Financial Econometrics I (2021/2022)

September 1, 2021

Lecture: Introduction

Summary

- Structure of the course and of the exam
- News about the 2021 edition of the course (changes in the syllabus, exams)
- Why a (first) course of Financial Econometrics
- Why Probability and Statistics and Finance

Structure of the course

- Let us begin with the general structure of the course

Structure of the course

- 24 lectures
- + 6 exercise classes (these are classes dedicated to the solution and discussion of past exams)
- No midterms
- No group work
- (Note: in the course timetable you may see dates for more than 24 lectures and 6 exercise classes. This is to avoid having to reschedule classes due to any possible reason)

The Exam

- Note: what follows is indicative, in case of any relevant change you shall promptly be informed
- 3 calls per year
- Written closed books final exam, live or in remote according to the University decisions
- 20 closed questions, 15 multiple choice (4 answers, 2 points), 5 true-false (1 point)
- No point loss for errors, no partial points
- Time: 45(live version)+5(if in remote) mins time

The Exam

Grading: raw points shall be converted into grades according to the following table

Raw Points	Grade
30-35	30 cum laude
28-29	30
26-27	29
24-25	28
14-23	raw points+4
11-13	18
below 11	raw points

Note: minor (upward) modifications are possible according to the circumstances

The Exam

- Read the general Bocconi rules about exams with care
- In particular:
- In case of a pass grade, this shall be recorded and you cannot do the exam again
- You can enrole in any of the calls of the exam: no constraints
- You MUST be enroled to take part into the exam
- This exam is a prerequisite for the Econometrics II exam

The Exam

- A word about "curving"
- For this exam no "curving" is applied
- Bocconi requires, as a reference rule, not to be applied mechanically, some quantiles of the grade frequency distribution (for the full year: 3 calls) not to be **above** some level
- (The implicit objective is to avoid grade inflation)
- I report here the exact requirements of the University:

1

Assessment

Merit and rigor in the assessment process

- One of Bocconi University's most important values is merit.
- In particular, Bocconi places emphasis on the rigor of the student assessment process. Fair assessment and appropriate grade distribution are crucial to guaranteeing high quality education for students and defending the university's credibility.
- In addition, an appropriate grade distribution that is in line with international standards facilitates an understanding and transfer of grades between institutions in the Erasmus/ECTS system.



Assessment

The role of grade distribution

- An appropriate distribution among the various grade ranges is the result of tests/exams designed to mirror the different levels of student preparation. Such distribution is not applied mechanically, nor is it the result of forcing the way the grades are assigned.
- Passing grades in courses with a high number of students should be distributed, on the whole (out of all the exam sessions held each year), according to the benchmark distribution below.

Grade range	Marginal frequency	Cumulative frequency	
18-20	10%	≤ 20	At least 10%
21-23	25%	≤ 23	At least 35%
24-27	30%	≤ 27	At least 65%
28-29	25%	≤ 29	At least 90%
30-30L	10%		

Università Bocconi MILANO

The Exam

- Notice that the rule explicitly mentions "levels of student preparation" and not "relative levels"
- Moreover it states: "such distribution is not applied mechanically, nor is it the result of forcing the way the grades are assigned":

- An appropriate distribution among the various grade ranges is the result of tests/exams designed to mirror the different levels of student preparation. Such distribution is not applied mechanically, nor is it the result of forcing the way the grades are assigned.

- This course aims at giving you a defined absolute level of knowledge: relative grades would be meaningless
- In accordance with the rules, the grade distribution comes naturally from the different level of difficulty of the questions
- In fact, e.g. in '18/'19, all constraints are respected with the exception of the last ($F(28) \approx .845 < .90$)

Course materials

- Available on BBoard (also on Prof. Corielli webpage)
- Course and preliminary course programs/syllaboi
- Handouts covering all lectures
- Past exams with solutions (very useful, you are required to study these!)
- Notice: both the old (open questions) and the new (closed questions) version of the exam are available: they are both useful

Course materials

- As a rule, Handouts shall not be updated during the course (except for possible misprints)

Syllabus

- The Syllabus is available in pdf form with the rest of the course material and in a web version thru You@Bocconi and BBoard
- Important: check the preliminary course syllabus and related material

Provisions for September 2021

- What follows is provisional and to be confirmed according to the University policy which should be checked frequently
- The University policy for classes starting in September 2021 can be summarised as follows
- Barred possible changes due to the epidemic situation, teachers can choose between two classes profiles

Provisions for September 2021

- First profile:
- Classroom lectures with students in the classroom
- Each lecture shall be live recorded and made available to students (for a short period of time)
- If the number of students enroled in a classroom and willing to be present to the lecture is above the room capacity, the class shall be divided in two groups which shall follow each one half of the live lectures

Provisions for September 2021

- Second profile:
- About one half to 60% of the course shall be available online
- The rest of the course shall be held in the classroom as in the first profile
- In this case, however, should the class be splitted each lecture shall be held twice, once for each class

Provisions for September 2021

- The 20191 course shall follow the first profile
- The number of classes, exercise classes, and the syllabus shall remain the same, barred interruptions of didactics
- All classes should be streamed live and recorded
- The 6 exercise classes shall be held twice, one for class 16 and one for class 17

Provisions for September 2021

- Until further notice, office hours shall be held online, either by personal appointment via e-mail or, if necessary, by live Bcollaborate or Teams sessions
- Simply: write an e-mail and we shall organize a Teams appointment
- This applies to office hours dedicated to the exam "paper review"
- Full solutions shall be published and for any further question please write an e-mail

Provisions for September 2021

- VERY IMPORTANT: The anti-Covid rules in Bocconi prevent us any access to the University in case of "flu symptoms"
- A list of such symptoms (Wikipedia) is :
- Fever and chills; Cough; Nasal congestion; Runny nose; Sore throat; Hoarseness; Earache; Muscle pains; Fatigue; Headache; Irritated, watering eyes; Reddened eyes; skin (especially face), mouth, throat and nose Petechial rash; gastrointestinal symptoms such as vomiting, diarrhea, and abdominal pain
- These are "cold related" symptoms which are very frequent in autumn (also independently on Covid-19 or Flu)

Provisions for September 2021

- It is then possible that we shall have to modify, cancel or go to online lectures
- This may be a last minute thing
- This means you should frequently check for messages

Course contents

- Now, something more specific about the overall course structure:

Course contents

- The course is divided in two parts.
- The divide point should roughly correspond with the October break (circa lesson 11/12)
- The first part of the course deals with single (stock) return series, the second with vectors (portfolio)
- The main difference between the two parts is that the second shall require some matrix algebra

Course contents

- In the first part we shall use elementary probability and statistics tools in order to discuss:
- Different definition of returns and their properties
- The simple log random walk model for stock returns and its implications
- In particular:
- Expected return and return variance as a function of time
- The "stocks for the long run" argument
- How to estimate expected return and variance
- Why is it easier to estimate return variances than return expected values

Course contents

- We pass then to the definition of VaR and its computation according to several methods:
- Classic parametric Gaussian VaR (RiskMetrics)
- Non parametric VaR
- Semi parametric VaR
- For the first and the second method, confidence intervals for the estimates are discussed

Course contents

- In the second part of the course we begin with a summary of useful matrix algebra and some probability connected with matrix algebra
- We give then a quick review of linear factor models for asset pricing
- Several lectures are dedicated to a detailed analysis of the linear model, its statistical properties and its interpretation
- The linear model is used for introducing Sharpe's Style Analysis
- We then consider the problem of estimating factor models using the principal component method

Course contents

- Minor program change:
- The last topic in the course use to be the Black and Litterman model
- Beginning from this year we shall drop this topic and dedicate more time to examples (in particular in the Linear Model and in the Principal Components sections of the course)
- In past exams you shall find questions on this topic and you can simply ignore these
- This is valid for past years students too

"Mission" of this course

- The mission of this course is to give you a basic idea of how Statistics and Probability are used in financial markets
- The course does not aim at giving you ready made solutions but tools useful to build such solutions
- For this reason the course relies very little on advanced tools of Probability and Statistics and concentrates in understanding the basics
- To gain such an understanding is both the most practically useful and, perhaps, the most difficult aspect of this course

Prerequisites

- Basic notions of Probability and Statistics, the kind usually developed during a BA in Economics
- Some slightly more advanced maths: in particular some matrix algebra
- We are going to begin quite slow, so there is time for filling possible gaps
- Browse thru the preparatory course syllabus and, when necessary, the related slides
- If what you see sounds familiar, even if you do not understand each single step, fine: it is enough
- Otherwise, ask for help, I am your teacher and your consultant

Software

- This is a very short course, I would like to teach you how to work with financial data using "real" programs (R, Matlab, Stata etc) but there is no time
- All our examples shall be in Excel and very simple
- However, those of you who may be interested in statistical computing are welcome to ask (I do this things since 1979 so I think I know something about the topic)
- All students, if willing, can install Matlab with a time limited license but this is absolutely not necessary for the course
- Bocconi offers many courses in computing and is currently engaged in an effort to make its students current with Python and R
- And this is all for the introduction with a last point: why these slides

Why these slides

- The slides contain, in a condensed form, the same material as the handouts
- Hence, you can use slides, handouts or both (?) for preparing the exam (together with the overimportant past exams we shall mainly discuss in the exercise classes): nothing changes
- The real reason for the slides is the need to have a clear text for use in the remote/recorded version of the course
- I could write all formulas and relevant passages by hand using an electronic whiteboard (I'll do this sometime) but I did verify that the result of this, in particular the recorded result of this, is not good: if you follow the lecture in remote the result is passable, but resolution degrades somewhat in recording
- Moreover, it is probably useful for some of you to have a text to follow during the live and recorded lectures
- Now, let's begin

Let's begin: Probability, Statistics and Finance

- Modern Finance studies are broadly divided in two (interconnected) fields: asset pricing and Corporate Finance
- This roughly corresponds to the markets+asset management vs corporate+capital markets division in practical Finance
- In the llast 30 years a somewhat new topic: risk management (in various flavours as: market, credit, operational, legal, reputational, etc.) has been intertwined to both fields, among academicians, practitioners and regulators

Let's begin: Probability, Statistics and Finance

- Asset pricing studies the observed statistics on financial securities
- The aim of the study is twofold:
- empirical: to characterize and summarize the joint distribution of observed prices and its evolution.
- theoretical: to justify the observed data by connecting it with the overall evolution of the economic system.
- From the point of view of applications results of AP are essential for pricing, hedging, portfolio management, risk management.

