricts this statistical description by linking some of the free parameters,

$$E\mathbf{r}^e=\gamma cov(\Delta c,\mathbf{r}^e)-\frac{1}{2}\text{diag}\Sigma \tag{168}$$

The standard ML method then is to estimate the restricted model, calculate Wald statistics for the parameters, and test the restriction by comparing the likelihood of the restricted model to that of an unrestricted model, one that freely estimates the mean and covariance matrix in (167)

We can substitute the restriction in the likelyhood function to eliminate the parameter $E(\mathbf{r}^e)$ and to write the restricted likelihood function as

$$\begin{aligned}\mathcal{L} &= -\frac{T}{2}\ln\left(2\pi\left|V\right|\right)-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t' V^{-1}\boldsymbol{\varepsilon}_t\\ \boldsymbol{\varepsilon}_t &= \left[\begin{array}{c}\Delta c_t-E\Delta c\\ \mathbf{r}^e_t-\gamma cov\left(\Delta c,\mathbf{r}^e\right)+\frac{1}{2}\text{diag}\Sigma\end{array}\right]\end{aligned}$$

The arguments of the likelihood function are $\mathcal{L}(\{\Delta c_t,\mathbf{r}^e_t\};\gamma,E\Delta c,cov\left(\Delta c,\mathbf{r}^e\right),\Sigma)$, including both statistical and economic parameters.

When we maximize the likelihood function you fairly easily see that $E\Delta c=\frac{1}{T}\Sigma_{t=1}^{T}\Delta c_t$ is the maximum likelihood estimate. Taking the derivative with respect to γ,

$$\frac{\partial\mathcal{L}}{\partial\gamma}=\left[\begin{array}{cc}0 & cov(\Delta c,\mathbf{r}^{e\prime})\end{array}\right]V^{-1}\sum_{t=1}^{T}\varepsilon_t=0$$

and hence

$$cov(\Delta c,\mathbf{r}^{e\prime})\Sigma^{-1}\,\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{r}^e_t-\gamma cov\left(\Delta c,\mathbf{r}^e\right)+\frac{1}{2}\text{diag}\Sigma\right)=0. \tag{169}$$

(to derive this last equation you have to use the partitioned matrix inverse formula on V and then recognize that using the sample mean for $E\Delta c$ means the first row is automatically satisfied.)

As we saw above, ML is equivalent to GMM with a specific choice of moments. In this case, the moments prescribed by ML are just a specific linear combination of the pricing errors. The leading terms in (169) are a weighting matrix in GMM language. $cov(\Delta c\,\mathbf{r}^{e\prime})\Sigma^{-1}$ is a $1\times N$ matrix that tells you which linear combination of pricing error moments to set to zero in order to estimate γ. In fact, if you had followed GMM, you might have started with the terms following $\frac{1}{T}$ in the last equation and treated those as your moment conditions for estimating γ. (169) is precisely the "optimal" second-stage GMM estimate in this case.

We can solve equation (169) for the estimate of γ, and it is

$$\hat{\gamma} = \left(cov(\Delta c, \mathbf{r}^{e\prime})\Sigma^{-1} cov\left(\Delta c, \mathbf{r}^{e}\right)\right)^{-1}\ cov(\Delta c, \mathbf{r}^{e\prime})\Sigma^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t^e + \frac{1}{2}\text{diag}\Sigma\right)$$

Once again, this estimate is a cross-sectional GLS regression of average returns, with a lognormal variance correction term, on covariances.

True ML is not quite so simple. The $cov\left(\Delta c, \mathbf{r}^{e}\right)$ and Σ are also parameters that must be estimated. Since they enter the model restrictions, (168), their ML estimates will not be the usual unconstrained estimates. (Similarly, the ML estimate of a time series regression is not the usual estimate, since we force the constant to zero.)

14.3 ML vs. GMM

I have emphasized the many similarities between ML GMM and regressions. In the classic environments we have examined, all methods basically pick parameters to minimize the pricing errors, and test the model's overall fit by whether the minimized pricing errors are larger than sampling variation would suggest.

However, there are differences, and it is time to stop and think about which technique to use. Furthermore, though ML, GMM and regression are quite similar in the classic case of a factor model, excess return, and i.i.d. normal returns and factors, they can suggest quite different procedures in other situations including more sophisticated consumption-based models, explicit term structure models, or option pricing models that require thoughtful treatment of changing volatility and non-normality.

As we have seen, ML is a special case of GMM: it gives a particular choice of moments that are statistically optimal in a well-defined sense. GMM can be used to derive an asymptotic distribution theory for ML as well as lots of other estimation procedures. Thus, the issue is really about when it is important to use ML estimates or whether it is a good idea to use other moments. As with OLS vs. GLS, sub-optimal estimation methods (OLS) can be more robust to model misspecifications. On the other hand, if the statistical model is tractable, right, and if one is unsure about which moments are informative, ML can be an important guide.

14.3.1 Specification errors

ML is often ignored

As we have seen, ML plus the assumption of normal i.i.d. disturbances leads to easily interpretable time-series or cross-sectional regressions. However, asset returns are *not* normally distributed or i.i.d.. They have fatter tails than a normal, they are heteroskedastic (times of high and times of low volatility), they are autocorrelated, and predictable from a variety of variables. If one were to take the ML philosophy seriously, one should model these features of returns. The result would be a different likelihood function, and its scores would prescribe different moment conditions than the familiar and intuitive time-series or cross-sectional regressions.

Interestingly, most empirical workers practically never do this. (The exceptions tend to be papers whose primary point is illustration of econometric technique rather than substantive issues.) ML seems to be fine when it suggests easily interpretable regressions; when it suggests something else, people use the regressions anyway. For example, as we have seen, a ML estimation of the CAPM prescribes that one estimate βs using time-series regressions *without* a constant, exploiting that prediction of the theory. Yet βs are almost universally estimated with a constant. Despite ML's specification of a GLS cross-sectional regression, most empirical work uses OLS cross-sectional regressions. And of course, the above "ML" estimates and test statistics continue to be used, despite the technical feasibility of addressing non-normal and non-i.i.d. returns.

This fact tells us something interesting about the nature of empirical work: researchers don't really believe that their null hypotheses, statistical and economic, are exactly correct. They want to produce estimates and tests that are *robust* to reasonable model misspecificiations. They also want to produce estimates and tests that are easily interpretable, that capture intuitively clear stylized facts in the data. Such estimates are persuasive in large part because the reader can see that they are robust.

ML does not necessarily produce robust or easily interpretable estimates. It wasn't designed to. The point and advertisement of ML is that it provides *efficient* estimates; it uses every scrap of information in the statistical and economic model in the quest for efficiency. It does the "right" efficient thing if model is true. It does not necessarily do the "reasonable" thing for "approximate" models.

Examples

For example, we have seen that ML specifies a time-series regression when the factor is a return, but a cross-sectional regression when the factor is not a return. The time-series regression gains one degree of freedom, but we have also seen that an even slight proxy error in the factor leads to the more intuitive cross-sectional regression. We have also discussed reasons why researchers use OLS cross-sectional regressions rather than more "efficient" GLS. GLS requires modeling and inverting an $N \times N$ covariance matrix, and then focuses attention on portfolios with strong positive and negative weights that seem to have lowest variance in a sample. But such portfolios may be quite sensitive to small transactions costs, and the sampling error in large covariance matrices may ruin the asymptotic advantages of GLS in a finite sample. Similarly, if one asked a researcher why he included a constant in estimating a beta while applying the CAPM, he might well respond that he doesn't believe the theory that much.

In estimating time-series models such as the AR(1) example above, maximum likelyhood minimizes one-step ahead forecast error variance, $\sum \varepsilon_t^2$. But any time-series model is only an approximation, and the researcher's objective may not be one-step ahead forecasting. For example, one may be interested in the long-run behavior of a slow-moving series such as the short rate of interest. The approximate model that generates the smallest one-step ahead forecast error variance may be quite different from the model that best matches long-run autocorrelations, so ML will pick the wrong model and make very bad predictions for long-run responses. (Cochrane 1986 contains a more detailed analysis of this point.)

Models of the term structure of interest rates and real business cycle models in macroeconomics give even more stark examples. These models are *stochastically singular*. They generate predictions for many time series from a few shocks, so the models predict that there are combinations of the time series that leave no error term. Even though the models have rich and interesting implications, ML will seize on this economically uninteresting singularity to reject any model of this form.

The simplest example of the situation is the linear-quadratic permanent income model paired with an AR(1) specification for income. The model is

$$\begin{aligned} y_t &= \rho y_{t-1} + \varepsilon_t \\ c_t - c_{t-1} &= (E_t - E_{t-1}) \frac{1}{1-\beta} \sum_{j=0}^{\infty} \beta^j y_{t+j} = \frac{1}{(1-\beta\rho)(1-\beta)} \varepsilon_t \end{aligned}$$

This model generates all sorts of important and economically interesting predictions for the joint process of consumption and income. Consumption should be roughly a random walk, and should respond only to permanent income changes; investment should be more volatile than income and income more volatile than consumption. Since there is only one shock and two series, however, the model taken literally predicts a deterministic relation between consumption and income.

$$c_t - c_{t-1} = \frac{r\beta}{1-\beta\rho} (y_t - \rho y_{t-1}).$$

ML will notice that this is the *statistically* most informative prediction of the model. In any real data set there is *no* configuration of the parameters r, β, ρ that make this restriction hold, data point for data point. The probability of observing a data set $\{c_t, y_t\}$ is exactly zero, and the log likelyhood function is $-\infty$ for any set of parameters. ML says to throw the model out.

The popular affine yield models of the term structure of interest rates act the same way. They specify that all yields at any moment in time are deterministic functions of a few state variables. Such models capture much of the important qualitative behavior of the term structure, including rising, falling and humped shapes, the time-evolution of those shapes (i.e. that a rising yield curve forecasts changes in future yields and bond holding period returns), and they are very useful for derivative pricing. But it is never the case in actual yield data that yields of all maturities are *exact* functions of three yields. Actual data on N yields always require N shocks, even if the last $N - 3$ have very small variances. Again, a ML approach reports a $-\infty$ log likelyhood function for any set of parameters.

Addressing model mis-specification.

The ML philosophy offers an answer to the model mis-specification question: specify the *right* model, and then do ML. If regression errors are correlated, model and estimate the covariance matrix and do GLS. If one is worried about proxy errors in the pricing factor, short sales costs or other transactions costs in the test assets, time-aggregation or mismeasurement of consumption data, or small but nonzero violations of the model simplifications such as time-varying betas and factor risk premia; additional pricing factors and so on, write them down, and then do ML.

For example, researchers have added "measurement errors" to real business cycle models and affine yield models in order to break the predictions of stochastic singularity. The trouble is, of course, that the assumed structure of the measurement errors now drives what moments ML pays attention to. Also, modeling and estimating stochastic structure of measurement errors takes us further away from the economically interesting parts of the model.

More generally, as we have seen, authors tend not to follow this advice, for the simple reason that it is infeasible. Economics in general and financial economics in particular necessarily studies quantitative parables rather than completely specified models. It would be nice if we could write down completely specified models, if we could quantitatively describe all the possible economic and statistical model and specification errors, but we can't.

The GMM framework, used judiciously, offers an alternative way to address model misspecification. Where ML only gives us a choice of OLS, whose standard errors are wrong, or GLS, GMM allows us to keep an OLS estimate, but correct the standard errors (at least asymptotically) for any statistical problems. More generally, GMM allows one to specify an economically interesting set of moments, or a set of moments that one feels will be robust to misspecifications of the economic or statistical model, *without* having to spell out exactly what is the source of model mis-specification that makes those moments "optimal". It allows one to accept the lower "efficiency" of the estimates if the null really is exactly true, in return for such robustness.

At the same time, it allows one to flexibly incorporate statistical model misspecifications in the distribution theory. For example, knowing that returns are not i.i.d. normal, one may want to use the time series regression technique to estimate betas anyway. This estimate is not inconsistent, but the *standard errors* that ML formulas pump out under this assumption are. GMM gives a flexible way to derive at least and asymptotic set of corrections for statistical model misspecifications of the time-series regression coefficient. Similarly, a pooled time-series cross-sectional OLS regression is not inconsistent, but standard errors that ignore cross-correlation of error terms are far too small.

The "calibration" of real business cycle models is really nothing more than GMM, using economically sensible moments such as average output growth, consumption/output ratios etc. to avoid the stochastic singularity. Calibration exercises usually do not compute standard errors, nor do they report any distribution theory associated with the “evaluation” stage when one compares the model’s predicted second moments with those in the data. (I guess reporting *no* distribution theory is better than reporting a *wrong* distribution theory, but not much!) Following Christiano and Eichenbaum (19xx) however, it’s easy enough to calculate such a distribution theory by listing the first and second moments together. A J_T test probably doesn’t make much sense in this case, since we know the model can be rejected at any level of significance by choosing different moments.

“Used judiciously” is an important qualification. Many GMM estimations and tests suffer from lack of thought in the choice of moments, test assets and instruments. For example, early GMM papers tended to pick assets and especially instruments pretty much at random. Authors often included many lags of returns and consumption growth as instruments to test a consumption-based model. However, the 7th lag of returns really doesn’t predict much about future returns given lags 1-12, and the first-order serial correlation in seasonally adjusted, ex-post revised consumption growth may be economically uninteresting. Therefore, more recent tests tend to emphasize a few well-chosen assets and instruments that capture important and economically interesting features of the data.

14.3.2 Other arguments for ML vs. GMM

Finite sample distributions

Many authors say they prefer regression tests and the GRS statistic in particular because it has a finite sample distribution theory, and they distrust the finite-sample performance of the GMM asymptotic distribution theory.

This is not useful argument. First, the “finite sample” theory is, as usual in regression, only true *conditional* on the factor return. If you want to include sampling variation in the factor return in the conceptual sampling experiment, then even regression tests can only provide asymptotic answers. Second, the finite sample distribution only holds if returns really are normal and i.i.d., and if the factor is perfectly measured. Since these assumptions do not hold, it is not obvious that a finite-sample distribution that ignores all these effects will be a better approximation than an asymptotic distribution that corrects for them.

It is true that the GMM asymptotic distribution theory can be a poor approximation to a finite-sample distribution theory, especially when one asks “non-parametric” corrections for autocorrelation or heteroskedasticity to provide large corrections and when the number of moments is large compared to the sample size. However, a “finite sample” distribution theory that ignores the effects for which GMM is correcting is not obviously better.

As detailed in section xx, an underused idea (at least in my opinion) is to describe the cross-correlation, autocorrelation, heteroskedasticity, etc. by parametric models as one would in ML when calculating the GMM distribution theory. For example, rather than calculate $\sum_{j=-\infty}^{\infty} E(u_t u_{t-j})$ from its sample counterpart, model $u_t = \rho u_{t-1} + \varepsilon_t$, estimate ρ, and then calculate $\sigma^2(u)\sum_{j=-\infty}^{\infty}\rho^j = \sigma^2(u)\frac{1+\rho}{1-\rho}$. This approach may give better small sample performance than the "nonparametric" corrections.