- Corporate Finance, in its origins, mostly deals with the capital structure of a firm as connected with its investments.
- The original problem was that of characterizing those investments which could be valuable for the firm and how to Finance these.
- During time, further problems were considered by the field mainly concerned with organizational and managerial issues.
- How to choose and compensate management, how to implement strategies which are viable for the stakeholder, when and why to reorganize the firm, and so on.
- Due to these developments modern Corporate Finance has much in common with organization theory and, in particular with the field of industrial organization.
- Any action which involves the capital structure of a firm: IPO, mergers, acquisitions, investments, gathering of capital etc is based on CF

- On the other hand, most research in the field of Corporate Finance has a strong empirical twist being concerned on the consequences of financing, organizational and investment decision on the "value" of the firm.
- Huge databases containing corporate data have been build in the last 20/30 years and the statistical applications to corporate study have grown to and amazing size.
- The Reader should notice that by "applications" here it is not intended "academic applications"
- Any Corporate Finance action: issue of new stock, bond issues, IPOs, M&A operations, investment planning etc is strongly based an a quantitative analysis of the relevant data.

- The rate of actual real world implementation of quantitative methods is a striking factor where Finance is different from Economics in general
- Perhaps the most relevant reason for this is that most "risk factors" (we shall discuss the term) are actually quoted in the market
- This allows people working in Finance to avoid, frequently, explicit an very difficult modelling of economic agents decision at least in the common case where successful actions only depend on relative prices of securities
- (This statement may seem somewhat obscure. It shall become clearer after this semester courses)

Probability, Statistics and Finance

- The intense use of data justifies the presence of (several) courses in applied Probability and Statistics within a graduate curriculum in Finance.
- While it would be foolish to downplay the immense relevance of organizational, accounting, legal and regulatory/institutional matters, a visit to any trading floor, participation to meetings for an M&E operation, spending some time in the risk management office of a bank or simply a reading of laws and regulations concerning the management of financial companies, could easily convince you of the relevance of quantitative methods in Finance
- Another possible way to understand the role of quantitative methods in Finance (and in financial regulation) could be browsing through the program of institutional exams required in order to relate with clients in international markets.
- As an example, consider the "Series" of the FINRA Registered Representative levels in the USA

- But the simplest and most direct way, at least in our opinion, is to point out the fact that most of Finance has to do with deciding today prices for economic entities whose precise future value is unknown, be these a securitis, firms, product lines etc.
- In such a field, the availability of a language for speaking about uncertain events which, at least in principle, satisfies simple requirements of no contradiction (usually tagged as "coherence" or "no arbitrage") is obviously a necessity.
- Up to present time, the most successful language devised with such a purpose is Probability theory

Probability and Statistics

- The use of Probability in Finance is a direct offspring of the classical origin of Probability in the context of gambling.
- However, gambling problems are usually much easier to deal with than, say, asset pricing problems.
- In most "games of chance" two elements are usually agreed upon:

1. The nature of the game is such that the probabilities of its results are agreed upon by the vast majority of participants
2. In typical situations, the betting decisions of players do not change the probability of results

Probability and Statistics

- In typical games of chance, Statistics, as a tool for estimating/choosing probabilities, is not required
- (It could be required in other betting settings, the like, e.g., of horse racing, football matches etc.).
- It happens that the basic input required by Probability theory, namely the probabilities of simple events, are agreed upon in most games of chance
- As a rule, this happens because almost everybody agree on the validity of simple symmetry arguments, from which numerical values of probabilities are derived
- In this sense (beware: nobody, ever, did observe a probability) they are "known" and do not require estimation

Probability and Statistics

- Some statistics is, in fact, used for periodical checking of the "randomness" of chance generating engines used in gambling (e.g. roulettes, fortune's wheels or dice)
- In words familiar to those working in financial risk management, we could say that in chance games there is no model and estimation risk
- This is not true in the Finance milieu.

Probability and Statistics

- In the case of financial security prices, simple symmetry arguments are, as a rule, not applicable
- Prices do not reflect invariable chances arrived at by simmetry arguments
- Prices reflect, in some unknown way, information, opinions, fads and bubbles
- For this reason, we often need to estimate probabilities using past observations and models
- In other words, we need Statistical Inference, with the implied estimation risk

- Moreover, the statistical models we use are not "common knowledge"
- It is not the case that probabilities are unknown parameters to be estimated within known (viz: agreed upon) models
- Models are under continuous scrutiny, they change for each single agent and differ across agents based on theyr ex post good or bad performance
- This "model instability" is called "model risk"

- Let us pass to the second point where Finance differs from games of chance: independence of probabilities on the strategic behaviour of players.
- The fact that, say, the future price of an asset is directly dependent on the bets made on it (maybe in an irrational way) by traders is a mainstay of Finance (and Economics).
- It embodies the complex interaction between judgment of values and expectations of prices which, through the concept of equilibrium, is both the stuff of economic theory and of day by day work in the markets.
- This interaction by itself contributes in determining the probabilities of market events, entities these which cannot at any time be considered as "given" as those in typical games of chance.

- In the end, this is what makes the field of Finance so different (and much more difficult and far more interesting) than the study of games of chance.
- We do not only run risks implied by the fact that results of bets are uncertain, we are also uncertain on how to model such risk (model risk) and we need data in order to estimate parameters in the models we decide to use (estimation risk).
- Ultimately, the decision of how much to bet on a given future scenario requires both an assessment of economic value AND an evaluation of the consequences of the interacting opinions of agents.

- This is a very difficult task, which, as observed, cannot rely on simple symmetry arguments common in standard games of chances
- Tools are required for economic evaluation
- Tools are required for connecting past observations of market and, more in general, economic events, to the statement of probabilities useful for choosing among actions whose results depend on future events
- Uncertainty, in Economics and Finance, is not only complex to model as, e.g. in Physics
- It also is evolving in its structure, as it responds to the continuous evolution of events and opinions
- In the financial field this makes Probability intimately connected with Statistics and more in general with Economics

Further Reasons for the use of Probability and Statistics

- To end this lecture we shall consider two further reasons for the use of probability and Statistics in Finance
- The first is that of "Antidotes to Delusions"
- The second that of "Legal Disclaimer
- These are becoming more and more important with time

Antidotes to delusions

- Financial markets are an environment of unsolvable uncertainty: uncertainty never stops to be and to change
- The human mind does not seem to adapt well to environments of this kind
- Our brain, which craves for stable patterns, shall, if uncontrolled, tend to create such patterns out of nothing and be fooled into believing in illusions
- (Even in the simpler setting of games of chance, there exists an immense literature substantiating irrational gambling behavior and perception fallacies)
- Such ineffectivenes of human mind confronted with chance may explain at least a subset of observed "irrational" behaviours of investors.

Antidotes to delusions

- Statistics and Probability are also relevant because they can be seen as antidotes to such delusions
- They may not make us right most of times and it may quite be that some lucky dumbo shall enjoy results better than ours
- But, at least, they may help us in not being upset if something which a priori we considered likely does indeed happen or prevent us in changing our decision rule due to events which confirm the optimality of such rule
- Maybe it is not much but in the not too long run it counts for much
- They help us, for instance, in understanding that, given the amount of variance in the market ant the huge number of "dumb" investors, the fact some "dumb" investor was or shall be better off than us is so likely to be almost sure and, for this reason, this fact should not upset us or induce us in "dumb" behavior

Antidotes to delusions

- There exists a classic literature on the topics of gambling, luck and delusions.
- It goes back at the very least to Heraclitus, Cicero and Plutarch but traces can be seen in the Bible itself.
- As in other important fields, time and again the principles of this literature are "rediscovered" by well meaning, while perhaps not sufficiently well read, Authors.
- It is peculiar of Finance the fact that such rediscovery always surprises the readers as if it presented true new ideas.
- This tells us much about our inability of learning how to deal with uncertainty.
- No matter how many times delusions are explained, they are bound to come back again.

Quantitative methods as legal disclaimer

- In recent years, first, in the US and then all over the world, a new role of quantitative models has surfaced and is becoming pervasive.
- That of legal disclaimer.
- Suppose you are an asset manager and your client is not satisfied by your results.
- The client is going to question you about what did you do in order to get such results.
- Maybe such questioning could take the form of a legal procedure where you are sued for malpractice.

Quantitative methods as legal disclaimer

- This is common to most professions, just consider the medical profession as a classic instance.
- Answering to such actions requires some definition of "best" practice, which, in most professional fields, is very difficult
- (In principle, if such definition were really possible, it would take the form of a computer program which could take the place of the involved professionals).
- In most fields, this has taken the shape of "protocols of intervention" which should define the"best practices" a professional should follow when dealing with a case.

Quantitative methods as legal disclaimer

- This was undoubtedly a useful development in some field, but was also accompanied by a huge increase of the bureaucratic aspect of any profession
- Also, this did imply an increase of rigidity of behaviour when dealing with situations where following the protocol is potentially dangerous for the client/patient.
- In these cases, the "protocol following" behaviour is useful for the professional, which could defend on this basis, but could be dangerous for the client.

Quantitative methods as legal disclaimer

- In Finance, in particular in the asset management field, the establishment of best practices and protocols takes the form, in most circumstance, of quantitative models of asset selection and evaluation which, in order to get an easy legal validity, are mostly based on standard academic arguments.
- It much easier to defend yourself in a Court of Justice by saying:
- "The asset allocation model we follow is *** and is based on published research by ***. The model gave a small probability of negative results which, alas, did actually happen",
- Than saying:
- "I choose the asset allocation on the basis of my gut feeling and past experience. it could go well, it could go bad. Though luck" it did go bad".

Quantitative methods as legal disclaimer

- While there may be not much difference between the asset allocations, the first defense is certainly stronger than the second, at least in a court
- This attitude is so widespread the, in some fields of Finance, it has become law: just think about the risk management rules derived from the sequel of the Basel agreements.
- Just like in the field of medicine, such developments had some positive consequences (beyond the disclaimer effect) in that they require discipline and clarity of analysis on the part of agents.

Quantitative methods as legal disclaimer

- They also contributed to spreading a "formal" use of quantitative models which, when not continuously questioned and updated, could increase the risk of decision taken "because the model says so" just in view of the legal disclaimer a model may offer.
- Whatever may be your evaluation of such a development, this is the world of Finance you are going to enter.
- If you want to be clever agents in this kind of work, you need to understand it and understanding requires (also!) the understanding of the quantitative tools applied for quantifying and for taking risks.

In the end a warning

- In this lecture we gave a general description of the course
- We did this is words only and this may have create the impression of some imprecision
- Do not worry
- This is the first and last lecture of the course mainly based on words and not on formulas

Lecture: Returns

Summary

- Different definitions of return
- Linear (percentage) returns and log "returns"
- Linear returns and across security additivity
- Log returns and time additivity
- Why returns if what we observe are prices

Returns

- Let P_{it} be the price of the $i-th$ stock at time t.
- The linear or simple return between times t_{j-1} and t_j is defined as:

$$r_{it_j} = P_{it_j}/P_{it_{j-1}} - 1$$

- The log return is defined as:

$$r^*_{it_j} = \ln(P_{it_j}/P_{it_{j-1}})$$

- In both these definitions of return we do not consider possible dividends.
- There exist corresponding definitions of "total return" where, in the case a dividend D_j is accrued between times t_{j-1} and t_j, the numerator of both ratios becomes $P_{t_j} + D_j$.

- Clearly, we can transform back returns into prices, if we have a starting (or ending) price
- If we know $P_{it_{j-1}}$ we have

$$P_{it_j} = P_{it_{j-1}}(1 + r_{it_j})$$

$$P_{it_j} = P_{it_{j-1}} \exp(r^*_{it_j})$$

- If we know P_{it} a simple manipulation of the two formulas allows us to move backward in time to $P_{it_{j-1}}$
- Notice that, obviously, the use of log returns in the first formula and/or linear returns in the second would give wrong results

- Here we do not apply any accrual convention to our returns, that is: we just consider period returns and do not transform these on a, say, yearly basis.
- It is to be noticed that, while P_{t_j} means "price at time t_j", r_{t_j} is a shorthand for "return between time t_{j-1} and t_j" so that the notation is not really complete and its interpretation depends on the context.
- When needed for clarity, we shall specify returns as indexed by the start and the end point of the interval in which they are computed as, for instance, in $r_{t_{j-1};t_j}$.

Returns

- The two definitions of return yield different numbers, the more so the more the ratio between consecutive prices is far from 1
- Consider the Taylor formula for $\ln(x)$ for x near to 1:

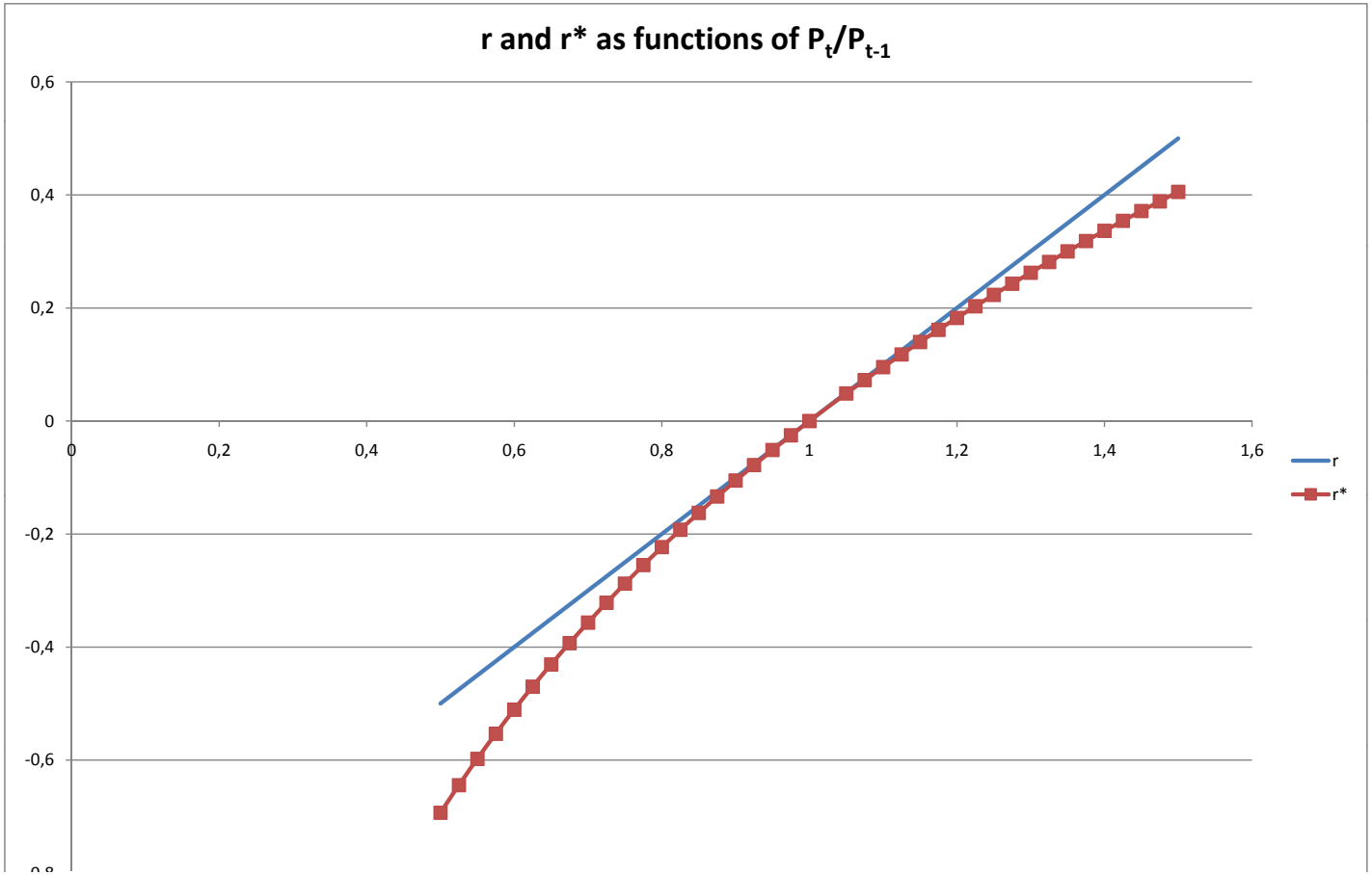
$$\ln(x) = \ln(1) + (x-1)/1 - (x-1)^2/2 + ...$$

- If we truncate the series at the first order term we have:

$$\ln(x) \cong 0 + x - 1$$

- Now, let x be the ratio between consecutive prices, only for x near one the linear approximation is good and the two definitions give similar values.

Returns


r and r* as functions of P_t/P_{t-1}
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
0
0,2
0,4
0,6
0,8
1
1,2
1,4
1,6
r
r*

- It is clear that $\ln(x) \leq x - 1$ for all x (in symbols $\forall x$)
- In fact, $x - 1$ is equal to $\ln(x)$ in $x = 1$ and has the same derivative for $x = 1$ so it is tangent to $\ln(x)$ for $x = 1$ (the derivative of $\ln(x)$ is $1/x$ which is 1 for $x = 1$, and that of $x - 1$ is always 1)
- Moreover, the second derivative of $x - 1 - \ln(x)$ is: $1/x^2$.
- This is always positive (and defined for $x \neq 0$. In our case we deal with positive prices so the price ratio $x > 0$) so that, the more we move to the left or to the right of the tangency point, the bigger becomes the difference

- Notice that, while the most negative possible value of the linear return (with nonnegative prices) is -1 (you cannot lose more that 100% from a limited liability security), if P_{it}/P_{it-1} goes to 0 the log return goes to minus infinity
- You clearly see how the log return is NOT a "percentage return" in any sense (again: you cannot loose "infinity precent" from a limited liability exposure)

Returns

- The two definitions yield different numbers if x is far from 1. When do we espect this to happen?
- In Finance, the ratio of consecutive prices over a short time interval (maybe corrected by taking into account accruals) is often modeled as a random variable with an expected value very near 1.
- This implies that the two definitions shall give different values with sizable probability only when the variance (or more in general the dispersion) of the price ratio (x) is non negligible, so that observations far from the expected value have non negligible probability.
- Standard models in Finance, as we are going to see shortly, assume that the variance of returns increases when the time interval on which the return is computed increases
- This means that, when such models are valid, the two definitions shall more likely correspond to different values when applied to longer horizon returns.