Once you have picked the estimation method – how you will generate a number from the data; or which moments you will use – finding its finite sample distribution, given an auxiliary statistical model, is simple. Just run a Monte Carlo or bootstrap. Thus, picking an estimation method because it delivers analytic formulas for a finite sample distribution (under false assumptions) should be a thing of the past. Analytic formulas for finite sample distributions are useful for comparing estimation methods and arguing about statistical properties of estimators, but they are not necessary for the empiricists' main task.

Auxiliary model

ML requires an auxiliary, parametric, statistical model. In studying the classic ML formalization of regression tests, we had to stop to assume that returns and factors are jointly i.i.d. normal. In the ML estimate of a consumption-based model, we had to worry equally about estimating statistical parameters $(V, \Delta c)$ of the consumption-return distribution along with the economic parameters (γ in this case) that we really care about. As the auxiliary statistical model becomes more and more complex and hence realistic, more and more effort is devoted to estimating the auxiliary statistical model. ML has no way of knowing that some parameters (risk aversion γ, β and λ) are more "important" than others.

A very nice feature of GMM is that it does not require such an auxiliary statistical model. For example, in studying GMM we went straight from $p = E(mx)$ to moment conditions, estimates, and distribution theory. This is most important as a saving of the researcher's and the reader's time effort and attention.

All of ML's complexity buys us one thing: a parametric expression for the optimal linear moments to set to zero. If one judges that the regular S matrix calculation does a good enough job of squeezing statistical information out of the sample, perhaps already trading too much efficiency for robustness, ML is not very attractive.

However, the absence of statistical modeling in GMM does rest on the asymptotic normality of sample means, together with "nonparametric" corrections for correlation and heteroskedasticity. The nonparametric corrections don't work that well in small samples, so one may want to model correlation and heteroskedasticity explicitly; in doing so one will again have to worry about the specification and estimation of an auxiliary statistical model.

The case for ML

There are cases in which ML, or a statistically motivated choice of moments, has important advantages. For example, Jacquier, Polson and Rossi (1994) study the estimation of a time-series model with stochastic volatility. This is a model of the form

$$\begin{aligned} dS_t/S_t &= \mu dt + V_t dZ_{1t} \\ dV_t &= \mu_V(V_t)dt + \sigma(V_t)dZ_{2t}, \end{aligned}$$

and S is observed but V is not. The obvious and easily interpretable moments include the autocorrelation of squared returns, or the autocorrelation of the absolute value of returns. However, they find in a simulation study that the resulting estimates are far less efficient than those resulting from the scores.

In advocating GMM so far, I have implicitly assumed that the economic model is approximate, the true economic model is unknown and the statistical model is approximate, and that the efficiency gain from ML is small. This is often true, but, as in this example, not always. ML's suggestion of moments can be valuable when the model is right (exactly right in any simulation study) so there is no tension between the moments in which one is interested and the scores on which ML focuses, when economically important moments are not obvious, and when the efficiency gain can be large.

Even in the canonical OLS vs. GLS case, a wildly heteroskedastic error covariance matrix can mean that OLS spends all its time fitting unimportant data points. A judicious application of GMM (OLS) in this case would require at least some transformation of units so that OLS is not wildly inefficient.

Conditioning information

Another advantage of the GMM approach with pricing error moments comes when we take seriously time-variation in mean returns and their standard deviation, and the fact that agents have a lot more information than we do. As we saw above, the GMM-pricing error method accommodates both features easily: $E(E(mx|I)) = E(mx)$. To model time-variation in returns in a ML context, you have to write out a parametric model of the time-varying return distribution; the scores will now be related to forecast errors rather than the returns themselves. Scores in such models are typically not easily interpretable as pricing errors, as the scores for simple i.i.d. models were. ML really doesn't allow us to think easily about agents who might have more information than we do.

General comments on statistical arguments

The history of statistical work that has been persuasive – that has changed people's understanding of the facts in the data and which economic models understand those facts – looks a lot different than the statistical theory preached in econometrics textbooks.

The CAPM was taught and believed in and used for years despite formal statistical rejections such as the GRS test. It only fell by the wayside when other, coherent views of the world were offered in the multifactor models. And the multifactor models are also rejected! It seems that "it takes a model to beat a model." Even when evaluating a specific model, most of the interesting tests come from examining specific alternatives rather than overall pricing error tests. The original CAPM tests focused on whether the intercept in a cross-sectional regression was higher or lower than the risk free rate, and whether individual variance entered into cross-sectional regressions. The CAPM fell when it was found that characteristics such as size and book/market do enter cross-sectional regressions, not when generic pricing error tests rejected.

In the history of finance and economics – actually in the history of science generally – no important issue has been decided by purely statistical considerations when methods of varying power disagree. Issues are decided when the profession collectively decides that the data sing a clear song, no matter what the t-statistics say.

Influential empirical work tells a story, not a t-statistic. The most efficient procedure does not seem to convince people if they cannot transparently see what stylized facts in the data drive the result. A test of a model that focuses on its ability to account for the cross section of average returns of interesting portfolios via their covariances with a state variable will in the end be much more persuasive than one that (say) focuses on the model's ability to explain the fifth moment of the second portfolio, even if ML finds the latter moment much more statistically informative. The papers that convinced the profession that returns are predictable at long horizons, or that factors past the market return are important in accounting for the cross-section of average returns, used no techniques past regression, but they made crystal clear what stylized and robust fact in the data drives the results. On the other hand, I can think of no case in which substantial changes in the way people thought about an issue resulted from the application of clever statistical models that wrung the last ounce of efficiency out of a dataset, changing t statistics from 1.5 to 2.5.

Given the non-experimental nature of our data, the inevitable fishing biases of many researchers examining the same data, and the unavoidable fact that our theories are really quantitative parables more than literal descriptions of the way the data are generated, the way the profession has decided things makes a good deal of sense. Statistical inference – classical or Bayesian – provides a poor description of the decision process we face in evaluating asset pricing models or any economic theory for that matter. Our objective is not to "accept" or "reject" a theory invented out of the blue, but always to refine it, to take a theory generated with some knowledge of the data, find out what aspects of the data it captures and what aspects it does not capture, and think about how it might be improved. To that end, lots of calculations are more revealing than test statistics.

Furthermore, the pretense of statistical purity is an illusion. Classical statistics requires that nobody ever looked at the data before specifying the model. Yet more regressions have been run than there are data points in the CRSP database. Bayesian econometrics can in principle incorporate the information of previous researchers, yet it never applied in this way – each study starts anew with a "uninformative" prior. Statistical theory draws a sharp distinction between the *model* – which we know is right; utility is exactly power; and the *parameters* which we estimate. But this distinction isn't true; we are just as uncertain about functional forms as we are about parameters.

We spend a lot of time on statistical theory, but we must realize that it is really a subsidiary question. The first question is, what is your economic model or explanation? Second, how did you produce your numbers from the data at hand, and was that a reasonable way to go about it? Third, are the model predictions robust to the inevitable simplifications? (Does the result hinge on power utility vs. another functional form? What happens if you add a little measurement error, or if agents have an information advantage, etc.) Finally, someone in the back of the room might raise his hand and ask, "if the data were generated by (say) a draw of i.i.d. normal random variables over and over again, how often would you come up with a number this big or bigger?" That's an interesting and important robustness check on what you did, but not necessarily the first such check, and not the central question in your and the profession's evaluation of whether your analysis of the data and models should change their minds. Similarly, statistical testing answers a very small and perhaps not very important question. It answers the question, "if your model were exactly true, and given an auxiliary statistical model, how often would you see a result this big (a parameter estimate, or a sum of squared pricing errors) due only to sampling variation?"

As we have seen, a lot of the arguments for GMM vs. maximum likelihood are statistical. The asymptotic distribution theory for GMM estimators and test statistics does not require one to use an explicit parametric model of distributions, and can therefore be robust to non-normality, conditional heteroskedasticity, serial correlation, and other statistical problems in the data. On the other hand, if the auxiliary statistical models are right, maximum likelyhood is more "efficient," and the "nonparametric" corrections often used in GMM applications may have poor small sample properties. However, if auxiliary statistical models are wrong, maximum likelyhood can provide very misleading estimates.

But in the end, *statistical* properties may be a poor way to choose statistical methods. I prefer GMM in most cases because it is a tool that allows me to evaluate the model in the simplest, most natural and most transparent way– just use sample averages in place of the population moments that are most economically important to the quantitative parable of the theory. Each step of the way has a clear intuition, and it is easy to trace results back to stylized facts of the data that generate them. There is no need to separate theorists from empirical workers with that approach. Even more important, the procedures one follows in constructing GMM estimates and tests are very easy. (Proving that GMM works in very general setups is hard, which is where it gets its unfortunate high-tech reputation.)

Both ML and GMM are best thought of tools that a thoughtful researcher can use in learning what the data says about a given asset pricing model, rather than as stone tablets giving precise directions that lead to truth if followed literally. If followed literally and thoughtlessly, both ML and GMM can lead to horrendous results.

14.4 Problems

1. When we express the CAPM in excess return form, can the test assets be differences between risky assets, $R_i - R_j$? Can the market excess return also use a risky asset, or must it be relative to a risk free rate?
(A: The test assets can be risky return differences, but the market excess return must be relative to a risk free rate proxy (which may be an estimated parameter). $E(R_i) - R_f = \beta_{i,m}\left(E(R_m) - R^f\right)$ implies $E(R^i - R^j) = \beta_{i-j,m}E(R^m - R^f)$ but not $= \beta_{i-j,m-j}\left(E(R^m - R^j)\right)$
2. Can you run the GRS test on a model that uses industrial production growth as a factor?
3. Show that if CAPM holds for a set of test assets it holds for market, IF the market is

spanned by the test assets. Is this true for any return-based factor model? (A: no).

4. Try to formulate a ML estimator based on an unrestricted regression when factors are not returns, equation (120). i.e. add pricing errors α_i to the regression as we did for the unrestricted regression in the case that factors are returns. What is your estimate of $\mathbf{B}$, $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, $E(\mathbf{f})$? (Treat V and Σ as known to make the problem easier.)
Answer: Adding pricing errors to (156), we obtain

$$R_t^{ei} = \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{\lambda} + \boldsymbol{\beta}_i'\left[\mathbf{f}_t - E(\mathbf{f_t})\right] + \varepsilon_t^i.$$

Stacking assets $i = 1, 2, ...N$ to a vector

$$\mathbf{R}_t^e = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\lambda} + \mathbf{B}\left[f_t - E(f_t)\right] + \boldsymbol{\varepsilon}_t$$

where $\mathbf{B}$ denotes a $N \times K$ matrix of regression coefficients of the N assets on the K factors.
If we fit this model, maximum likelyhood will give asset-by asset OLS estimates of the intercept $\mathbf{a} = \boldsymbol{\alpha} + \mathbf{B}(\boldsymbol{\lambda} - E(\mathbf{f}_t))$ and slope coefficients $\mathbf{B}$. It will not give separate estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. The most that the regression can hope to estimate is one intercept; if one chooses a higher value of λ, we can obtain the same error term with a lower value of α. The likelyhood surface is flat over such choices of α and λ. One could do an ad-hoc second stage, minimizing (say) the sum of squared $\boldsymbol{\alpha}$ to choose λ given $\mathbf{B}$, $E(\mathbf{f}_t)$ and $\mathbf{a}$. This intuitively appealing procedure is exactly a cross-sectional regression. But it would be ad-hoc, not ML.

5. Instead of writing a regression, build up the ML for the CAPM a little more formally. Write the statistical model as just the assumption that individual returns and the market return are jointly normal,

$$\left[\begin{array}{c} \mathbf{R}^e \\ R^{em} \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} E(\mathbf{R}^e) \\ E(R^{em}) \end{array}\right], \begin{array}{cc} \Sigma & cov(R^{em}, \mathbf{R}^{e\prime}) \\ cov(R^{em}, \mathbf{R}^e) & \sigma_m^2 \end{array}\right)$$

The model's restriction is

$$E(\mathbf{R}^e) = \gamma cov(R^{em}, \mathbf{R}^e).$$

Estimate γ and show that this is the same time-series estimator as we derived by presupposing a regression.

6. Fama and French (19xx) report that pricing errors are correlated with betas in a test of a factor pricing model on industry portfolios. How is this possible?
A:Yes with time-series regressions. No with a cross-sectional OLS regression.

14.5 References

Hamilton (1994) p.142-148 are a nice summary of maximum likelyhood facts. The appendix in Campbell Lo MacKinlay (199x) is also a nice maximum likelyhood reference, and their

Chapter 5 and 6 treat regression based tests and maximum likelyhood in more depth than I do here.

E. Jacquier N. Polson and Peter Rossi, “Bayesian Analysis of Stochastic Volatility Models,” Journal of Business and Economic Statistics (1994), 12 , 371-418.

PART III
General equilbrium and derivative pricing

Chapter 15. General Equilibrium

So far, we have not said where the joint statistical properties of the payoff x_{t+1} and marginal utility m_{t+1} or consumption c_{t+1} come from. We have also not said anything about the fundamental exogenous shocks that drive the economy. The basic pricing equation $p = E(mx)$ tells us only what the price should be, *given* the joint distribution of consumption (marginal utility, discount factor) and the asset payoff.

Similarly, there is nothing that stops us from writing the basic pricing equation as

$$u'(c_t) = E_t\left[\beta u'(c_{t+1})x_{t+1}/p_t\right].$$

Now, we can think of this equation as determining today's *consumption* given asset prices and payoffs, rather than determining today's *asset price* in terms of consumption and payoffs. Thinking about the basic first order condition in this way, with asset prices as given and consumption as the quantity to be determined, is exactly the basis of the permanent income model of consumption. Which is the chicken and which is the egg? Which variable is exogenous and which is endogenous?

The answer for now is, neither. The first order conditions characterize any equilibrium; if you happen to know $E(mx)$, you can use them to determine p; if you happen to know p, you can use them to determine consumption and savings decisions.

An obvious next step, then is to complete the solution of our model economy; to find c and p in terms of truly exogenous forces. The results will of course depend on what the rest of the economy looks like, in particular what the *production technology* is and what the set of markets is.

Figure 22 shows one possibility for a general equilibrium. Suppose that the production technologies are linear: the real, physical rate of return (the rate of intertemporal *transformation*) is not affected by how much is invested. Now consumption must adjust to these technologically given rates of return. If the rates of return on the intertemporal technologies were to change, the consumption process would have to change. This is, implicitly, how the permanent income model works. More explicitly, this is how many finance theories such as the CAPM and (more explicitly) the Cox, Ingersoll and Ross (1986) model of the term structure work.

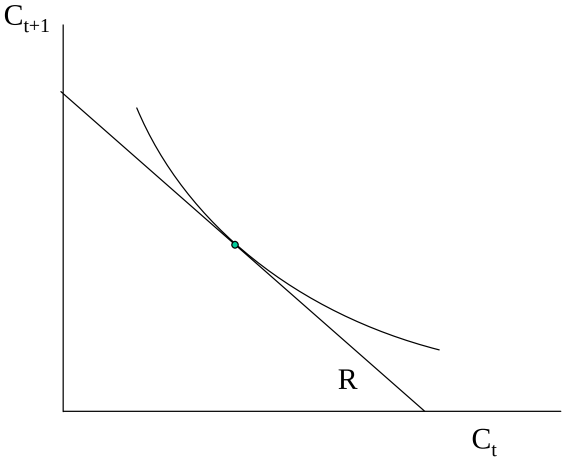


Figure 22. Consumption adjusts when the rate of return is determined by a linear technology.

Figure 23 shows another extreme possibility for the production technology. This is an "endowment economy." Nondurable consumption appears (or is produced) every period. There is nothing anyone can do to save, store, invest or otherwise transform consumption goods this period to consumption goods next period. Hence, asset prices must adjust until people are just happy consuming the endowment process. In this case consumption is exogenous and asset prices adjust. Lucas (1978) and Mehra and Prescott (1985) are two very famous applications of this sort of "endowment economy."

Which of these possibilities is correct? Well, neither of course. The real economy and all serious general equilibrium models look something like figure 24: one can save or transform consumption from one date to the next, but at a decreasing rate. As investment increases, rates of return decline

Does this observation invalidate any modeling we do with the linear technology (CAPM, CIR) model, or endowment economy model? No. Start at the equilibrium in figure 24. Suppose we model this economy as a linear technology, but we happen to choose for the rate of return on the linear technologies exactly the same stochastic process that emerges from the general equilibrium. The resulting joint consumption, asset return process is exactly the same as in the original general equilibrium! Similarly, suppose we model this economy as an endowment economy, but we happen to choose for the endowment process exactly the stochastic process for consumption that emerges from the equilibrium with a concave technology. Again, the joint consumption-asset return process is exactly the same.

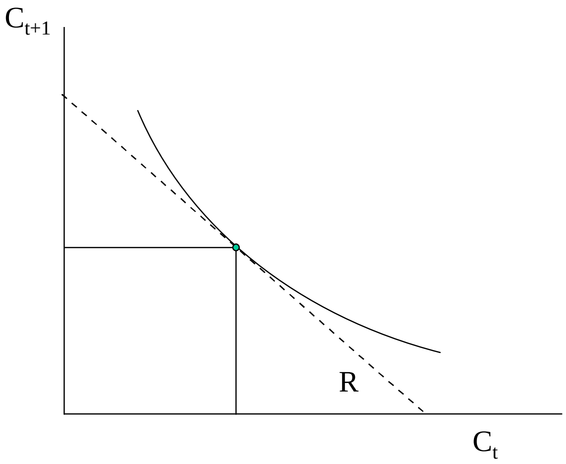


Figure 23. Asset prices adjust to consumption in an endowment economy.

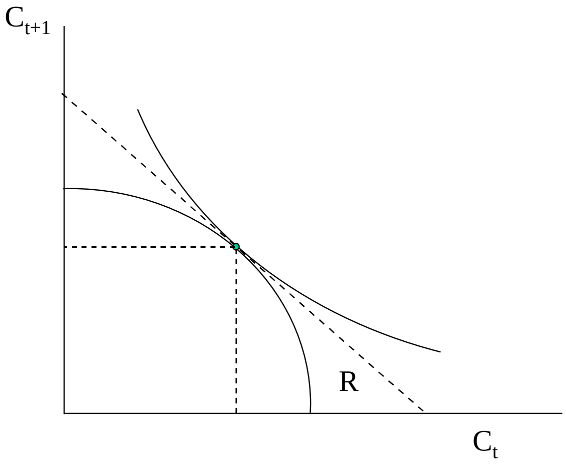


Figure 24. General equilibrium. The solid lines represent the indifference curve and production possibility set. The dashed straight line represents the equilibrium rate of return. The dashed box represents an endowment economy that predicts the same consumption-asset return process.

Therefore, there is nothing empirically wrong in adopting one of the following strategies: 1) Form a statistical model of bond and stock returns, solve the optimal consumption-portfolio decision. 2) Form a statistical model of the consumption process, calculate asset prices and returns from the basic pricing equation $p = E(mx)$. If the statistical models are right, and if they coincide with the equilibrium consumption or return process generated by the true economy, with concave technology, each approach will give correct predictions for the joint consumption-asset return process.

As we will see, most finance models, developed from the 1950s through the early 1970s, take the return process as given, implicitly assuming linear technologies. The endowment economy approach, introduced by Lucas (1978), is a breakthrough because it turns out to be much easier. It is much easier to evaluate $p = E(mx)$ for fixed m than it is to solve joint consumption-portfolio problems for given asset returns. To solve a consumption-portfolio problem we have to model the investor's entire environment: we have to specify *all* the assets to which he has access, what his labor income process looks like (or wage rate process, and include a labor supply decision). Once we model the consumption stream directly, we can look at each asset in isolation, and the actual computation is almost trivial.

Most uses of $p = E(mx)$ do not require us to take any stand on exogeneity or endogeneity, or general equilibrium. This is a condition that must hold for any asset, for any production technology. Having a taste of the extra assumptions required for a general equilibrium model, you can now appreciate why people stop short of full solutions when they can address an application using only the first order conditions, using knowledge of $E(mx)$ to make a prediction about p.

It is enormously tempting to slide into an interpretation that $E(mx)$ *determines* p. We routinely think of betas and factor risk prices – components of $E(mx)$ – as *determining* expected returns. For example, we routinely say things like "the expected return of a stock increased *because* the firm took on riskier projects, thereby increasing its β". But the whole consumption process, discount factor, and factor risk premia change when the production technology changes. Similarly, we are on thin ice if we say anything about the effects of policy interventions, new markets and so on. The equilibrium consumption or asset return process one has modeled statistically may change in response to such changes in structure.

15.0.1 Normal-exponential example.

The normal-exponential model is a nice place in which to see how the general equilibrium aspects of our models work out. If you recall, by assuming normally distributed returns R with mean $E(R)$ and covariance matrix Σ, a riskfree rate R^f, a stock of initial wealth W that must be divided among assets $W = y^f + y'1$, and exponential utility $Eu(c) = \exp(-\alpha c)$, we derived first order conditions

$$E(R) - R^f = \alpha \Sigma y = \alpha \; cov(R, R^m). \tag{170}$$

There are two ways to complete this model and describe the general equilibrium. This is the same thing as thinking about the supply of assets. First, we can think of the returns as corresponding to linear technologies. Then the returns are invariant to the amounts invested y. How can this model then "determine" expected returns, if they are immutable features of the technology? It doesn't. If an expected return $E(R^i)$ is really high, then investors will buy more of it, raising y^i. As they keep doing this, the share of R^i in the market return keeps rising, and $cov(R^i, R^m)$ keeps rising, until the two sides are equated. Thus, if we pair these preferneces with a linear technology, the amounts invested y are endogenous, and it is the market return rather than the individual expected returns or covariances which adjust.

Second, let us complete this model with a fixed supply of assets, or a Leontief technology. Each asset corresponds to a payoff x^i, which is normally distributed and fixed. Larger amounts invested in each asset do not raise the payoff at all. Now, the prices of each payoff x^i or equivalently the returns $R^i = x^i/p^i$ are the quantities that adjust to achieve equilibrium. The total quantity invested in each technology is equal to the price of payoff x^i, $y^i = p^i$. Substituting, and multiplying through by prices, (170) is equivalent to

$$E(x) - pR^f = \alpha\Sigma_x\mathbf{1}$$

where Σ_x is the covariance matrix of the payoffs x, or,

$$p = \frac{1}{R^f}\left(E(x) - \alpha\Sigma_x\mathbf{1}\right). \tag{171}$$

Again, we have a beautiful equation. The first term is of course the risk neutral present value formula. The second term is a risk correction. Prices are lower if risk aversion is higher, and if a security has a higher covariance with wealth = $x'\mathbf{1}$. This equation is also linear, which is a reason that this model is very popular for theoretical work.

(170) and (171) are of course the same thing; each holds under the other's general equilibrium assumptions. All I have done in each case is solve for the variable that is endogenous, y or p in each case.

15.1 General Equilibrium Recipe

1) find quantity dynamics 2) price assets from consumption-based model

15.2 Endowment economies

15.2.1 Mehra-Prescott style

15.2.2 Arbitrary law of motion for consumption

15.2.3 Show how b's etc. are all determined in the model

15.2.4 Beware stochastic singularities

dividend example

15.3 Production economies

15.3.1 Log Cobb-Douglas model

15.3.2 Linear - quadratic model

15.3.3 Any model

15.4 Justification for the procedure

15.4.1 Welfare theorems

15.4.2 Aggregation to representative consumer

15.4.3 Asset pricing in distorted economies

15.5 Risk sharing and security design.

Complete risk sharing.

CAT story

15.6 Incomplete markets economies.

Risk sharing in an incomplete market

Heaton/Lucas, Constantinides/Duffie.. saving up to avoid constraints,

15.7 Outlook

(move to appropriate sections)

GE fundamental question seems farther off than ever.

rationality has hardest time at every turn. Perhaps illiquid, sloping demand curves.

Testing: individual securities rather than portfolios. covariance matrix restrictions

Chapter 16. Continuous time and derivative pricing

Continuous time models have a fearsome reputation. Unfortunately the language of much continuous time finance is often so different that much of the profession has separated into discrete-time and continuous-time camps that communicate very little. This is as unnecessary as it is unfortunate. In this chapter, I'll show how all the ideas of the previous chapters extend naturally to continuous time.

The choice of discrete vs. continuous time is merely one of modeling convenience. The richness of the theory of continuous time processes often allows you to obtain analytical results that would be unavailable in discrete time. On the other hand, in the complexity of most practical situations, one ends up resorting to numerical simulation of a discretized model anyway. In those cases, it might be clearer to start with a discrete model. But I emphasize this is all a choice of language. One should become familiar enough with discrete as well as continuous time representations of the same ideas to pick the one that is most convenient for a particular application.

16.1 Diffusion models

---

dz is defined by $z_{t+\Delta t} - z_t \ \sim \ \mathcal{N}(0, \Delta t)$.

Diffusion models $dx = \mu(\cdot)dt + \sigma(\cdot)dz$

Diffusion models are the standard way to represent random variables in continuous time.

16.1.1 Brownian motion and diffusion

The simplest example is a *Brownian motion* which is the natural generalization of a random walk in discrete time. For a random walk

$$z_t - z_{t-1} = \varepsilon_t$$

the variance scales with time; $var(z_{t+2} - z_t) = 2var(z_{t+1} - z_t)$. Thus, define a Brownian motion as a process z_t for which

$$z_{t+\Delta t} - z_t \ \sim \ \mathcal{N}(0, \Delta t).$$

We have added the normal distribution to the usual definition of a random walk. Brownian motions have the interesting property that the sample paths (z_t plotted as a function of t) are *continuous* but nowhere *differentiable*.

In discrete time, uncorrelated random variables are the basic building blocks of time series. We can use the increments to Brownian motions the same way in continuous time. Construct a series

$$x_{t+\Delta t} - x_t = \mu(\cdot)\Delta t + \sigma(\cdot)(z_{t+\Delta t} - z_t)$$

or, using d to denote arbitrarily small increments,

$$dx = \mu(\cdot)dt + \sigma(\cdot)dz$$

μ and σ can be functions of time directly, or of state variables. For example, we might have $\mu(\cdot) = \mu(x,t)$. In discrete time, we are used to analogs to μ that are only linear functions of past values, but in continuous time we can tractably handle nonlinear functions μ and σ as well.

It's important to be clear about the notation. dx means $x_{t+\Delta t} - x_t$. We often bandy about dx thinking about the derivative of a function, but since a Brownian motion is not a differentiable function of time, $dz = \frac{dz(t)}{dt}dt$ makes no sense. So let dx mean the increment; if x is a differentiable function of time, $dx = dx(t)/dt\ dt$ will be meaningful, but not otherwise. (We'll soon see how to modify this equation so it does make sense.)

A natural step is to take a *differential equation* like this one and simulate (integrate) it forward through time to obtain the finite-time random variable $x_{t+\Delta t}$. Sticklers for precision in continuous time prefer to always think of random variables this way rather than through the differential notation. Putting some arguments in for μ and σ for concreteness, you can think of evaluating the integral

$$x_T - x_0 = \int_0^T dx_t = \int_0^T \mu(x_t, t, ..)dt + \int_0^T \sigma(x_t, t, ..)dz_t.$$

Initially, it's a little disconcerting to see dz_t and dt as separate arguments, but we have to do this. z is not differentiable, so you can't write $dz = (dz/dt)dt$. But you can add up the increments to dz to find out where z ends up just as you can add up the increments to dt to find out where t ends up, and you can multiply each increment dz by some amount $\sigma(x_t, t, ..)$ before you add it up. The notation $\int_0^T \sigma(x_t, t, ..)dz_t$ just tells you to add up increments.

If you have functional forms for μ and σ and are good at integrating, you can see this procedure will give us the *distribution* of x at some future date, or at least some characterizations of that distribution such as conditional mean, variance etc.

16.1.2 A little toolkit of processes.

$dx = \mu dt + \sigma dz.$

$dx_t = -\phi(x-\mu)\ dt + \sigma\ dz$

$dx_t = -\phi(x-\mu)\ dt + \sigma\sqrt{x}\ dz$

$\frac{dp}{p} = \mu dt + \sigma dz.$

Like the AR(1) and MA(1), there are some standard useful workhorse examples of diffusion models.

Random walk with drift. The simplest example is to let μ and σ be constants. Then

$$dx = \mu dt + \sigma dz.$$

It's easy to figure out discrete time implications for this process,

$$x_{t+s} = x_t + \mu s + \sigma(z_{t+s} - z_t)$$

or

$$x_{t+s} = x_t + \mu s + \varepsilon_{t+s};\ \ \varepsilon_{t+s}\tilde{}\mathcal{N}(0, \sigma s)$$

a random walk with drift.

AR(1). The simplest discrete time process is an AR(1); this is its obvious continuous time counterpart. In discrete time,

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

can be written

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

In continuous time, write

$$dx_t = -\phi(x-\mu)\ dt + \sigma\ dz$$

The drift $-\phi(x-\mu)$ pulls x back to its steady state value μ.

Square root process. Like its discrete time counterpart, the continuous time AR(1) ranges over the whole real numbers. It would be nice to have a process that was always positive, so it could capture a price or an interest rate. A natural extension of the continuous time AR(1) is a workhorse of such applications,

$$dx_t = -\phi(x-\mu)\ dt + \sigma\sqrt{x}\ dz.$$

Now, as x approaches zero, the volatility declines. At $x=0$, the volatility is entirely turned off, so x drifts up to $\dot{\mu}$. We will show more formally below that this behavior keeps $x \geq 0$ always; the conditional and unconditional distributions of such a process stop at zero.

This is a nice example because it is decidedly *nonlinear*. We could write its discrete time analogue, but standard time series tools would fail us. We could not, for example, give a pretty equation for the distribution of x_{t+s} for finite s. We can do this in continuous time.

Price processes A modification of the random walk with drift is the most common model for prices. We want the *return* or *proportional* increase in price to be uncorrelated over time. The most natural way to do this is to specify

$$dp = p\mu dt + p\sigma dz$$

or more simply

$$\frac{dp}{p} = \mu dt + \sigma dz.$$

We most easily capture dynamics – variation in expected returns or conditional variance of returns – by making the μ and σ in this representation vary over time or in response to state variables.

16.2 Ito's lemma

Do second order Taylor expansions, keep dz, dt,and $dz^2 = dt$ terms.