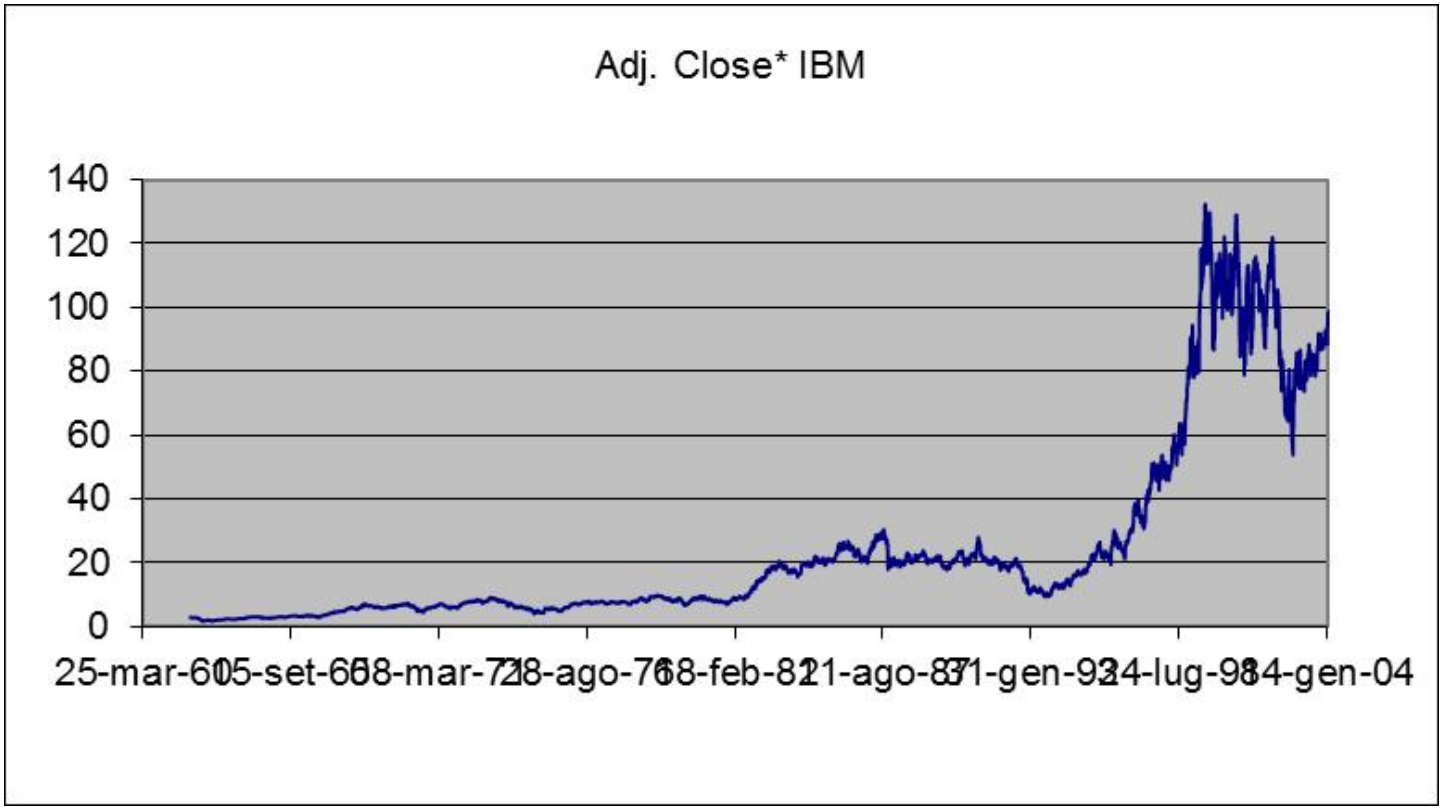
Returns

- However... why bother with two definitions?
- Let us go a step back and ask: why we observe prices, but discuss models based on returns
- This depends on the fact that, while prices and returns tell the "same story"...
- ...returns seem to have a "more homogeneous" behavior over time
- A model of prices shall be easier to build if we begin with a model of returns

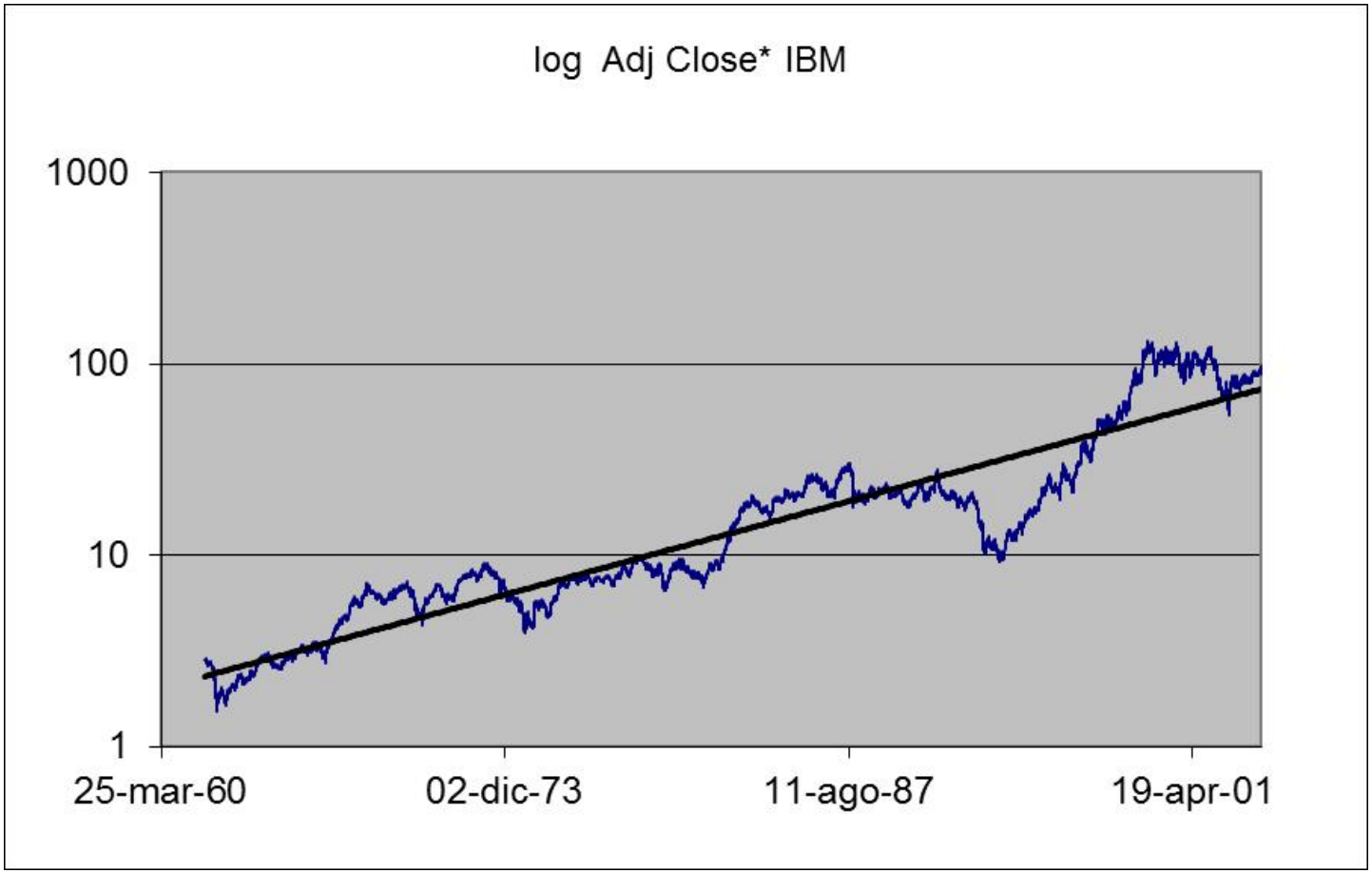
Returns

- Consider, e.g., the following prices, log prices and returns plots for a long series of IBM daily prices
- All four plots show some lack of homogeneous behaviour in time:
- The price plot shows an "exponential trend": you could never swap different time sections of the plot without realizing you did this
- The log price shows a linear trend, however, around this trend the beaviour of the series is ore homogeneus that that of price
- Linear and, even better, log returns only show, at first sight, "variable volatility" but you could swap sections of the series without realizing you did this (almost)
- Exactly this makes easier to model prices by first modelling returns

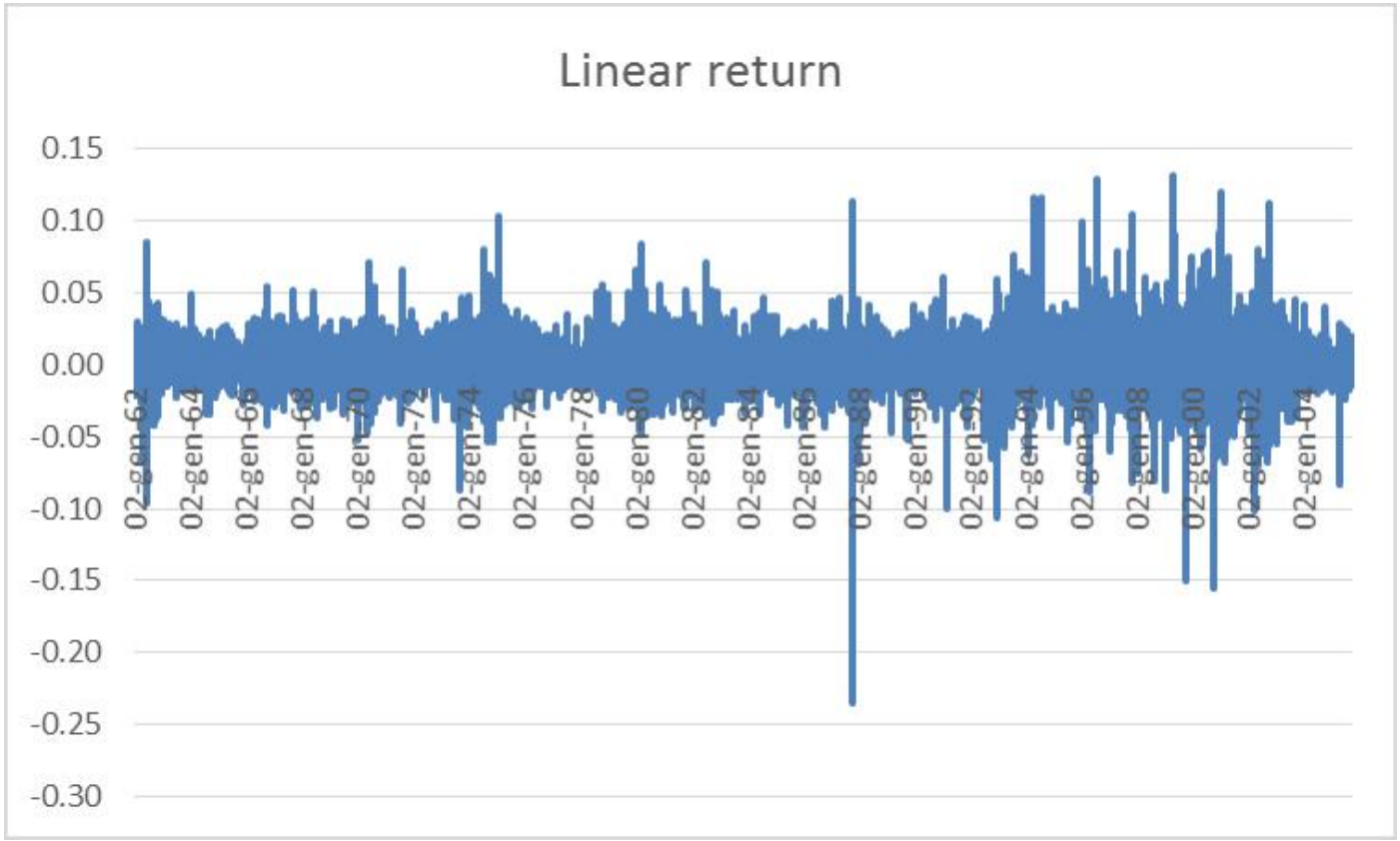
Returns


Adj. Close* IBM
140
120
100
80
60
40
20
0
25-mar-60
05-set-66
08-mar-72
08-ago-76
08-feb-82
21-ago-87
31-gen-93
24-lug-98
04-gen-04