$dy = f'(x)dx + \frac{1}{2}f''(x)dx^2$

$dy = \left(f'(x)\mu_x + \frac{1}{2}f''(x)\sigma_x^2\right) dt + f'(x)\sigma_x dz$

You often have a diffusion representation for one variable, say

$$dx = \mu_x(\cdot)dt + \sigma_x(\cdot)dz.$$

Then you define a new variable in terms of the old one,

$$y = f(x).$$

Naturally, you want a diffusion representation for y. Ito's lemma tells you how to get it. It says,

Use a *second* order Taylor expansion, and think of dz as $\sqrt{dt}$; thus as $\Delta t \rightarrow 0$ keep terms dz, dt, and $dz^2 = dt$, but terms $dtdz$ and beyond go to zero.

Let's go step by step. Start with the second order expansion

$$dy = \frac{df(x)}{dx}dx + \frac{1}{2}\frac{d^2f(x)}{dx^2}dx^2$$

Now

$$dx^2 = [\mu_x dt + \sigma_x dz]^2 = \mu_x^2 dt^2 + \sigma_x^2 dz^2 + 2\mu_x \sigma_x dt dz.$$

But $dt^2 = 0$, $dz^2 = dt$ and $dtdz = 0$. Thus,

$$dx^2 = \sigma_x^2 dt$$

Substituting for dx and dx^2,

$$\begin{aligned} dy &= \frac{df(x)}{dx}[\mu_x dt + \sigma_x dz] + \frac{1}{2}\frac{d^2 f(x)}{dx^2}\sigma_x^2 dt \\ &= \left(\frac{df(x)}{dx}\mu_x + \frac{1}{2}\frac{d^2 f(x)}{dx^2}\sigma_x^2\right) dt + \frac{df(x)}{dx}\sigma_x dz \end{aligned}$$

Thus, *Ito's lemma*.

$$dy = \left(\frac{df(x)}{dx}\mu_x(\cdot) + \frac{1}{2}\frac{d^2 f(x)}{dx^2}\sigma_x^2(\cdot)\right) dt + \frac{df(x)}{dx}\sigma_x(\cdot) dz$$

The surprise here is the second term in the drift. Intuitively, this term captures a "Jensen's inequality" effect. If a is a mean zero random variable and $b = a^2 = f(a)$, then the mean of b is higher than the mean of a. The more variance of a, and the more concave the function, the higher the mean of b.

The only thing you have to understand is, why is $dz^2 = dt$? Once you know $dz^2 = dt$ it's clear we have to keep the dz and dz^2 terms in an expansion, and we need second order expansions to do so. Think of where dz came from. $dz = z_{t+\Delta t} - z_t$ is a normal random variable with variance Δt. That means its *standard deviation* is $\sqrt{\Delta t}$. Thus, clearly dz^2 is of *order* dt, and dz of order, or "typical size" $\sqrt{dt}$. In fact, dz^2 really equals dt; in the limit dz^2 becomes deterministic.

16.2.1 Examples

1) Log. A classic example and a common fallacy. Suppose a price follows

$$\frac{dp}{p} = \mu dt + \sigma dz$$

What is the diffusion followed by the log price,

$$y = \ln(p)?$$

Applying Ito's lemma,

$$dy = \frac{1}{p}dp - \frac{1}{2}\frac{1}{p^2}dp^2 = \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dz.$$

Not

$$dy = \mu dt + \sigma dz.$$

It is *not* true that $dy = d(\ln(p)) = dp/p$. You have to include the second order terms.

2) xy. Usually, we write

$$d(xy) = xdy + ydx$$

But this expression comes from the usual first order expansions. When x and y are diffusions, we have to keep second order terms. Thus,

$$d(xy) = xdy + ydx + dydx$$

16.3 Densities

One of the nice things about continuous time processes is that we can analytically characterize the distributions of nonlinear processes.

16.3.1 Forward and backward equations

16.3.2 Stationary density

The *stationary density* of a stationary process is the unconditional density, or the limit of the conditional density as time increases. The stationary density $f(x)$ of a diffusion $dx = \mu(x)dt + \sigma(x)dz$, if it exists, satisfies

$$\mu(x)f(x) = \frac{1}{2}\frac{d}{dx}\left[\sigma^2(x)f(x)\right]$$

or

$$\frac{d}{dx}\left(e^{-2\int^x dv \frac{\mu(v)}{\sigma^2(v)}}\sigma^2(x)f(x)\right) = 0.$$

Hence,

$$f(x) = \frac{Ke^{2\int^x dv \frac{\mu(v)}{\sigma^2(v)}}}{\sigma^2(x)}.$$

More simply, let

$$s(x) = \frac{e^{2\int^x dv \frac{\mu(v)}{\sigma^2(v)}}}{\sigma^2(x)}$$

then $1 = \int f(x)$ implies

$$f(x) = \frac{s(x)}{\int s(x)dx}.$$

16.4 Tricks

If x follows a diffusion, there are no attracting boundaries, and the process is stationary, then the drift diffusion and stationary density are realted by

$$\mu(x) = \frac{1}{2}\left[\frac{f'(x)}{f(x)}\sigma^2(x) + \frac{d}{dx}\sigma^2(x)\right]$$

(Ait-sahalia 1986 uses this fact to esimate the diffusion function from the drift and stationary density.

16.5 Tricks

If x follows a diffusion, there are no attracting boundaries, and the process is stationary, then the drift diffusion and stationary density are realted by

$$\mu(x) = \frac{1}{2}\left[\frac{f'(x)}{f(x)}\sigma^2(x) + \frac{d}{dx}\sigma^2(x)\right]$$

(Ait-sahalia 1986 uses this fact to esimate the diffusion function from the drift and stationary density.

16.6 Black Scholes with discount factors

Write a process for stock and bond, then use Λ^* to price the option. The Black-Scholes formula results.

As an immediate application we can derive the Black-Scholes formula. This case shows some of the interest and engineering complexity of continuous time models. Though at each instant the analysis is trivial law of one price, chaining it together over time is not trivial either mathematically or in the result we get. I also want to show how thinking of the world in terms of a discount factor is (at least) as easy as other approaches.

The standard approach to the Black-Scholes formula rests on explicitly constructing portfolios: at each date we construct a portfolio of stock and bond that replicates the instantaneous payoff of the option; we reason that the price of the option must equal the price of the replicating portfolio. Instead, I'll follow the discount factor approach: at each date construct a discount factor that prices the stock and bond, and use that discount factor to price the option.

A stock follows

$$\frac{dS}{S} = \mu_S dt + \sigma_S dz.$$

There is also a money market security that pays the real interest rate rdt.

We use the theorem of the last section: to price the stock and interest rate, the discount factor must be of the form

$$\frac{d\Lambda}{\Lambda} = -rdt - \frac{(\mu_S - r)}{\sigma_s} dz - dw; \quad E(dwdz) = 0.$$

(you might want to check that this set of discount factors does in fact price the stock and interest rate.) Now we price the call option with this discount factor, and show that the Black-Scholes equation results. Importantly, the choice of discount factor via choice of dw has *no* effect on the resulting option price. *Every* discount factor that prices the stock and interest rate gives the same value for the option price. The option is therefore priced using the law of one price alone.

16.6.1 Method 1: Price using discount factor

Let us use the discount factor to price the option directly:

$$C_t = E_t \left\{ \frac{\Lambda_T}{\Lambda_t} \max\left(S_T - X, 0\right) \right\} = \int \frac{\Lambda_T}{\Lambda_t} \max\left(S_T - X, 0\right) \; df\left(\Lambda_T, S_T\right)$$

where Λ_T and S_T are solutions to

$$\frac{dS}{S} = \mu_S dt + \sigma_S dz$$

$$\frac{d\Lambda}{\Lambda} = -rdt - \frac{\mu_S - r}{\sigma_S} dz - dw.$$

I start by setting dw to zero, and then I show that adding dw does not change the option price.

We can find analytical expressions for the solutions to these differential equations, (Arnold,

P. 138):

$$\frac{dX}{X} = \mu dt + \sigma dz$$

has solution

$$\ln X_t = \ln X_0 + \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma \left(z_t - z_0\right)$$

i.e., $\ln X$ is conditionally normal with mean $\ln X_0 + \left(\mu - \frac{\sigma^2}{2}\right) t$ and variance $\sigma^2 t$.

Thus,

$$\ln S_T = \ln S_t + \left(\mu_S - \frac{\sigma_S^2}{2}\right) (T-t) + \sigma_S \left(z_T - z_t\right)$$

$$\ln \Lambda_T = \ln \Lambda_t - \left(r + \frac{1}{2}\left(\frac{\mu_S - r}{\sigma_S}\right)^2\right) (T-t) - \frac{\mu_S - r}{\sigma_S} \left(z_T - z_t\right)$$

or, with

$$x = \frac{z_T - z_t}{\sqrt{T-t}} \sim \mathcal{N}\left(0,1\right),$$

we have

$$\ln S_T = \ln S_t + \left(\mu_S - \frac{\sigma_S^2}{2}\right) (T-t) + \sigma_S \sqrt{T-t} x$$

$$\ln \Lambda_T = \ln \Lambda_t - \left(r + \frac{1}{2}\left(\frac{\mu_S - r}{\sigma_S}\right)^2\right) (T-t) - \frac{\mu_S - r}{\sigma_S} \sqrt{T-t} x.$$

Then, we evaluate the call option from the integral

$$\begin{aligned} C_t &= \int_{S_T=X}^{\infty} \frac{\Lambda_T}{\Lambda_t} \left(S_T - X\right) \, df\left(\Lambda_T, S_T\right) = \\ &= \int_{S_T=X}^{\infty} \frac{\Lambda_T}{\Lambda_t} S_T \, df\left(\Lambda_T, S_T\right) - \int_{S_T=X}^{\infty} \frac{\Lambda_T}{\Lambda_t} X \, df\left(\Lambda_T, S_T\right). \end{aligned}$$

The objects on the right hand side are *known*. We know the distribution of the terminal stock price S_T and discount factor Λ_T. To find the call price, we just have to evaluate an expectation or integral. Often this is done numerically, but this example has enough structure that we can find an analytical formula at some cost in algebra.

Doing the integral

In general, we have to find a joint distribution for Λ_T and S_T. But S_T and Λ_T are transforms of the same Normal(0,1), which I'll denote $x,$ so we can reduce the problem to a single integral over x. Plugging in the above expressions for S_T and Λ_T,

$$C_t = \int_{S_T=X}^{\infty} \exp\left(-\left(r+\frac{1}{2}\left(\frac{\mu_S-r}{\sigma_S}\right)^2\right)(T-t)-\frac{\mu_S-r}{\sigma_S}\sqrt{T-t}x\right)\times$$

$$\times S_t \exp\left(+\left(\mu_S-\frac{\sigma_S^2}{2}\right)(T-t)+\sigma_S\sqrt{T-t}x\right)\ df\left(x\right)-$$

$$-X\int_{S_T=X}^{\infty} \exp\left(-\left(r+\frac{1}{2}\left(\frac{\mu_S-r}{\sigma_S}\right)^2\right)(T-t)-\frac{\mu_S-r}{\sigma_S}\sqrt{T-t}x\right)\ df\left(x\right)$$

I change variables to express the result as the integral of a $\mathcal{N}\left(0,1\right)$ rather than the expectation of a function of an $\mathcal{N}\left(0,1\right)$. Organize in powers of x,

$$C_t = \frac{1}{\sqrt{2\pi}}S_t\int_{S_T=X}^{\infty} \exp\left\{\left[\mu_S-r-\frac{\sigma_S^2}{2}-\frac{1}{2}\left(\frac{\mu_S-r}{\sigma_S}\right)^2\right](T-t)\right.$$

$$\left.+\left[\sigma_S-\frac{\mu_S-r}{\sigma_S}\right]\sqrt{T-t}x-\frac{1}{2}x^2\right\}dx$$

$$-\frac{1}{\sqrt{2\pi}}X\int_{S_T=X}^{\infty} \exp\left(-\left[r+\frac{1}{2}\left(\frac{\mu_S-r}{\sigma_S}\right)^2\right](T-t)-\frac{\mu_S-r}{\sigma_S}\sqrt{T-t}x-\frac{1}{2}x^2\right)\ dx.$$

Express as quadratic functions of x,

$$C_t = \frac{1}{\sqrt{2\pi}}S_t\int_{S_T=X}^{\infty} \exp\left\{-\frac{1}{2}\left(x-\left[\sigma_S-\frac{\mu_S-r}{\sigma_S}\right]\sqrt{T-t}\right)^2\right\}dx$$

$$-\frac{1}{\sqrt{2\pi}}Xe^{-r(T-t)}\int_{S_T=X}^{\infty} \exp\left\{-\frac{1}{2}\left(x+\frac{\mu_S-r}{\sigma_S}\sqrt{T-t}\right)^2\right\}dx.$$

The lower bound $S_T = X$ is, in terms of x,

$$\ln X = \ln S_T = \ln S_t + \left(\mu_S-\frac{\sigma_S^2}{2}\right)(T-t)+\sigma_S\sqrt{T-t}x$$

$$x = \frac{\ln X - \ln S_t - \left(\mu_S - \frac{\sigma_S^2}{2}\right)(T-t)}{\sigma_S\sqrt{T-t}}$$

Finally, we use

$$\frac{1}{\sqrt{2\pi}}\int_a^\infty e^{-\frac{1}{2}(x-\mu)^2}dx = \Phi\left(\mu - a\right)$$

i.e., $\Phi()$ is the area under the left tail of the normal distribution, to get

$$C_t = S_t\Phi\left(-\frac{\ln X - \ln S_t - \left(\mu_S - \frac{\sigma_S^2}{2}\right)(T-t)}{\sigma_S\sqrt{T-t}} + \left[\sigma_S - \frac{\mu_S - r}{\sigma_S}\right]\sqrt{T-t}\right)$$

$$-Xe^{-r(T-t)}\Phi\left(-\frac{\ln X - \ln S_t - \left(\mu_S - \frac{\sigma_S^2}{2}\right)(T-t)}{\sigma_S\sqrt{T-t}} - \frac{\mu_S - r}{\sigma_S}\sqrt{T-t}\right)$$

Simplifying, we get the Black-Scholes formula

$$C_t = S_t\Phi\left(\frac{\ln S_t/X + \left[r + \frac{1}{2}\sigma_S^2\right](T-t)}{\sigma_S\sqrt{T-t}}\right) - Xe^{-r(T-t)}\Phi\left(\frac{\ln S_t/X + \left[r - \frac{1}{2}\sigma_S^2\right](T-t)}{\sigma_S\sqrt{T-t}}\right).$$

16.6.2 Method 2: Derive Black-Scholes differential equation

Guess that the solution for the call price is a function of stock price and time to expiration, $C_t = C(S,t)$. We can use the basic pricing equation $0 = E_t\left(d\Lambda C\right)$ to derive a differential equation for the call price *function* of stock price and time to expiration, .

$$0 = E_t\left(d\Lambda C\right) = CE_t d\Lambda + \Lambda E_t dC + E_t d\Lambda dC.$$

We use Ito's lemma to find derivatives of $C(S,t)$,

$$dC = C_t dt + C_S dS + \frac{1}{2}C_{SS}dS^2$$

$$dC = \left[C_t + C_S S\mu_S + \frac{1}{2}C_{SS}S^2\sigma_S^2\right]dt + C_S S\sigma_S dz$$

Plugging into the first order condition and canceling Λdt, we get

$$0 = -rC + C_t + C_S S\mu_S + \frac{1}{2}C_{SS}S^2\sigma_S^2 - S\left(\mu_S - r\right)C_S.$$

$$0 = -rC + C_t + SrC_S + \frac{1}{2}C_{SS}S^2\sigma_S^2.$$

This is the Black-Scholes differential equation (Duffie p.238); solved with boundary condition

$$C = \max\{S_T - X, 0\}$$

it yields the familiar formula.