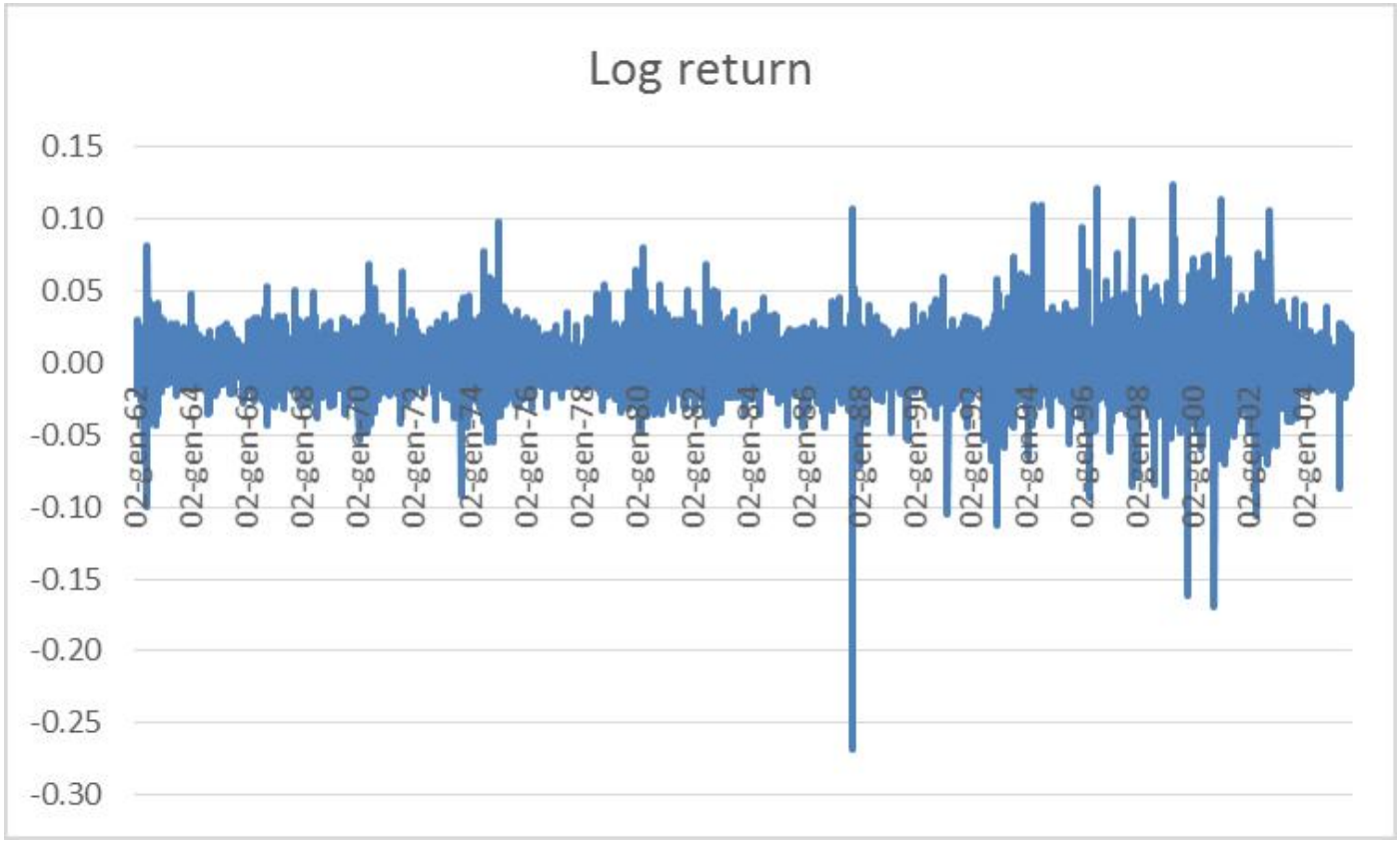
Returns


log Adj Close* IBM
1000
100
10
1
25-mar-60
02-dic-73
11-ago-87
19-apr-01

Returns


Linear return
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
02-gen-62
02-gen-64
02-gen-66
02-gen-68
02-gen-70
02-gen-72
02-gen-74
02-gen-76
02-gen-78
02-gen-80
02-gen-82
02-gen-84
02-gen-86
02-gen-88
02-gen-90
02-gen-92
02-gen-94
02-gen-96
02-gen-98
02-gen-00
02-gen-02
02-gen-04

Returns


Log return
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
02-gen-62
02-gen-64
02-gen-66
02-gen-68
02-gen-70
02-gen-72
02-gen-74
02-gen-76
02-gen-78
02-gen-80
02-gen-82
02-gen-84
02-gen-86
02-gen-88
02-gen-90
02-gen-92
02-gen-94
02-gen-96
02-gen-98
02-gen-00
02-gen-02
02-gen-04

Returns

- This shall be made more precise in the following lectures
- Now, back to the "why two returns" question
- Returns are not only more "stable" in behaviour, using returns simplify some common computations
- Two different definitions exist because, both in modelling and in simplifyng computations, each is useful in different circumstances
- To undestand this is the purpose of what follows
- We anticipate that linear returns shall be useful for describing the evolution of portfolios over a single time interval, while log returns shall be useful when dealing with single securities over many time intervals

- From now on, for simplicity, let us only consider times t and $t-1$.
- So, each time interval shall be of the same length (albeit unspecified: 1 day, week, month...)
- Let the value of a buy and hold portfolio, composed of k stocks each for a nominal quantity n_i, at time t be:

$$\sum_{i=1..k} n_i P_{it}$$

- Let us show that the linear return of the portfolio, over one time interval, is a linear function of the returns of each stock.

$$r_t = \frac{\sum_{i=1..k} n_i P_{it}}{\sum_{j=1..k} n_j P_{jt-1}} - 1 = \sum_{i=1..k} \frac{n_i P_{it}}{\sum_{j=1..k} n_j P_{jt-1}} - 1$$

$$= \sum_{i=1..k} \frac{n_i P_{it-1}}{\sum_{j=1..k} n_j P_{jt-1}} \frac{P_{it}}{P_{it-1}} - 1 =$$

$$= \sum_{i=1..k} w_{it}(r_{it}+1) - 1 = (\sum_{i=1..k} w_{it} r_{it} + \sum_{i=1..k} w_{it} 1) - 1 =$$

$$= \sum_{i=1..k} w_{it} r_{it} + 1 - 1 = \sum_{i=1..k} w_{it} r_{it}$$

- Where $w_{it} = \frac{n_i P_{it-1}}{\sum_{j=1..k} n_j P_{jt-1}}$ are non negative "weights" (comment on short positions) summing to 1 which represent the percentage of the portfolio invested in the i-th stock at time $t-1$.

- This simple result is very useful.
- Suppose, for instance, that you know at time $t-1$ the expected values for the returns between time $t-1$ and t.
- Since the weights w_{it} are known, hence non stochastic at time $t-1$, we can easily compute (at time $t-1$) the expected return for the portfolio as:

$$E(r_t) = \sum_{i=1..k} w_{it} E(r_{it})$$

- Moreover if we know all the covariances between r_{it} and r_{jt} (if $i = j$ we simply have a variance) we can find the variance of the portfolio return as:

$$V(r_t) = \sum_{i=1..k} \sum_{j=1..k} w_i w_j Cov(r_{it}; r_{jt})$$

- For log returns this is not so easy. In fact we have:

$$r_t^* = \ln\left(\frac{\sum_{i=1..k} n_i P_{it}}{\sum_{j=1..k} n_j P_{jt-1}}\right) = \ln\left(\sum_{i=1..k} \frac{n_i P_{it-1}}{\sum_{j=1..k} n_j P_{jt-1}} \frac{P_{it}}{P_{it-1}}\right) =$$

$$= \ln\left(\sum_{i=1..k} w_{it} \exp(r_{it}^*)\right)$$

- The log return of the portfolio is not a linear function of the log (and also of the linear) returns of the components.
- In this case assumptions on the expected values and covariances of the components cannot be directly translated into assumptions on the expected value and the variance of the portfolio.
- Think how difficult this could make any standard portfolio optimization procedure the like of Markowitz model.

Returns

- Before dealingn with log returns we stress, again, an important point.
- All computations we made here suppose that prices at time $t-1$ are known, that is: non stochastic.
- The same easy computation would not be possible at any previous time $t-q$ with $q>1$ as at that time w_{it} would still be stochastic (go back and see again what exactly is w_{it})
- In this case, we could not make passages as:

$$E_{t-1}(r_t) = \sum_{i=1..k} E_{t-1}(w_{it} r_{it}) = \sum_{i=1..k} w_{it} E_{t-1}(r_{it})$$

- We should be satisfied by the almost useless

$$E_{t-q}(r_t) = \sum_{i=1..k} E_{t-q}(w_{it} r_{it})$$

- (Notice the indexes which, only in this slide, stress the "moment in time" you compute the expected value

- Now, log returns.
- These are much easier to use than linear return when we consider a single stock price evolving over time.
- Suppose we observe the price P_{t_i} at time $t_1, \ldots t_n$ the log return between t_1 and t_n shall be:

$$r^*_{t_1,t_n} = \ln \frac{P_{t_n}}{P_{t_1}} = \ln \frac{P_{t_n} P_{t_{n-1}}}{P_{t_{n-1}} P_{t_1}} = \ldots = \ln \prod_{i=2\ldots n} \frac{P_{t_i}}{P_{t_{i-1}}} = \sum_{i=2\ldots n} r^*_{t_i}$$

- It is then easy, given the expected values and the covariances of the sub period returns, to compute the expected value and the variance of the full period return (from t_1 to t_m)

$$E(r^*_{t_1,t_n}) = \sum_{i=2\ldots n} E(r^*_{t_i}); \quad V(r^*_{t_1,t_n}) = \sum_{i=2\ldots n} \sum_{j=2\ldots n} Cov(r^*_{t_i}; r^*_{t_j})$$

- On the other hand, this does not happen for the linear return. We have:

$$r_{t_1,t_n} = \frac{P_{t_n}}{P_{t_1}} - 1 = \frac{P_{t_n} P_{t_{n-1}}}{P_{t_{n-1}} P_{t_1}} - 1 = ... =$$

$$= ... = \prod_{i=2...n} \frac{P_{t_i}}{P_{t_{i-1}}} - 1 = \left[\prod_{i=2...n} (r_{t_i} + 1) \right] - 1$$

Returns

- In general, the expected value of a product (and notice that the factors in the product are 1 plus the returns, not the returns) is difficult to evaluate and does not depend only on the expected values of the terms.
- For the computation of the variance, the problem is even bigger.
- A noticeable special case of some practical value is that of independence (non correlation is not enough) among returns.
- In this case the expected value of the product is the product of the expected values.
- We have a formula for the variance, but this is in general messy and is NOT the product of variances, except in the zero expected value case.

Returns

- Too cumbersome: much better the use of log returns.
- In the end, we shall typically use log returns for single series of prices over time and linear returns for a portfolio of securities over a single time period.
- It is clear that, when very realistic "mixed" problems, that is: problems involving the modeling of portfolio evolution over time, are considered, we shall often see promiscuous and inexact use of the the two definitions.
- Or we shall drop returns and go back to prices.
- You may be more used to the linear then the log return, at least in basic portfolio theory.
- This because standard "introductory" portfolio allocation models are one period buy and hold models.

- To sum up: the two definition of returns yield different values when the ratio between consecutive prices is not equal to 1.
- The linear definition works very well for portfolios over a single period and conditional to the knowledge of prices at time $t-1$:
- Expected values and variances of portfolios can be derived by expected values variances and covariances of the components
- For analogous reasons the log definition works very well for single securities over time.

Returns

- We conclude this section with three warnings.
- First:
- Many other definitions of return exist.
- Each definition origins from either traditional accounting behavior (and typically is connected with some specific asset class) or from specific computational needs.
- Most of these definitions are based on linear returns, but use different conventions on rate compounding and on the accounting for the number of days between two prices and the accrual of possible dividends and coupons.

- Second:
- No single definition is the "correct" or the "wrong" one.
- In fact such a statement has no meaning.
- The correctness in the use of a definition depends on the context in which it is applied (accounting uses are to be satisfied) and, obviously, on avoiding naive errors the like of exponentiating linear returns for deriving prices or summing exponential returns over different securities in order to get portfolio returns.

- Third:
- This multiplicity of definitions requires that, when we speak about any properties of "returns", it should be made clear which return definition we have in mind.
- For instance:
- The expected value of log returns must not be confused with the expected value of linear returns.
- The probability distribution of log returns shall not be the same as the probability distribution of linear returns, and so on.

Returns

- Practitioners are very precise in specifying all minute details of such definitions in financial contracts
- You do not see the same precision in financial newspapers and, frequently, in academic papers

Lecture: Returns: Some Empirical Facts

Summary

- Need for knowing the "history of numbers" in Finance
- Vague and precise questions: "how did the US stock market behave since 19th century?"
- A comparison between Jeremy Siegel and Robert Shiller
- Understalding long run and short run data
- More, sometimes surprising, data on major asset classes

Some empirical facts

- Finance is "full of numbers", price data and related statistics are gathered for commercial and institutional (infrequently also for research) reasons and are readily available on free and commercial databases.
- As in any field a researcher must be cautious before using these numbers.
- What does the number mean? How was it recorded? Did it always mean the same thing?
- This is not the place for a detailed discussion but it could be useful for us to try and analyze a very simple example.

Some empirical facts

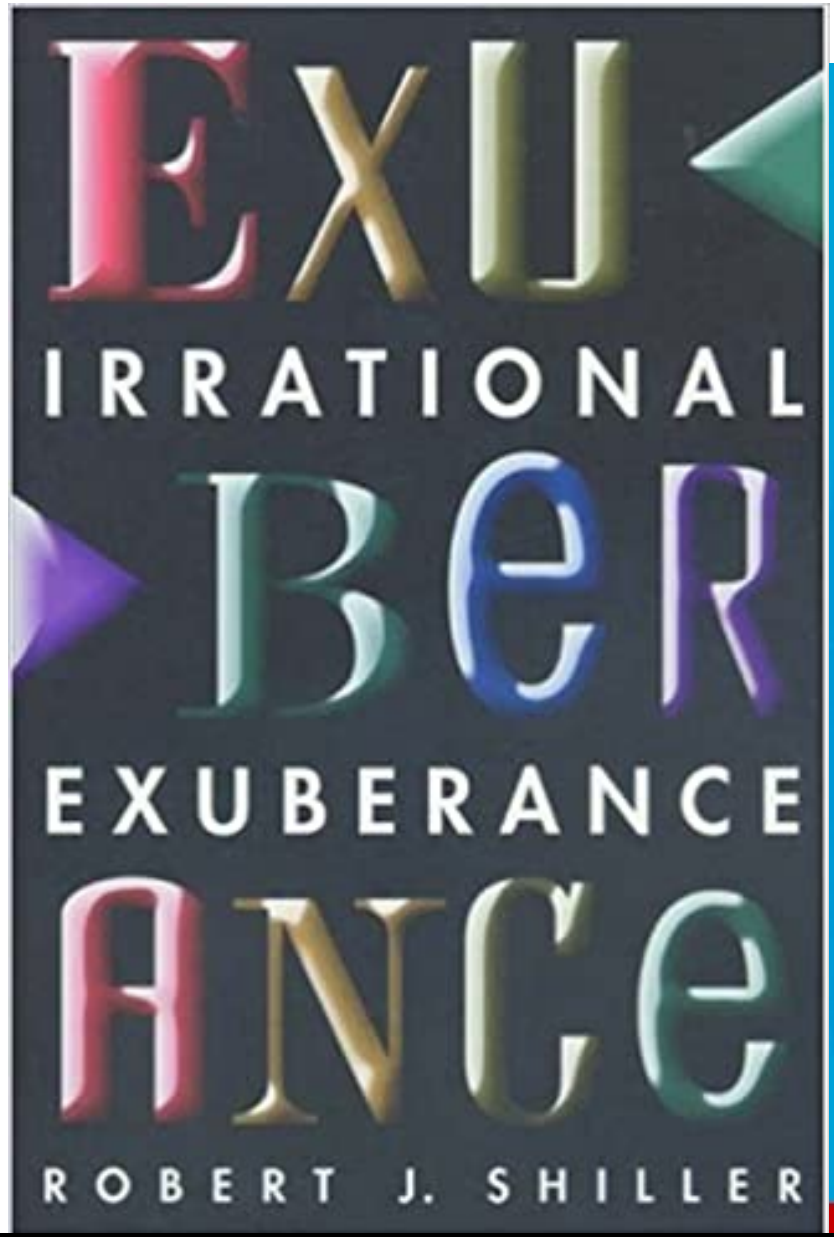
- Suppose you wish to answer the following question: "how did the US stock market behave since 19th century?".
- You browse the Internet and run a search for literature on the topic.
- Suppose you are able to shunt off conspiracy theorists, Finance fanatics, quack doctors and serpent oil sellers, Ponzi scheme advertising and the like.
- Let us say that you concentrate on academic and academic linked literature.

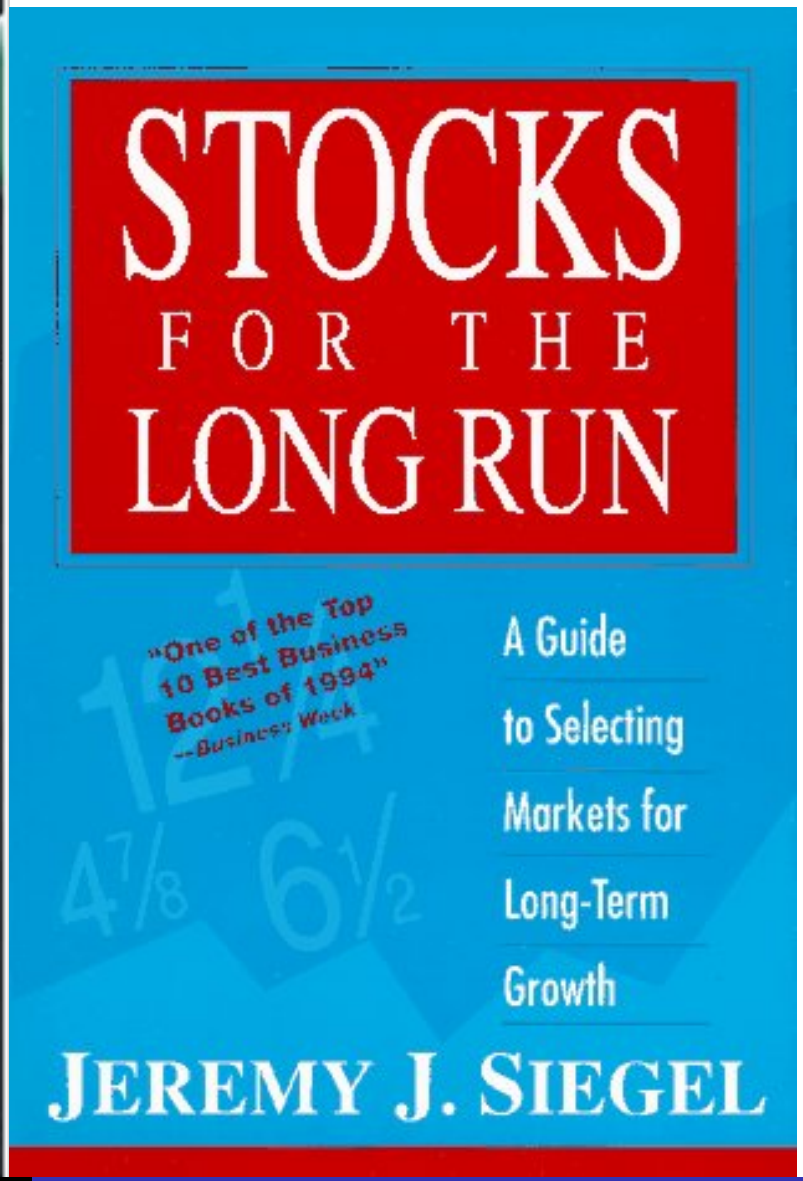
Some empirical facts

- You find a whole jungle of academic and semi academic references among which you choose two frequently quoted expository books, authored by famous academicians:
- "Irrational exuberance" by Robert J. Shiller (of Yale) (first edition january 1994)
- and
- "Stocks for the long run" By Jeremy J. Siegel (of Wharton) (first edition march 2000)

Some empirical facts

First edition covers





Some empirical facts

- These two Authors share a similar academic background: both were PhD students at MIT.
- Siegel, (born in 1945) got his PhD in 1971 (with Samuelson, solow and Modigliani as advisors)
- Shiller (born in 1946) got his PhD in 1972 (with Modigliani as advisor).
- Siegel thesis in in monetary policy and deals with the effect of expectations on the results of monetary policy.
- Shiller thesis deal with how rational expectations shape the term structure of interest rates (a fashionable topic at the time).
- Shiller got the Nobel medal in 2013.
- Siegel, quite likely, got much more money in his consulting work.