16.7 Arbitrage bounds using discount factors

Chapter 17. “Good-deal” pricing

Chapter 18. Term structure of interest rates

18.1 Overview

18.2 Models based on a short rate process

18.3 Ho-Lee approach

18.4 Use to price options

Chapter 19. Frictions

Short sale borrowing constraints transactions costs.

all cases of short sale constraints

Sublinear extension of basic theorems.

Luttmer T bill results

Constantinides trade less often results

PART IV
Empirical survey

Chapter 20. Return Predictability

20.1 Stocks

20.1.1 Univariate: Long horizon regressions and variance ratios,

20.1.2 Multivariate: Term, d/p, and anomalies.

20.2 Term structure

Fama/Bliss, Campbell

20.3 Comparison to continuous-time models

20.4 Foreign exchange

20.5 Econometric issues

Big picture: yield differences don’t predict right changes.

20.6 Conditional variance

20.7 Conditional Sharpe ratios and portfolio implications

Conditional mean vs. conditional variance. Some ARCH evidence on c var

Want to estimate E(R)/sigma(R). Brandt

Chapter 21. Present value tests

Issue ex-post volatility, R^2

- Volatility tests. bound and decomposition.
- Equivalence to forecastability
- Bubbles and sunspots
- Use ks to illustrate identity.
- Cross-sectional

Chapter 22. Factor pricing models

22.1 Capm

22.2 Chen Roll Ross model

22.3 Investment and macro factors

Jagannathan Wang

Campbell

22.4 Book to market

22.5 Momentum and more

22.6 Digesting the tests

22.6.1 Statistics versus plots

Statistics never convinced anyone. But plots depend on portfolios.

Solution: individual securities?

22.6.2 Portfolios

Recently followed anomalies. Why size etc portfolios? some portfolio based on characteristic. Natural characteristic is beta, Fama McBeth did beta. But ad-hoc seem to give better spread in E(R), so chasing ad-hoc characteristics.

What happened to APT?

Chapter 23. Consumption based model tests

23.1 CRRA utility

Hansen and Singleton, updates

23.2 Durable goods

23.3 Habits

23.4 State-nonseparabilities

Utility function modifications Epstein Zin, habits, etc.

Chapter 24. Hansen-Jagannathan bounds and equity premium puzzle.

Much work on the consumption-based model has proceeded by shooting in the dark. A model is rejected, for reasons that are unclear. One then uses introspection to dream up a new utility function, tries it out on the data, rejects it, and iterates. Progress is slow in this loop. It would clearly be more productive to find what qualitative properties of the data drive rejections of a given model. This knowledge could give our search for new models of m some target. The GMM diagnostics described above are one approach to this characterization. Hansen-Jagannathan bounds are another approach to *diagnosing* the failures of a model.

As another motivation, it is desirable to say more about the performance of a model than just to "reject" or "fail to reject" it. Statistical tests answer the questions "is this model literally true, except for sampling variation?" That's often not a useful *question.* We are interested in ways of characterizing the performance of false models as well as testing for truth. For instance, it is interesting to know if a rejected model produces expected return errors of 0.001% rather than 10%. Above, I advocated examination of the pricing errors, along with GMM-based standard errors and ad-hoc weighting matrices to this end. The Hansen-Jagannathan bound provides another set of characterization tools.

The basic idea is to summarize a set of asset data by "what discount factors are consistent with this set of asset data?" Then, we can try on each model in turn, to see if its discount factor satisfies the characterization. Instead of performing (#data sets $\times$ # models) tests, we need only perform #data sets + #models calculations. Better yet, knowing what characteristics of the discount factor we need should be helpful information in constructing new models. For example, Campbell and I (1997) reverse-engineered a utility function to generate the conditional heteroskedasticity in the discount factor that we knew we needed, from this kind of diagnostic.

The basic Hansen-Jagannathan bound characterizes discount factors by mean and variance. Knowledge that standard data sets require a large discount factor variance been a great spur to development of the consumption-based model. I'll also survey extensions of the bound from Cochrane and Hansen (1991) that characterize the correlation of discount factors with asset returns, and the predictability and conditional heteroskedasticity of the discount factor. Much work remains to be done on finding other interesting moments or characterizations that will be useful for constructing asset pricing models.

Shiller (198x) made the first calculation that showed either a large risk aversion coefficient or counterfactually large consumption variability was required to explain means and variances of asset returns. Mehra and Prescott (198x) labeled this fact the "equity premium puzzle" and described the risk free rate puzzle discussed below. However, they described these puzzles in the context of a two-state Markov model for consumption growth, identifying a stock as a claim to consumption and a risk free bond. As we will see, the equity premium-risk free rate puzzle can be deduced from $1 = E(mR)$ and the basic moments of asset returns, without all the rest of the Mehra-Prescott structure.

24.1 The basic HJ bound and equity premium

$$\frac{\sigma(m)}{E(m)} \geq \frac{E(R^e)}{\sigma(R^e)}.$$

In postwar US data, this calculation implies $\sigma(m) \geq 50\%$ on an annual basis, requiring huge risk aversion or consumption growth volatility.

Recall that in chapter (ref), we started with a consumption-based model, and related the slope of the mean-variance frontier to the volatility of the discount factor. Reviewing the logic,

$$0 = E(mR^e) = E(m)E(R^e) + \rho\sigma(m)\sigma(R^e).$$

implies

$$\sigma(m) = E(m)\frac{1}{(-\rho)}\frac{E(R^e)}{\sigma(R^e)}.$$

Correlation coefficients must be less than one, so *any discount factor m that prices the excess return R^e must have standard deviation*

$$\frac{\sigma(m)}{E(m)} \geq \frac{E(R^e)}{\sigma(R^e)}.$$

In chapter (ref) we took the *discount factor* as given, and used this equation to characterize the mean-variance frontier. Here, we use the opposite interpretation. Given the *Sharpe ratio* of assets, what do we learn about *discount factors* that might price them? The answer is a restriction on their means and variances. As graphed in figure (ref), $\sigma(m)$ must lie above a line with slope $E(R^e)/\sigma(R^e)$. The latter is the slope of the mean-variance frontier or Sharpe ratio.

Numbers. The essence of the Hansen-Jagannathan distillation of the equity premium puzzle is straightforward now. The postwar mean value weighted NYSE is about 8% per year over the T-bill rate, with a standard deviation of about 16%. Thus, $E(R^e)/\sigma(R^e)$ is about 0.5 in annual data, or 0.25 in quarterly data. (Standard deviations scale with the square root of the horizon.) If there was a constant risk free rate, $E(m) = 1/R^f$ would nail down $E(m)$. The t-bill rate is not very risky, so $E(m)$ is not far from the inverse of the mean T-bill rate, or about 0.99. Thus, these basic facts about the mean and variance of stocks and bonds imply $\sigma(m) > 0.5 = 50\%$, or 25%, in quarterly data.

In the standard consumption-based model, $m_t = \beta(c_t/c_{t-1})^{-\gamma}$. Per capita consumption growth has standard deviation about 1% per year. With log utility, that implies $\sigma(m) = 0.01 = 1\%$ which off by a factor of 50!. Raising the risk aversion coefficient helps. But, to first order $\sigma[(c_t/c_{t-1})^{-\gamma}] = \gamma\sigma[c_t/c_{t-1}]$ so huge risk aversion coefficients are required.

The difference between the bound and log utility $\sigma(m)$ poses a huge challenge for the consumption-based model. Mismeasurement of consumption data or lack of aggregation due to uninsurable individual risk are often mentioned as possible solutions to asset pricing puzzles, but we can see they won't help here. It's not credible that perfectly measured aggregate or individual consumption growth varies by *50% per year!* (25% per quarter, or $50/\sqrt{12} \simeq 15\% =$ per month.) Mine doesn't. Does yours?

Retreating to the CAPM or other models really doesn't help, either. For example, the best derivation of the CAPM starts with the consumption-based model and log utility. The log utility consumption-based model is in there! Most implementations of the CAPM take the market premium as given; but to believe the market Sharpe ratio of 0.5 and the CAPM, you have to believe that properly measured consumption growth has a 50% per year standard deviation!

I now digress into better ways of making the calculation. Then I return to the numbers and extensions of the calculation.

24.2 Many returns–formulas

Technically, it's clear we want to calculate a bound on $\{\sigma(m), E(m)\}$ using a vector of returns rather than a single return. I present several ways to make this calculation. I return to the results later.

24.2.1 A quick argument.

A quick derivation of the Hansen-Jagannathan bound with no restriction $m \geq 0$,

$$\sigma^2(m) \geq (\mathbf{p} - E(m)E(\mathbf{x}))\,\Sigma^{-1}\,(\mathbf{p} - E(m)E(\mathbf{x}))\,.$$

Hansen and Jagannathan give a quick regression derivation. Take any valid m, i.e. an m such that $\mathbf{p} = E(m\mathbf{x})$. Think of running a regression of m on the asset payoffs in question,

$$m_t = E(m) + (\mathbf{x}_t - E(\mathbf{x}))'\boldsymbol{\beta} + \epsilon_t$$

We can infer the regression coefficient $\boldsymbol{\beta}$ by the requirement that m correctly price the assets.

$$\mathbf{p} = E(m\mathbf{x}) = E(m)E(\mathbf{x}) + \Sigma\boldsymbol{\beta}$$

where

$$\sum = cov(\mathbf{x}, \mathbf{x}')$$

and, by definition of a regression, $E(\varepsilon_t \mathbf{x}) = E(\varepsilon_t) = 0$. Solving,

$$\boldsymbol{\beta} = \Sigma^{-1} \left(\mathbf{p} - E(m)E(\mathbf{x}) \right).$$

Again, the latter equality holds because m must price the assets.

Now, regression residuals are uncorrelated with right hand variables, by construction. Thus,

$$\sigma^2(m) = \sigma^2 \left[(\mathbf{x}_t - E(\mathbf{x}))'\boldsymbol{\beta} \right] + \sigma^2(\epsilon)$$

and finally, the *Hansen-Jagannathan Bound:*

$$\sigma^2(m) \geq \left(\mathbf{p} - E(m)E(\mathbf{x}) \right) \Sigma^{-1} \left(\mathbf{p} - E(m)E(\mathbf{x}) \right). \tag{172}$$

This is a parabolic region in $\left\{ E(m), \sigma^2(m) \right\}$ space, or a hyperbola in $\{E(m), \sigma(m)\}$ as illustrated in Figure (ref)

24.2.2 A projection argument.

The mean-variance frontier of discount factors can be characterized analogously to the mean-variance frontier of asset returns,

$$m = x^* + we^* + n$$

The projection or regression of m onto asset payoffs ought to remind you of the geometric arguments used in Chapter (ref) to discuss arbitrage and mean-variance frontiers. In fact, same geometry and arguments generate the mean-variance frontier of *discount factors*[9].

Recall that there is always an x^* in the payoff space $\underline{X}$ that prices all payoffs in $\underline{X}$. Furthermore, any m must be equal to x^* plus some random variable orthogonal to $\underline{X}$, and x^* is the projection of any m on the space of payoffs; $p = E(x^*x) \Leftrightarrow p = E[(x^* + \varepsilon)x]$ where $E(\varepsilon x) = 0$.

[9] As a note in the history of thought, this was the argument in the first draft of Hansen and Jagannathan's paper; the more intuitive arguements came later. This presentation can be found in Gallant Hasnen and Tauchen (1989).

We can construct a three-way orthogonal decomposition of discount factors m just as we did for returns, as illustrated in figure 25. (In fact this is the same drawing with different labels.) Any m must line in the plane marked $\underline{M}$, perpendicular to $\underline{X}$ through x^*. Any m must be of the form

$$m = x^* + we^* + n$$

e^* is defined as the residual from the projection of 1 onto $\underline{X}$ or, equivalently the projection of 1 on the space $\underline{E}$ of "excess m's", random variables of the form $m - x^*$.

$$e \equiv 1 - proj(1|\underline{X}) = proj(1|\underline{E}).$$

e^* generates means of m just as R^{e*} did for returns:

$$E(m - x^*) = E[1 \times (m - x^*)] = E[proj(1|\underline{E})(m - x^*)]$$

Finally n, defined as the leftovers, has mean zero since it's orthogonal to 1 and is orthogonal to $\underline{X}$. As with returns, then, the mean-variance frontier of m's is given by

$$m^* = x^* + we.$$

(If all this seems a bit rushed, go back to Chapter (ref). It is *exactly* the same argument.)

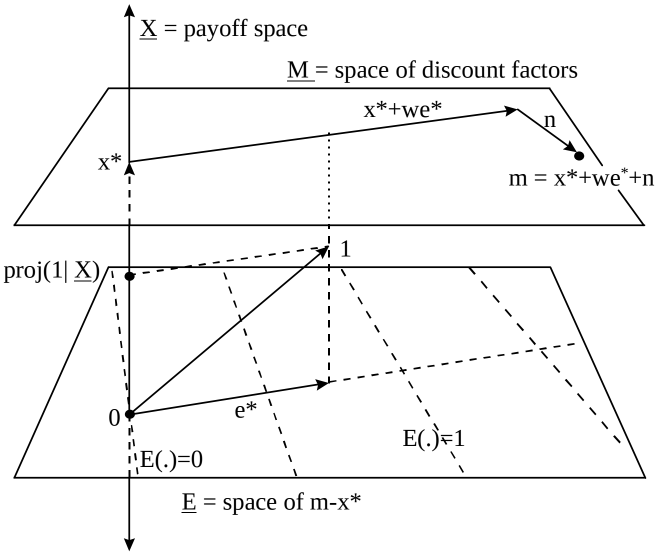


Figure 25. Decomposition of any discount factor $m = x^* + we + n$.

This construction can be used to derive a formula for the Hansen-Jagannathan bound for the finite-dimensional cases discussed above. It's more general, since it can be used in infinite-dimensional payoff spaces as well. It extends to conditional vs. unconditional bounds in the same way, and more easily than the regression approach given above.

Now, to give equations for the construction and find the mean-variance frontier of discount factors. We find x^* as before, it is the portfolio $\mathbf{c}'\mathbf{x}$ in $\underline{X}$ that prices $\mathbf{x}$ (see (ref)):

$$x^* = \mathbf{p}'\mathbf{E}(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}.$$

Similarly, let's find e^*. Using the standard OLS formula and remembering that $E(\mathbf{x}1) = E(\mathbf{x})$, the projection of 1 on $\underline{X}$ is

$$proj(1|\underline{X}) = E(\mathbf{x})'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}.$$

(After a while you get used to the idea of running regressions with 1 on the left hand side and random variables on the right hand side!) Thus,

$$e^* = 1 - E(\mathbf{x})'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}.$$

Again, you can construct time-series of $x^*$and e^* from these definitions.