Some empirical facts

- You browse through the first chapter of both books (in what follows I mostly use first editions) and find this:
- Figure 1-1 of Siegel tells you that 1 dollar invested in stock in 1802 would have become 7,500,000 dollars by 1997.
- Moreover you read that 1 dollar of 1802 is equivalent (according to Siegel) to 12 dollars in 1997.
- The real return should have been of about 625,000 times (7,500,000/12) in real terms (62,500,000% !)
- On the other hand Figure 1.1 of Shiller's book gives the following information:
- Between 1871 and 2000 the S&P composite index, corrected for inflation, grew from (roughly) 70 to (roughly) 1400 with a real return of roughly 20 times (2000%).

Some empirical facts. Siegel Figure 1-1

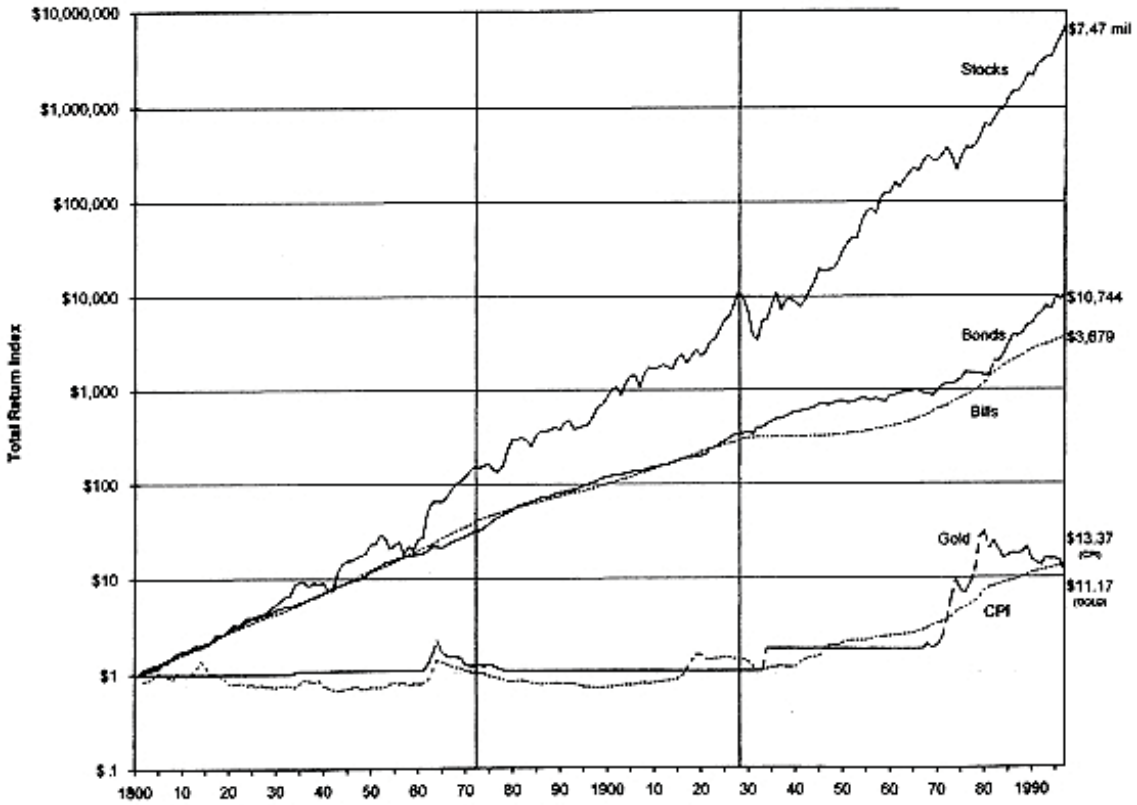


FIGURE 1-1
Total Nominal Return Indexes, 1802-1997

Some empirical facts. Shiller Figure 1.1



Figure 1.1

Stock Prices and Earnings, 1871–2000

Real (inflation-corrected) S&P Composite Stock Price Index, monthly, January 1871 through January 2000 (upper series), and real S&P Composite earnings (lower series), January 1871 to September 1999. *Source:* Author's calculations using data from S&P Statistical Service; U.S. Bureau of Labor

Some empirical facts

- Both numbers are big, but also quite different.
- Now you are puzzled.
- Sure: a part of the difference is due to the different time basis.
- Looking to Siegel picture, you see that the dollar value of the investment around 1870 was about 200.
- Exaggerate inflation: attribute the full 12 times devaluation to the 1870-2000 period (see below on this point)
- In this case 200 dollars of 1870 would be equivalent to 2400 1997 dollars
- The real increase is now of 3125 times (625,000/200) which is still more than 150 times Shiller number (20 times).

Some empirical facts

- This, obviously, cannot come from the difference in the end date of the sample.
- The period: 1997-2000 was a bull market period (internet bubble) and should reduce, not increase, the difference.
- Now, both Authors are famous Finance professors and both are media gurus for Finance and markets.
- So the problem must be in the reader (us).
- Let us try and improve our understanding by reading the details.
- First: we notice that Siegel quotes as source for the raw data the Cowles series as reprinted in Shiller book "Market volatility" for the 1871-1926 period and the CRISP data for the following period, while Shiller speaks about the S&P composite index.

Some empirical facts

- Reading with care we see another difference: Shiller speaks about a "price" index while Siegel about a reinvested dividends "total return" index.
- Is this the trick?
- Browsing the Internet we see that Shiller's data are actually available for downloading.
- This is the link: Shiller data web page
- We can compute the total return for Shiller data between 1871 and 1997 and the real increase now is from 1 dollar to 3654 dollars in real terms (now bigger than Siegel's).
- We also see that the CPI passed from 12 to 154 in the same time interval so that the "12 times" rule for the value of the dollar used by Siegel seems a good approximation.

Some empirical facts

- There is still some disagreement between the numbers
- Siegel 3125, but with exaggerated inflation, and Shiller 3654
- These two numbers correspond, roughly, to: yearly log rates of 5.37 and 5.46 (do not say %!) for 150 years
- Not so different, if we compare with the starting point
- So we think that, at least for solving our puzzle, we understood the basics.

Some empirical facts

- In this very short and summary analysis we did learn some important things.
- First: understand your question.
- "How did the US market behave since 19th century" is, now we realize, not quite a well specified question.
- Are we looking for a summary of the history of prices, or for the history of one dollar invested in the market?
- The two different questions have two different answers and require different data.
- Second: understand your data. Price data? Total return data? Raw or inflation corrected?

Some empirical facts

- There are many subtle but relevant points that should be made, we only mention three of these
- First: What do we really learn from this answer? Is the result "good" or "bad" and compared with... what?
- Do you know the growth rate, in real terms, of the US economy in the same interval?
- Estimates of the real growth rate of US GDP per person between 1820 and 200 are about 1.7% per year on average
- This is much less than 5.37 or 5.46% but this is a growth of GDP not of wealth
- So, you should drop dividend reinvestment from the comparison and if you do this you get back something of the order of 1.5%

Some empirical facts

- Second: Can we extrapolate our observation of the US market to other markets and/or times?
- Can we extrapolate it to, say, the next 200 years of US market?
- Survivorship Bias: we are considering the US market, today, because, today, this is the most relevant stock market
- Things were completely different 200 years ago
- Had I randomly chosen a market existing 200 years ago and study its evolution in 200 years, with all probability I would not find something similar
- The comparison with any other market could be unfair: it may be that we are simply studying a "monster" market just because "it has been a monster" so that it became the most relevant in the world
- This could tell us very little about the potential evolution of an "average" market

Some empirical facts

- Third: We deflate data...
- Inflation rates in the long run are a big puzzle, how to define a relevant basket to be priced thru time?
- E.g.: (data from the Internet, no guaranty at all!) a very good riding horse in 1865 had a cost of about 200 dollars (of 1865)
- A comparable car today could be priced at about 50000 dollars. Not 12 times but 250 times more
- On the other hand: 12 times 200 is 2400 and this could be the price of a scooter (or of a good bicycle) today
- Which is the right comparison?

Some empirical facts

- The standard fee for the home call of a doctor was 2 dollars. This (if still possible) did increase a little more than 12 times
- The same could be argued for the rent of a four-room house in most eastern US cities. This ran at about $4.50 per month
- Important "goods" simply disappear: around 1860, 1/3 of the US southern states population was slave.
- The cost of a slave was, I read, minimum, 800 dollars and a skilled one, e.g. a blacksmith, could cost 2500 dollars (equivalent to paying the salary of a free blacksmith for five years)

Some empirical facts

- A last note on the overall hystory of inflation in USA
- This is, maybe, not what you expect
- (btw: know all these facts and make good impression on interviewers!)
- According to the Bureau of Labor Statistics data (from 1913) and related research (before 1913) the price level oscillated between 1800 and 1917 with no overall trend
- The overall level was about constant (so our hypothesis about inflation is not so bad)
- It doubled (+100%) between 1917 and 1950 and from 1950 to 2019 (mostly after 1965) it increased tenfold (+1000%)

Some empirical facts. Agree to Disagree

- We stop here, for the moment, and do not further mention the fact that a lot of discussion has run about the relevance of the questions and of the answers and their interpretation.
- The fact is: Siegel and Shiller start with similar (almost identical) data but they reach quite different conclusions
- (At least, this is their opinion on their work and this is reflected by the very different way they use to summarize and present the same data).
- We can reconcile the data: we understand they are using the same data in two different ways.
- However, why each of them draw a different conclusion and, moreover, why they did "agree to disagree"?

Some empirical facts

- We are not discussing this point, for the moment but... remember model uncertainty and the idea that, if your models are different, you may draw different conclusions from the same data
- Or, maybe, a different model is simply another name for the different conclusions you draw from the same data.

Some empirical facts

- While we realize, if not fully understand, these differences of opinion, let us take the opportunity to state several empirical "facts" that underlie much of the discussion about the long run behaviour of US stock market.
- We do this with the yearly Shiller dataset (widely used in academic literature).
- We shall concentrate on the total log return series.
- The dataset starts with 1887 and is updated each year since, in the latest available version, Shiller uses the dataset up to 2013 included we shall limit our computations to the interval 1871-2013.