Finally, we now can construct our variance minimizing discount factors

$$m^* = x^* + we^* = \mathbf{p}'\mathbf{E}(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x} + w\left[1 - E(\mathbf{x})'E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}\right]$$

or

$$m^* = w + \left[\mathbf{p} - wE(\mathbf{x})\right]' E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x} \tag{173}$$

As w varies, we trace out discount factors m^*on the frontier with varying means and variances. It's easiest to find mean and second moment:

$$E(m^*) = w + \left[\mathbf{p} - wE(\mathbf{x})\right]' E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x})$$

$$E(m^{*2}) = \left[\mathbf{p} - wE(\mathbf{x})\right]' E(\mathbf{x}\mathbf{x}')^{-1}\ \left[\mathbf{p} - wE(\mathbf{x})\right];$$

variance follows from $\sigma^2(m) = E(m^2) - E(m)^2$. With a little algebra one can also show that these formulas are equivalent to equation (172).

What if there is a riskfree rate?

So far I have assumed that the payoff space does not include a (constant) riskfree rate. What if it does? The answer is, of course, that $R^f = 1/E(m)$ nails down the mean discount factor, so the "cup" reduces to a vertical line. In this case, the mean-variance frontier of discount factors is the *point* x^* alone. If this isn't clear from the picture, we can characterize any discount factor algebraically as $m = x^* + \varepsilon$ with $E(x\varepsilon) = 0$. With $1 \in \underline{X}$, $E(x\varepsilon) = 0$ implies $E(\varepsilon) = 0$. Thus, *any* m must have the same $E(m) = E(x^*)$. The minimum *second moment* discount factor, x^*, is then also the minimum *variance* discount factor.

This observation leads to another way of thinking about the Hansen-Jagannathan frontier for payoff spaces that do *not* contain a unit payoff. Add a unit payoff, supposing that its price is a prespecified value of $E(m)$, then find x^* and take its variance. Note that all the above formulas for Hansen-Jagannathan minimizers m^* consist of some unit payoff and some combination of the original asset returns. A good exercise is to go through the construction of x^* in augmented payoff spaces and show that you get the same answer

24.2.3 Brute force.

You can obtain the same result with a brute force minimization, picking m state-by-state or date-by-date to minimize variance.

A brute force approach to the bound is also useful. It is a technique that is easily adaptable to finding more interesting bounds on fancier moments, and bounds with frictions. It also shows graphically how we are constructing a *stochastic process* for the discount factor Finally, though the derivations given below are much more elegant and short, it's hard to see how one would ever have thought of them. It's comforting to see that one can get directly and constructively to the same answer. (In fact, I know that at least a few of the bounds in Cochrane and Hansen (199x) were first derived this way, and then presented with more beautiful arguments like the above!)

By brute force, I mean, solve the problem

$$\min_{\{m\}} var(m) \text{ given } E(m),\ \mathbf{p} = E(m\mathbf{x}).$$

where $\mathbf{x}$ is a vector of asset payoffs with price $\mathbf{p}$. We need to pick the *random variable* m, state-by-state or date-by-date. Since the mean is held fixed, we can minimize second moment as well as variance, or

$$\min_{\{m_t\}} \frac{1}{T}\sum_{t=1}^{T} m_t^2 \text{ given } \frac{1}{T}\sum_{t=1}^{T} m_t,\ \mathbf{p} = \frac{1}{T}\sum_{t=1}^{T} m_t\mathbf{x}_t.$$

If you prefer, you can get the same answer starting with

$$\min_{\{m(s)\}} \sum_s \pi(s)m(s)^2 \text{ given } \sum_s \pi(s)m(s),\ \mathbf{p} = \sum_s \pi(s)m(s)\mathbf{x}(s).$$

Introduce a Lagrange multiplier 2λ on the first constraint and $2\boldsymbol{\delta}$ on the second. Then the first order conditions are

$$\frac{\partial}{\partial m_t} : m_t^* = \lambda + \boldsymbol{\delta}'\mathbf{x}_t$$

Thus, *the variance minimizing discount factor m^* is a combination of a constant and a linear combination of $\mathbf{x}_t$.*

The next step (as in any Lagrangian minimization) is to determine λ and $\boldsymbol{\delta}$ to satisfy the constraints. It is more convenient to reparameterize the variance minimizing discount factor

$$m_t^* = E(m) + (\mathbf{x}_t - E(\mathbf{x}))'\boldsymbol{\beta}.$$

This is still some combination of a constant and a payoff of $\mathbf{x}'s$ so nothing has changed. Now we are back where we started with the regression derivation. (If you write $m_t^* = \lambda + \boldsymbol{\delta}'\mathbf{x}_t$ you are at the same point in the mean-variance characterization, and will get the same answer as equation (173).) The $E(m)$ constraint is obviously satisfied, so we only have to pick $\boldsymbol{\beta}$ to satisfy the pricing constraint

$$E(\mathbf{p}) = E\left[E(m)\mathbf{x}_t + \mathbf{x}_t(\mathbf{x}_t - E(\mathbf{x}))'\boldsymbol{\beta}\right] = E(m)E(\mathbf{x}) + \Sigma\boldsymbol{\beta}$$

where Σ is the variance-covariance matrix of the payoffs $\mathbf{x}$. Thus,

$$\boldsymbol{\beta} = \Sigma^{-1}\left[\mathbf{p} - E(m)E(\mathbf{x})\right]$$

and, finally,

$$m_t^* = E(m) + \left[\mathbf{x}_t - E(\mathbf{x})\right]'\Sigma^{-1}\left[\mathbf{p} - E(m)E(\mathbf{x})\right] \tag{174}$$

$$\sigma^2(m^*) = \left[\mathbf{p} - E(m)E(\mathbf{x})\right]'\Sigma^{-1}\left[\mathbf{p} - E(m)E(\mathbf{x})\right]$$

just as before.

Equation (174) is useful. You can use it to *plot* the time series of the variance-minimizing discount factor and see what it looks like.

24.2.4 Sharpe ratio intuition

How to connect the Hansen-Jagannathan bound to the Mean-variance frontier.

In a single excess return case, we found

$$\frac{\sigma(m)}{E(m)} \geq \frac{E(R^e)}{\sigma(R^e)}.$$

This suggests a graphical way to find a Hansen-Jagannathan bound with many assets: For any hypothetical risk-free rate, find the highest Sharpe ratio. That is, of course the tangency portfolio. Then the slope to the tangency portfolio gives the ratio $\sigma(m)/E(m)$. Figure 26 illustrates.

As we sweep through values of $E(m)$, the slope to the tangency becomes lower, and the Hansen-Jagannathan bound declines. At the mean return corresponding to the minimum

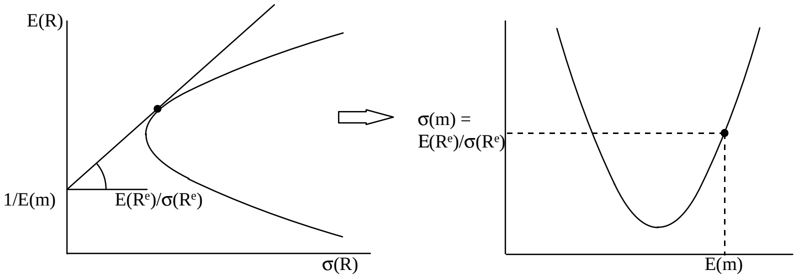


Figure 26. Graphical construction of the Hansen-Jagannathan bound.

variance point, the HJ bound attains its minimum. Continuing, the Sharpe ratio rises again and so does the bound.

This fact emphasizes the equivalence between Hansen-Jagannathan frontiers and mean-variance frontiers. For example, an obvious exercise is to see how much the addition of assets raises the Hansen-Jagannathan bound. This is *exactly* the same as asking how much those assets expand the mean-variance frontier. It was, in fact, this link between Hansen-Jagannathan bounds and mean-variance frontiers rather than the logic I described that inspired Knez and Chen (1996) and DeSantis (1994) to test for mean-variance efficiency using, essentially, Hansen-Jagannathan bounds.

24.2.5 A beautiful inequality

Another way of stating the relation between bounds and Sharpe ratios is the following:

$$\min_{\{\text{all } m \text{ that price } \mathbf{x}\in\underline{X}\}} \frac{\sigma(m)}{E(m)} = \max_{\{\text{all excess returns } R^e \text{ in } \underline{X}\}} \frac{E(R^e)}{\sigma(R^e)}. \tag{175}$$

24.2.6 Positive Discount factors

The discount factors produced by consumption based models are always positive, while the discount factors produced by the Hansen-Jagannathan procedure can be negative at times. (x^* is one of them.) We will get a tighter bound if we construct the space of all *positive* discount factors.

More generally, this is the main advantage of using the discount factor language rather than expected return and mean-variance frontier language. In that language, it is very hard to incorporate positivity and arbitrage, while it is seamless in discount factor language.

Direct approach

To derive this bound, try the brute force method.

$$\min\ E(m^2) \text{ given } E(m),\ \mathbf{p} = E(m\mathbf{x}),\ m \geq 0.$$

$$\min\ \frac{1}{T}\sum_t m_t^2 \text{ given } \frac{1}{T}\sum_t m_t,\ \mathbf{p} = \frac{1}{T}\sum_t m_t\mathbf{x}_t,\ m_t \geq 0.$$

Denote the Lagrange multipliers on the first two constraints 2λ and $2\boldsymbol{\delta}$, and the Kuhn-Tucker multiplier on the last constraint $2\eta_t$. The first order conditions are

$$\frac{\partial}{\partial m_t} :\ m_t^* = \lambda + \boldsymbol{\delta}'\mathbf{x}_t + \eta_t\ ;\ \begin{cases} \eta_t > 0 \text{ if } m_t = 0 \\ \eta_t = 0 \text{ if } m_t > 0 \end{cases}$$

Equivalently,

$$m^* = \max(\lambda + \boldsymbol{\delta}'\mathbf{x}, 0) \equiv \left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^+.$$

The last equality defines the $[\cdot]^+$ notation for the truncation of a random variable.

This formula has a nice interpretation: m_t^* is a *call option with zero strike price* on a portfolio of payoffs x augmented by a constant.

Note that if the previous formula happened to produce a positive m^*, then it solves this problem as well. If it did not, then we now get a tighter bound.

As before, we have to determine λ and $\boldsymbol{\delta}$ in order to satisfy the mean and pricing constraints. Since the problem is no longer linear, this must be done numerically. One approach is obvious: hand the two equations in two unknowns

$$E\left(\left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^+\right) = E(m)$$

$$E\left(\left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^+ \ \mathbf{x}\right) = \mathbf{p}$$

to a nonlinear equation solver. Good starting values are the solutions to the HJ bound without positivity, or the solution at a nearby $E(m)$, since you will typically do this for a grid of $E(m)'s$.

An approach using a minimization.

Nonlinear equations solvers often get stuck, so a second approach that uses straight minimization often works better. Remember that the variance minimizing m at a given $E(m)$ can be found by augmenting the payoff space with a unit payoff, supposing its price to be $E(m)$, and finding x^*. The same approach adapts easily to finding the bound with positivity, and yields a numerically more stable procedure.

Recall that $x^* = R^*/E(R^{*2})$ where R^* is the minimum second moment *return*. We know that the non-negative variance minimizing discount factor will be a truncated payoff of the from $\max(\lambda + \boldsymbol{\delta}'\mathbf{x}_t, 0)$. Thus, consider the payoff space composed of the original payoffs, the unit payoff with price $E(m)$ and truncations of these. All we have to do is find the minimum second moment return in this payoff space, find x^*, and take its variance.

To be specific, suppose we start with two returns, R^a and R^b. The return on the unit payoff is $1/E(m)$. Then, R^* solves

$$E(R^{*2}) = \min_{\{c^a, c^b\}} E\left[\max\left(c^a R^a + c^b R^b + (1 - c^a - c^b)\frac{1}{E(m)},\ 0\right)^2\right]$$

at the optimal choice of c^a, c^b. You still have to use a search routine to find c^a and c^b, but you are searching for the minimum of a “quadratic” function, which is a lot easier numerically than solving systems of equations. Finally, you construct $\sigma^2(x^{*2}) = E(x^{*2}) - E(x^*)^2 = 1/E(R^{*2}) - E(m)^2$. Do all this for a range of $E(m)$, and you trace out the Hansen-Jagannathan bound with positivity.

A duality approach.

Here's another way to calculate the bound. Write the original problem in Lagrangian form as

$$\min_{\{m\geq 0\}} \max_{\{\lambda,\boldsymbol{\delta}\}} E(m^2) - 2\lambda \left[E(m) - \mu\right] - 2\boldsymbol{\delta}' \left[E(m\mathbf{x}) - \mathbf{p}\right].$$

We can interchange $\min$ and max yielding

$$\max_{\{\lambda,\boldsymbol{\delta}\}} \min_{\{m\geq 0\}} E(m^2) - 2\lambda \left[E(m) - \mu\right] - 2\boldsymbol{\delta}' \left[E(m\mathbf{x}) - \mathbf{p}\right]$$

Do the inner minimization: For given λ and $\boldsymbol{\delta}$ state-by-state minimization gives

$$m^* = \left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^+.$$

Now the problem is simply

$$\max_{\{\lambda,\boldsymbol{\delta}\}} E\left[\left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^{+2}\right] - \lambda \left[E\left(\left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^{+}\right) - \mu\right] - \boldsymbol{\delta}' \left[E\left(\left[\lambda + \boldsymbol{\delta}'\mathbf{x}\right]^{+}\mathbf{x}\right) - \mathbf{p}\right]$$

While our original minimization problem $\min_{\{m\}}$ meant choosing m in every state (or date, in a finite sample), this conjugate or dual minimization only chooses one λ and $\boldsymbol{\delta}$, a vector with as many elements as payoffs. It is quite straightforward to hand this problem directly to a numerical maximizer.

Arbitrage bounds

While the original Hansen-Jagannathan bound is parabolic, the bound with positivity rises to infinity at finite values of $E(m)$. Thus, it appears that there are values of $E(m)$ for which we cannot construct any positive discount factor. It must be that they imply an arbitrage opportunity.

We can calculate the arbitrage bounds on $E(m)$ directly. Suppose a payoff x is always greater than 1, $x > 1$ Then price of this payoff must satisfy $p\left(x\right) > p(1) = E(m)$ by absence of arbitrage. Interpreting the equation backwards, we learn $E(m) < p(x)$. (Or, the minimum variance of a positive discount factor with mean greater than $p(x)$ is infinite.)

You will generally be able to construct such a payoff in a finite sample. For example, 10 times the T-bill return will undoubtedly be greater than 1 at every data point. (These are *payoffs*, not *portfolios* — the weights do not have to sum to one.) Then, we know that $E(m) < 10$.

If 1 was in the payoff space (if there was a real risk free rate) then we would know $E(m)$ exactly. If the payoff space does not contain 1, it may well contain a payoff (payoff, not return) that is always *greater* or *smaller* than one. The price of these payoffs gives definite limits on the range of $E(m)$.

Finding the arbitrage bounds is straightforward. We want the payoff greater than one with *smallest* price. So search for it. In our two-return example payoffs are combinations $c^a R^a + c^b R^b$, and have price $c^a + c^b$. Thus, you want to find

$$\min_{\{c^a,c^b\}} \left(c^a + c^b\right) \ s.t.\ c^a R_t^a + c^b R_t^b \geq 1 \text{ for all t.}$$

Again, we do *not* impose $c^a + c^b = 1$; these are payoffs, not necessarily returns. Similarly, you can find the lower arbitrage bound by finding the payoff with largest price that is always less than one.

You should already be concerned about the finite-sample performance of this procedure. For example, suppose returns are lognormally distributed. Then, any return will eventually be arbitrarily close to zero and arbitrarily large, so there is no arbitrage bound using the population moments. But of course, any sample will feature a minimum and a maximum return, so that it will look like there is an arbitrage bound in any sample.

24.3 Results

24.4 Beyond mean and variance.

The point of the bound is to characterize discount factors in ways that will be useful to the construction of asset pricing models. By now, it is well known that discount factors must be quite volatile. But what *other* moments must a successful discount factor posses? Work is just beginning on this question. Here I report two calculations, taken from Cochrane and Hansen (1992).

24.4.1 Correlation Puzzle

Go back to the derivation simplest one excess return bound, we manipulated $0 = E(mr^e)$ to obtain

$$\sigma(m) = -\frac{E(m)E(R^e)}{\rho\sigma(R^e)}. \tag{176}$$

At this point, we noted that $|\rho| < 1$, yielding the HJ bound. This means that the m on the bound is *perfectly* correlated with the excess return. More generally, Hansen-Jagannathan minimizers are of the form $m = v+R'\boldsymbol{\beta}$ and so are perfectly correlated with some portfolio of asset returns..

This problem is more general. The stock returns in the Mehra-Prescott model are also nearly perfectly correlated with consumption growth. Most general equilibrium models also feature consumption growth highly or perfectly correlated with stock returns. At an elemental level, most general equilibrium models use only one shock, so there is a sense in which *all* time series are perfectly correlated (stochastically singular).

But consumption growth is *not* highly correlated with returns. In quarterly data, the correlation of consumption growth with the VW excess return is about 0.2. Using equation (176), this observation raises the required variance of m by a factor of 5, to about 1.0 or 100% per quarter!

Furthermore, it's easy to modify a discount factor model to give you lots of variance: Just add an i.i.d. error. $E\left[(m+\epsilon)R\right] = E(mR)$ so this has no effect whatsoever on the pricing predictions; but obviously gives a much bigger variance.

Finally, correlation is what asset pricing is fundamentally all about. $E(R^e) = -cov(m, R^e)/E(m)$ so an m only explains expected return variation if it is correlated with returns. We don't want to produce models with variable discount factors uncorrelated with returns!

A simple bound

Based on equation (176) we could thus generalize the previous bound to a minimum variance of m given the mean of m *and* a correlation of m and excess returns not to exceed a given value ρ. This calculation yields bounds proportionately higher than the original bound.

A multivariate version

One can do the same thing in a multivariate context. Think of running any m on the set of asset returns under consideration and a constant,

$$m_t = E(m) + [\mathbf{x}_t - E(\mathbf{x})]\boldsymbol{\beta} + \epsilon_t$$

We know what $[\mathbf{x}_t - E(\mathbf{x})]\boldsymbol{\beta}$ should be; that's the HJ minimizer m^* at this value of $E(m)$. Thus,

$$\sigma^2(m) = \sigma^2(m^*) + \sigma^2(\epsilon).$$

The R^2 of the regression of m on $\mathbf{x}$ is

$$R^2 = \sigma^2(m^*)/\sigma^2(m)$$

Thus, for any discount factor,

$$\sigma^2(m) = \sigma^2(m^*)/R^2.$$

Here's what it means. Take the Hansen-Jagannathan bound, and divide by a value of R^2, say 0.2. This is the minimum variance of all $m's$ with the given mean $E(m)$ and that, when regressed on asset returns, have an R^2 no less than 0.2. Formally, denote by $R^{\max 2}$ the desired upper bound of the R^2 of the discount factor on asset payoffs. The problem

$$\min \sigma^2(m) \text{ s.t.} \mathbf{p} = E(m\mathbf{x}), R^2 \leq R^{\max 2}$$

has solution

$$\sigma^2(m) = \sigma^2(m^*)/R^2.$$

Obviously, you get a series of higher and higher bounds, as R^2 is lower and lower.

Changing the m rather than the bound.

Instead of changing the *bound*, we can achieve the same thing by changing the *candidates*. Consider the problem

$$\min \sigma^2(proj(m|\underline{X})) \text{ s.t.} \mathbf{p} = E(m\mathbf{x})$$

Since $m^* = proj(m|\underline{X}, 1)$, we have the answer to this problem,

$$\sigma^2(proj(m|\underline{X})) = \sigma^2(m^*)$$

The Hansen-Jagannathan bound doesn't only apply to the actual discount factor; it applies just as well to the *projection* of the discount factor on the space of asset payoffs (with a constant). If a discount factor model has lots of variance, but is uncorrelated with asset payoffs, it may fit in the HJ bound, but it will utterly fail this test, revealing (or diagnosing) its problem: lack of *correlation* with asset returns.

Conditional mean vs. variance

24.5 What do we know about discount factors: a summary

24.6 Comments on the Hansen-Jagannathan bound.

It's fairly easy to produce a discount factor that has a lot of *variance*. Given a miserable candidate, m, then $m + \epsilon$, $E(\epsilon x) = 0$ prices assets just as badly as m but has a lot more variance!

Thus, the HJ bound seems most useful for evaluating models such as the standard consumption-based model that produce a unique series for *the* discount factor m. One evaluates habit persistence with pricing objects in the class $m+\epsilon$ that may have a lot more variance than the true m. The HJ bound is basically useless for evaluating factor models; these make no pretence at being *the* m.

SECTION 24.6 COMMENTS ON THE HANSEN-JAGANNATHAN BOUND.

The most lasting impact of the HJ bound may come from its impact on other questions in Finance.

We can exploit HJ bounds to ask questions about mean-variance frontiers, and apply the above GMM testing methodology to those questions. It's natural to ask whether adding an asset return or group of asset returns makes bounds go up. This is *exactly* the same question as asking whether the addition of a return makes the mean-variance frontier expand. Snow (1991) uses this idea to test whether the addition of small firm returns expands the mean-variance frontier beyond what is available using large firm returns. This is a nice test of the "small firm effect." DeSantis (1993) uses the same idea to test whether one can really expand the mean-variance frontier by international diversification. Like adding domestic returns, *ex-post* mean variance frontier can enlarge a great deal by adding assets, but this is probably spurious. By *testing*, we can see if the *ex-ante* mean variance frontier is enlarged by adding some asset returns.. This is an important test of international diversification.

More generally, the Hansen-Jagannathan methodology is the inspiration for testing factor pricing models by testing models of the form $m = a + \mathbf{b}'\mathbf{f}$. These models look just like Hansen-Jagannathan candidate discount factors, and they are. By testing such a model, we are testing whether the factors $\mathbf{f}$ span the mean-variance frontier of the assets.

Chapter 25. Investment, q and asset pricing

PART V
Appendix

Chapter 26. Notation

I use bondface to distinguish vectors from scalars. For example x is a scalar, but $\mathbf{x}$ is a vector. I use capital letters for matrices, though not all capital letters are matrices (e.g. R). A partial list of frequently-used symbols:

$x =$ payoff

$p,\ p(x) =$ price, price of payoff x

$R = x/p =$ gross return

$r = R - 1$ or $\ln(R) =$ net or log return

$R^e =$ excess return

$R^i =$ notation to indicate one among many asset returns. $i = 1, 2, ..N$ is implicit.

$m =$ discount factor, $p = E(mx)$

$u(c),\ u'(c) =$ utility, marginal utility

$\underline{X} =$ space of all payoffs $x \in \underline{X}$

$x^* =$ payoff that acts as discount factor, $p = E(x^*x)$

$R^* =$ return that acts as discount factor $R^* = x^*/p(x^*)$

β subjective discount factor $u(c_t) + \beta u(c_{t+1})$ and to denote regression coefficient

$\beta, \beta_{y,x}$ regression coefficients.

$\lambda =$ factor risk premium in beta pricing model

$f_t =$ factor e.g., market return for the CAPM

$z_t,\ dz_t =$ standard Brownian motion

Returns I use capital R to denote a *gross return,* e.g.

$$R = \frac{\$\text{back}}{\$\text{paid}}.$$

For a stock that pays a dividend D, the gross return is

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} = \frac{\text{\$back}_{t+1}}{\text{\$paid}_t} \quad \text{(for example, } 1.10\text{)}$$

R is a number like 1.10 for a 10% return.

Several other units for returns are convenient. The *net* return is

$$r_{t+1} = R_{t+1} - 1 \text{ (For example, } 0.10).$$

The *percent* return is

$$100 \times r_{t+1} \text{ (For example, } 10\%)$$

The *log* or *continuously compounded* return is

$$r_t = \ln R_t \text{ (For example, } \ln(1.10) = 0.09531 \text{ or } 9.531\%)$$

The *real return* corrects for inflation,

$$R^{\text{real}}_{t+1} = \frac{\text{Goods back}_{t+1}}{\text{Goods paid}_t}.$$

The consumer price index is defined as

$$CPI_t \equiv \frac{\$_t}{\text{Goods}_t}; \;\; \Pi_{t+1} \equiv \frac{CPI_{t+1}}{CPI_t}$$

Thus, we can use CPI data to find real returns as follows.

$$R^{\text{real}}_{t+1} = \frac{\$_{t+1} \times \frac{\text{Goods}_{t+1}}{\$_{t+1}}}{\$_t \times \frac{\text{Goods}_t}{\$_t}} = \frac{\$_{t+1} \frac{1}{CPI_{t+1}}}{\$_t \frac{1}{CPI_t}} = R^{\text{nomial}}_{t+1} \frac{CPI_t}{CPI_{t+1}} = \frac{R^{\text{nominal}}_{t+1}}{\Pi_{t+1}}.$$

I.e., divide the gross nominal return by the gross inflation rate to get the gross real return.

You're probably used to *subtracting* inflation from nominal returns. This is exactly true for log returns. Since

$$\ln(A/B) = \ln A - \ln B,$$

we have

$$\ln R^{\text{real}}_{t+1} = \ln R^{\text{nomial}}_{t+1} - \ln \Pi_{t+1}.$$

For example, 10%-5% = 5%. It is approximately true that you can subtract net returns this way,

$$\frac{R_{t+1}^{\text{nominal}}}{\Pi_{t+1}} = \frac{\left(1+r^{nomial}\right)}{1+\pi} \approx 1 + r^{nom} - \pi.$$

The approximation is ok for low inflation (10%) or less, but really bad for 100% or more inflation.

Using the same idea as for real returns, you can find dollar returns of international securities. Suppose you have a German security, that pays a gross Deutchmark return

$$R_{t+1}^{DM} = \frac{DM \text{ back}_{t+1}}{DM \text{ paid}_t}$$

Then change the units to dollar returns just like you did for real returns. The exchange rate is defined as

$$e_t^{\$/DM} = \frac{\$_t}{DM_t}.$$

Thus,

$$R_{t+1}^{\$} = \frac{\$_{t+1}}{\$_t} = \frac{DM_{t+1}}{DM_t} \times \frac{\$_{t+1}/DM_{t+1}}{\$_t/DM_t} = R_{t+1}^{DM} \times \frac{e_{t+1}^{\$/DM}}{e_t^{\$/DM}}.$$

Compound returns Suppose you hold an instrument that pays 10% per year for 10 years. What do you get for a $1 investment? The answer is not $2, since you get "interest on the interest." The right answer is the *compound return.* Denote

$$V_t = \text{value at time t}$$

Then

$$V_1 = RV_0 = (1+r)\,V_0$$

$$V_2 = R \times (RV_0) = R^2 V_0$$

$$V_T = R^T V_0$$

Thus, R^T is the *compound return.*

As you can see, it's not obvious what the answer to 10 years at 10% is. Here is why log returns are so convenient. Logs have the property that

$$\ln(ab) = \ln a + \ln b;\ \ln\left(a^2\right) = 2\ln a.$$

Thus

$$\ln V_1 = \ln R + \ln V_0$$

$$\ln V_T = T \ln R + \ln V_0$$

Thus the compound *log* return is T times the one-period log return.

More generally, log returns are really handy for multi-period problems. The T period return is

$$R_1 R_2 ... R_T$$

while the T period log return is

$$\ln(R_1 R_2 ... R_T) = \ln(R_1) + \ln(R_2) + ... \ln(R_T)$$

Within period compounding This is best explained by example. Suppose a bond that pays 10% is compounded semiannually, i.e. two payments of 5% are made at 6 month intervals. Then the total annual gross return is

$$\text{compounded semi-annually:} \quad (1.05)(1.05) = 1.1025 = 10.25\%$$

What if it is compounded quarterly? Then you get

$$\text{compounded quarterly} \quad (1.025)^4 = 1.1038 = 10.38\%$$

Continuing this way,

$$\text{compounded N times: } \left(1 + \frac{r}{N}\right)^N$$

What if you go all the way and compound *continuously?* Then you get

$$\lim_{N\to\infty} \left(1 + \frac{r}{N}\right)^N = 1 + r + \frac{1}{2}r^2 + \frac{1}{3 \times 2}r^3 ... = e^r.$$

Well, if the gross return $R = e^r$, then $r = \ln R$. For example a stated rate of 10%, continuously compounded is really a gross return of $e^{0.10} = 1.1051709 = 10.517\%$.

Both kinds of compounding If you really want a headache, what is the two year return of a security that pays a stated rate R, compunded semiannually? Well, again with $r = R - 1$, it must be

$$\left(1 + \frac{r}{2}\right)^2.$$

Similarly, the continuously compounded T year return is

$$e^{rT}.$$

Chapter 27. Utility functions

The standard representation of investor preferences or *utility* is

$$E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}).$$

This maps a consumption stream into "utility" or "happiness." The *period utility function* $u(c)$ is an increasing function – more consumption makes you happier– but concave – the extra dollar of consumption increases happiness less and less the more you have.

Standard functional forms: *Power* or *constant relative risk aversion* utility is

$$u(c) = \frac{c^{1-\gamma}}{1-\gamma}$$

a special case when $\gamma = 1$ is *log utility*

$$u(c) = \ln(c).$$

We sometimes use *quadratic utility*

$$u(c) = -\frac{1}{2}(c - c^*)$$

it's convenient for solving problems, but obviously limited to $c < c^*$. It has the unattractive property that you get more risk averse as consumption rises.

This *concavity* of the utility function also generates *risk aversion.* For example if there are two possible events, $a =$ win \$100 and $b =$ lose \$100 then

$$U(\text{bet}) = \pi_a u(c_a) + \pi_b u(c_b) = E\left[u(x)\right].$$

the concavity of utility means *risk aversion*, people will pay to avoid fair bets. In equations,

$$E[u(\text{bet})] = \pi_a u(c_a) + \pi_b u(c_b) \leq u(\pi_a c_a + \pi_b c_b) = u[E(\text{bet})]$$

Risk aversion means that an equal chance of getting or losing \$100 (say) of consumption makes you worse off. If utility is flat (linear), the investor is *risk-neutral* and indifferent to such a bet. The utility of such a *bet* is expected *utility*, not utility of *expected consumption*.

The sum part of the utility function captures the effect of time. You prefer consumption today to consumption 10 years from now. To capture this, the momentary utility $u(c)$ is multiplied by a *subjective discount factor* β, somewhat less than one.

Chapter 28. Probability and statistics

28.1 Probability

28.1.1 Random variables

We model returns as *random variables.* A random variable can take on one of many values, with an associated probability. For example, the gross return on a stock might be one of four values.

	Value	Probability
	1.1	$1/5$
$R =$	1.05	$1/5$
	1.00	$2/5$
	0.00	$1/5$

Each value is a possible *realization* of the random variable. Of course, stock returns can typically take on a much wider range of values, but the idea is the same. Many finance texts distinguish the *random variable* from its *realization* by using $\tilde{R}$ for the random variable and R for the realization. I don't.

The *distribution* of the random variable is a listing of the values it can take on along with their probabilities. For example, the distribution of return in the above example is

(Real statisticians call this the *density* and reserve the word *distribution* for the *cumulative distribution*, a plot of values vs. the probability that the random variable is at or below that value.)

A deeper way to think of a random variable is a *function.* It maps "states of the world" into real numbers. The above example might really be

	Value	State of the world	Probability
	1.1	New product works, competitor burns down	$1/5$
$R =$	1.05	New product works, competitor ok.	$1/5$
	1.00	Only old products work.	$2/5$
	0.00	Factory burns down, no insurance.	$1/5$

The probability really describes the external events that define the state of the world. However, we usually can't name those events, so we just think about the probability that the stock return takes on various values.

In the end, all random variables have a discrete number of values, as in this example. Stock prices are only listed to 1/8 dollar, all payments are rounded to the nearest cent, computers can't distinguish numbers less than 10^{-300} or so apart. However, we often think of *continuous* random variables, that can be any real number. Corresponding to the discrete probabilities above, we now have a continuous probability *density*, usually denoted $f(R)$. The density tells you the probability per unit of R; $f(R_0)\,\Delta R$ tells you the probability that the random variable R lies between R_0 and $R_0 + \Delta R$.

A common assumption is that returns (or log returns) are *normally distributed.* This means that the density is given by a specific function,

$$f(R) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(R-\mu)^2}{2\sigma^2}\right].$$

The graph of this function looks like

About 30% (really 31.73%) of the probability of a normal distribution is more than one standard deviation from the mean and about 5% is more than two standard deviations from the mean (really 4.55%, the 5% probability line is at 1.96 standard deviations). That means that there is only one chance in 20 of seeing a value more than two standard deviations from the mean of a normal distribution. Stock returns have "fat tails" in that they are slightly more likely to take on extreme values than the normal distribution would predict.

28.1.2 Moments

Rather than plot whole distributions, we usually summarize the behavior of a random variable by a few *moments* such as the mean and variance.

I'll denote the values that R can take on as R_i with associated probabilities π_i. Then the *mean* is defined as

$$\text{Mean: } E(R) = \sum_{\text{possible values i}} \pi_i R_i.$$

The mean is a *measure of central tendency*, it tells you where R is "on average." A high mean stock return is obviously a good thing!

The *variance* is defined as

$$\text{Variance: } \sigma^2(R) = E\left[(R - E(R))^2\right] = \sum_i \pi_i \left[R_i - E(R)\right]^2$$

Since squares of negative as well as positive numbers are positive, variance tells you how far away from the mean R typically is. It measures the spread of the distribution. High variance is not a good thing; it will be one of our measures of *risk.*

The *covariance* is

$$\text{Covariance: } cov\left(R^a, R^b\right) = E\left[\left(R^a - E\left(R^a\right)\right)\left(R^b - E\left(R^b\right)\right)\right]$$

$$= \sum_i \pi_i \left[R_i^a - E\left(R^a\right)\right]\left[R_i^b - E\left(R^b\right)\right]$$

It measures the tendency of two returns to move together. It's positive if they typically move in the same direction, negative if one tends to go down when the other goes up, and zero if there is no tendency for one to be high or low when the other is high.

The size of the covariance depends on the units of measurement. For example, if we measure one return in cents, the covariance goes up by 100, even though the tendency of the two returns to move together hasn't changed. The *correlation coefficient* resolves this problem.

$$\text{Correlation: } corr\left(R^a, R^b\right) = \rho = \frac{cov\left(R^a, R^b\right)}{\sigma\left(R^a\right)\sigma\left(R^b\right)}.$$

The correlation coefficient is always between -1 and 1.

For continuously valued random variables, the sums become integrals. For example, the mean is

$$E(R) = \int R\ f(R)\, dR.$$

The normal distribution defined above has the property that the mean equals the parameter μ, and the variance equals the parameter σ^2. (To show this, you have to do the integral.)

28.1.3 Moments of combinations

We will soon have to do a lot of manipulation of random variables. For example, we soon will want to know what is the mean and standard deviation of a *portfolio* of two returns. The basic results are

1) Constants come out of expectations and expectations of sums are equal to sums of expectations. If c and d are numbers,

$$E\left(cR^a\right) = cE\left(R^a\right)$$

$$E\left(R^a + R^b\right) = E\left(R^a\right) + E\left(R^b\right)$$

or, more generally,

$$E\left(cR^a + dR^b\right) = cE\left(R^a\right) + dE\left(R^b\right).$$

2) Variance of sums works like taking a square,

$$var\left(cR^a + dR^b\right) = c^2 var\left(R^a\right) + d^2 var\left(R^b\right) + 2cd\; cov\left(R^a, R^b\right).$$

3) Covariances work linearly

$$cov\left(cR^a, dR^b\right) = cd\; cov\left(R^a, R^b\right)$$

To derive any of these or related rules, just go back to the definitions. For example,

$$E\left(cR^a\right) = \sum_i \pi_i c R_i^a = c\sum_i \pi_i R_i^a = cE\left(R^a\right).$$

28.1.4 Normal distributions.

Normal distributions have an extra property. Linear combinations of normally distributed random variables are again normally distributed. Precisely, if R^a and R^b are normally distributed, and

$$R^p = cR^a + dR^b$$

then, R^p is also normally distributed with the mean and variance given above.

28.1.5 Lognormal distributions

A variable R is *lognormally* distributed if $r \equiv \ln(R)$ is normally distributed. This is a nice model for returns since we can never see $R < 0$ and a lognormal captures that fact, where you can see $R < 0$ if it is normally distributed. Lognormal returns are like log returns, useful for handling multiperiod problems.

Since $R = e^{\ln R} = e^r$ by definition, wouldn't it be nice if $E(R) = e^{E(r)}$? Of course, that isn't true because $E[f(x)] \neq f[E(x)]$. But something close to it is true. By working out the integral definition of mean and variance, you can show that

$$E(R) = e^{E(r)+\sigma^2(r)/2}.$$

The variance is a little trickier. $R^2 = e^{2r}$ so this is also lognormally distributed. Then

$$\sigma^2(R) = E(R^2) - E(R)^2 = e^{2E(r)+2\sigma^2(R)} - e^{2E(r)+\sigma^2(R)} = e^{2E(r)+\sigma^2(R)}\left[e^{\sigma^2(R)} - 1\right].$$

As a linear combination of normals is normal, a product of lognormals (raised to powers) is lognormal. For example,

$$R_1R_2 = e^{r_1+r_2};$$

since r_1 and r_2 are normal so is $r_1 + r_2$, and therefore R_1R_2 is lognormal.

28.2 Statistics

28.2.1 Sample mean and variance

What if you don't know the probabilities? Then you have to *estimate* them from a *sample*. Similarly, if you don't know the mean, variance, regression coefficient, etc., you have to estimate them as well. That's what *statistics* is all about.

The *average* or *sample mean* is

$$\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t$$

where $\{R_0, R_1, ...R_t, ...R_T\}$ is a *sample* of data on a stock return. Just to be confusing, many people use μ for sample as well as population mean. Sometimes people use hats, $\hat{\mu}$ to distinguish estimates or sample quantities from true population quantities.

Keep the *sample mean* and the true, or *population* mean separate in your head. For example, the true probabilities that a coin will land heads or tails is 1/2, so the mean of a bet on a coin toss ($1 for heads, -$1 for tails) is 0. A *sample* of coin tosses might be {H,T,T,H,H}. In that *sample*, the frequency of heads is 3/5 and tails 2/5, so the *sample mean* of a coin toss bet is 1/5.

Obviously, as the sample gets bigger and bigger, the *sample mean* will get closer and closer to the true or *population mean*. That property of the sample mean (*consistency)* makes it a good estimator. But the sample and population mean are not the same thing for any finite sample! Also, sample means approach population means *only* if you are repeatedly doing the same thing, such as tossing the same coin. This may not be true for stocks. If there are days when expected returns are high and days when they are low, then the average return will not necessarily recover either expected return. The sample of the Peso/Dollar exchange rate was pretty useless the day before the Peso plunged.

The *sample variance* is

$$s^2 = \hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left[R_t - \bar{R}\right]^2.$$

Sample values of the other moments are defined similarly, as obvious analogs of their population definitions.

28.2.2 Variation of sample moments

The sample mean and sample variance vary from sample to sample. If I got {H,T,T,H,H}, the sample mean is 1/5, but if I happened to get {T,T,H,T,T}, the sample mean would be -4/5. Thus the sample mean, standard deviation, and other statistics are also *random variables*; they vary from sample to sample. They are random variables that depend on the whole sample, not just what happened one day, but they are random variables nonetheless. The population mean and variance, by contrast are just numbers.

We can then ask, “how much does the sample mean (or other statistic) vary from sample to sample?” This is an interesting question. If a mutual fund manager tells you “my mean return for the last five years was 20% and the S&P500 was 10%” you want to know if that was just due to chance, or means that his true, population mean, which you are likely to earn in the *next* 5 years is also 10% more than the S&P500. In other words, was the *realization* of the random variable called “my estimate of manager A’s mean return” near the mean of the true or population mean of the random variable “manager A’s return?”

Figuring out the variation of the sample mean is a good use of our formulas for means and variances of sums. The sample mean is

$$\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t.$$

Therefore,

$$E\left(\bar{R}\right) = \frac{1}{T}\sum_{t=1}^{T} E\left(R_t\right) = E\left(R\right)$$

assuming all the $R_t's$ are drawn from the same distribution (a crucially important assumption). This verifies that the sample mean is *unbiased*. On average, across many samples, the sample mean will reveal the true mean.

The variance of the sample mean is

$$\sigma^2\left(\bar{R}\right) = \sigma^2\left(\frac{1}{T}\sum_{t=1}^{T} R_t\right) = \frac{1}{T^2}\sum_{t=1}^{T}\sigma^2\left(R_t\right) + \text{(covariance terms)}$$

If we assume that all the covariances are zero, we get the familiar formula

$$\sigma^2\left(\bar{R}\right) = \frac{\sigma^2\left(R\right)}{T}$$

or

$$\sigma\left(\bar{R}\right) = \frac{\sigma\left(R\right)}{\sqrt{T}}.$$

For stock return, $cov\left(R_t, R_{t+1}\right) = 0$ is a pretty good assumption. It's a great assumption for coin tosses: seeing heads this time makes it no more likely that you'll see heads next time. For other variables, it isn't such a good assumption, so you shouldn't use this formula.

You don't know σ. Well, you can *estimate* the sampling variation of the sample mean by using your estimate of σ, namely the sample standard deviation. Using hats to denote estimates,

$$\hat{\sigma}\left(\bar{R}\right) = \frac{\hat{\sigma}\left(R\right)}{\sqrt{T}}.$$

The classic use of this formula is to give a standard error or measure of uncertainty of the sample mean, and to test whether the sample mean is equal to some value, usually zero.

The test is usually based on a *confidence interval*. Assuming normal distributions, the confidence interval for the mean is the sample mean plus or minus 2 (well, 1.96) standard errors. The meaning of this interval is that if the true mean was outside the interval, there would be less than a 5% chance of seeing a sample mean as high (or low) as the one we actually see.

Now that we have computers, there is an easier method. We can just calculate the probability that the sample mean comes out at its actual value (or larger) given the null hypothesis, i.e. calculate this area

This is called the *p-value.*

Usually, tests are run using the *t-distribution*. When you take account of sampling variation in $\hat{\sigma}$, you can show that the ratio

$$\sqrt{T}\frac{\bar{R}-E\left(R\right)}{\hat{\sigma}}$$

is not a normal distribution with mean zero and variance 1, but a *t distribution.*

28.3 Regressions

We will run regressions, for example of a return on the market return,

$$R_t = \alpha + \beta R_{m,t} + \epsilon_t;\ t = 1, 2...T$$

and sometimes multiple regressions of returns on the returns of several portfolios

$$R_t = \alpha + \beta R_{m,t} + \gamma R_{p,t} + \epsilon_t;\ t = 1, 2...T.$$

The generic form is usually written

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \epsilon_t;\ \ t = 1, 2, ...T$$

Both textbooks and regression packages give standard formulas for estimates of the regression coefficients β_i and standard errors with which you can construct hypothesis tests.

Several important facts about regressions.

1) The population value of a single regression coefficient is[10]

$$\beta = \frac{cov\left(y,x\right)}{var\left(x\right)}.$$

2) The regression recovers the true β (precisely, the estimate of β is *unbiased*) only if the error term is uncorrelated with the right hand variables. For example, suppose you run a regression

sales = $\alpha + \beta$ advertising expenses + ϵ.

Discounts also help sales, so discounts are part of the error term. If advertising campaigns happen at the same time as discounts, then the coefficient on advertising will pick up the effects of discounts on sales.

3) In a multiple regression, β_1 captures the effect on y of only movements in x_1 that are not correlated with movements in x_2. If you run a regression of price of shoes on sales of right shoes and left shoes, the coefficient on right shoes only captures what happens to price when right shoe sales go up and left shoe sales don't. I.e., it doesn't mean much.

28.4 Partitioned matrix inverse formulas

$$\left|\begin{bmatrix} A & b \\ a' & \alpha \end{bmatrix}\right| = (\alpha - a'A^{-1}b)\left|A\right|$$

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} A^{-1}+A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{bmatrix} \\ E &= D - CA^{-1}B \end{aligned}$$

or,

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} E^{-1} & -E^{-1}BD^{-1} \\ -D^{-1}CE^{-1} & D^{-1}+D^{-1}CE^{-1}BD^{-1} \end{bmatrix} \\ E &= A - BD^{-1}C \end{aligned}$$

[10] If you forgot why, start with

$$y_t = \alpha + \beta x_t + \epsilon_t$$

multiply both sides by $x_t - E\left(x_t\right)$ and take expectations, which gives you

$$cov\left(x_t, y_t\right) = \beta var\left(x_t\right).$$

If they are symmetric,

$$\begin{aligned}\begin{bmatrix} A & B' \\ B & D \end{bmatrix}^{-1} &= \begin{bmatrix} A^{-1}+A^{-1}B'E^{-1}BA^{-1} & -A^{-1}B'E^{-1} \\ -E^{-1}BA^{-1} & E^{-1} \end{bmatrix} \\ E &= D - BA^{-1}B' \end{aligned}$$

or,

$$\begin{aligned}\begin{bmatrix} A & B' \\ B & D \end{bmatrix}^{-1} &= \begin{bmatrix} E^{-1} & -E^{-1}B'D^{-1} \\ -D^{-1}BE^{-1} & D^{-1}+D^{-1}BE^{-1}B'D^{-1} \end{bmatrix} \\ E &= A - B'D^{-1}B \end{aligned}$$