Some empirical facts

- This and similar datasets concerning the long term evolution of the US market are the most studied datasets in the field
- As considered above, this could be dangerous as it may be that we are concentrating on something "exceptional"
- And, in fact, even simple empirical considerations create huge "empirical puzzles"
- In fact almost everything in the data is "puzzling"
- Let us see why

Some empirical facts

- During this time interval the average real log total return of the index was 6.33%.
- In the same period the average real one year interest rate was 1.03% so that the so called risk premium was about 5.3%.
- The standard deviation of the real log total return was 17.09% while the same statistic for one year real interest rates was 6.54%.
- This is ex post, but it was true for a long time period: at this conditions why should anyone invest much money in bonds and not in stocks?

Some empirical facts

- In fact, the 5.3% average real log total return in excess to the yearly rate (which was even higher up to year 2000) compared with the 17.09% standard deviation (even smaller than this up to 2000) did generate a literature concerned with the "equity premium puzzle".
- The 1.03% average one year rate compared with 6.54% std generates the "risk free rate puzzle".
- Buth there is more...

Some empirical facts

- The average of the real dividend yield (up to 2011 only) is 4.45% and the standard deviation of the same is 1.5%.
- The average real log price return was 2.16% and the standard deviation of the same 17.68%.
- So: no equity premium puzzle if you only consider prices.
- But then... Most of the growth in real value of stocks is due to dividends and dividends have very little std ("dividend puzzle"?)

Some empirical facts

- Beware: we can only approximately sum price and dividend growths and compare them with the total real log return
- However, we still see that most of the equity premium is associated with the dividend yield.
- Notice that the correlation coefficient between real dividend yield and real log price return is .10 (positive but small)
- This explains why the standard deviation of the total real log return is even smaller than the sum of the standard deviations of log real price return and real dividend yield.
- On the other, hand this small correlation is, by itself, a puzzle (after all, standard understanding is that today's value of a firm should be equal to the expected discounted value of its future dividends).

Some empirical facts

- You can see how many interesting puzzles are born from simple observation of a single series
- Perhaps this means that this series is, really, an anomaly?
- Who knows! We shall reconsider this in the near future
- (BTW: this is a very good topic to study and keep in mind so that, in an interview, when they ask you "why are you interested in markets" you may have a nice and engaging answer)

Some empirical facts

- A last piece of simple data analysis: the 1 year autocorrelation of the real total log return series is very small: 2.29/100.
- This is a first simple evidence of the fact that is is very difficult to forecast future returns on the basis of past returns.
- Some of these empirical facts are at the basis of the simple stock price evolution model we shall introduce in the next chapter.

Some empirical facts. The pictures

- The following slides contain some more information on the "size" of financial markets
- These are not required for the exam but very useful for your general knowledge
- I'll discuss these in the classroom if we have time, otherwise, read these by yourself (and recall the interview!)

Some empirical facts. The pictures

- Some more about the long term evolution of US market
- See here Siegel figure 1.1, from the first edition and the 2014 edition of Siegel book "Stocks for the Long Run", with a comparison of nominal values for one dollar in 1802 invested in:
- the stock market,
- a treasury bond,
- treasury bill
- gold
- These are compared with the Consumer Prices Index
- In the second version of the slide you do not see the CPI but the "value of 1 US dollar"

Some empirical facts. The pictures

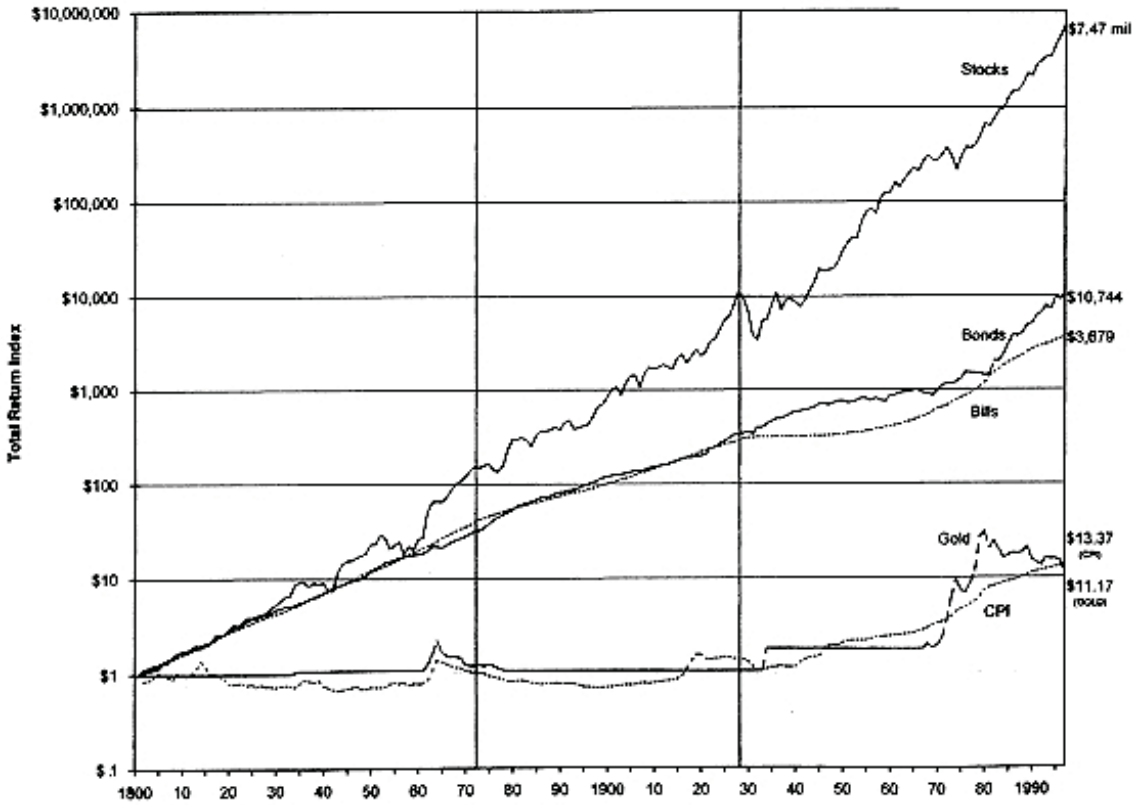
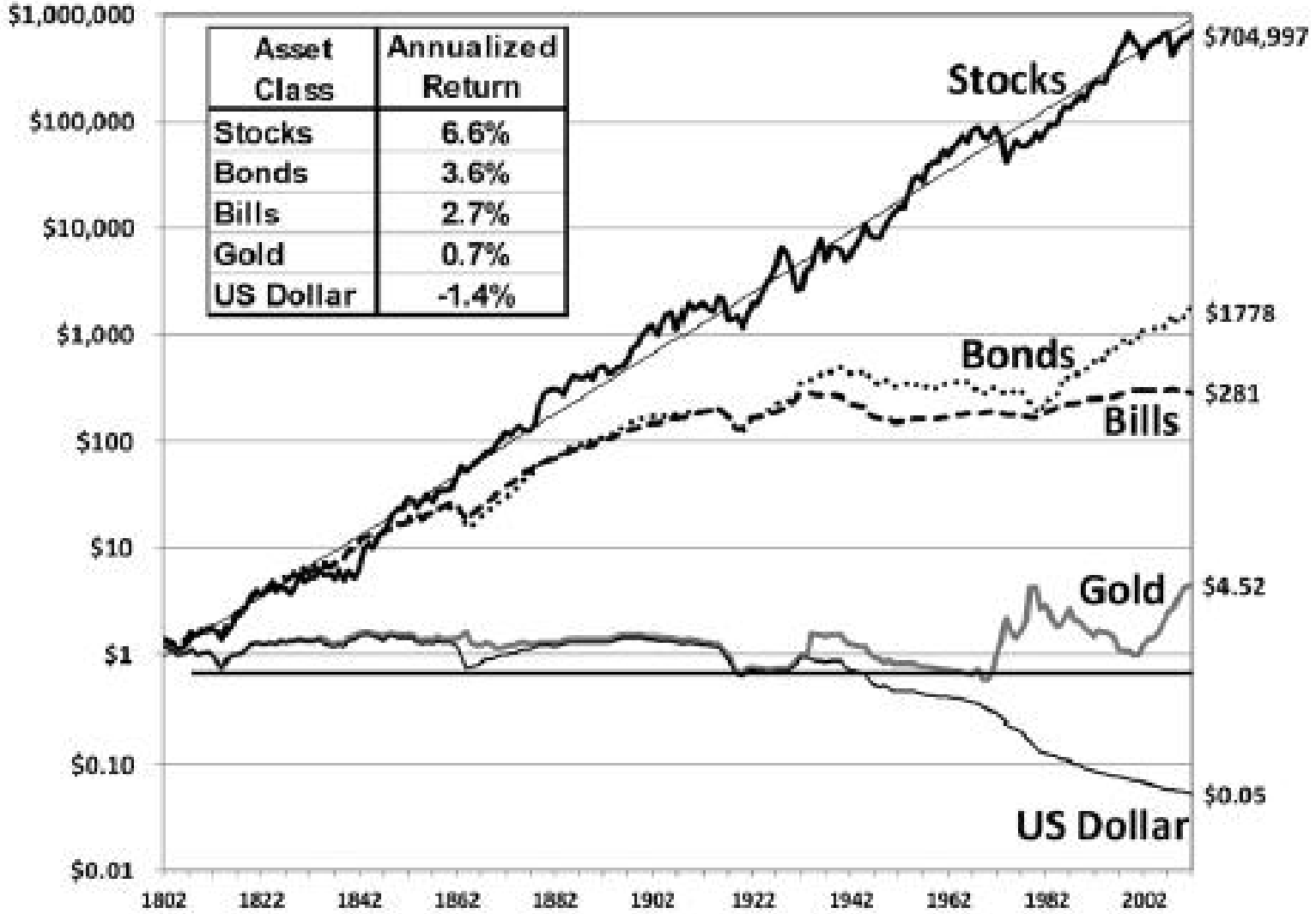


FIGURE 1-1
Total Nominal Return Indexes, 1802-1997

Some empirical facts. The pictures

FIGURE 1–1

Total Real Returns on U.S. Stocks, Bonds, Bills, Gold, and the Dollar, 1802–2012



Some empirical facts. The pictures

- In the following two slides you can see Shiller figure 1.1 in the first and last edition of "Irrational Exuberance"
- The figures give you the "real price index" compared with real earnings
- Notice the huge increase of price/earnings ratio since mid eighties
- In the third slide, from the 2015 edition of "Irrational Exuberance", you can see the evolution of P/E compared with the level of interest rates
- Do earnings justify prices?

Some empirical facts. The pictures



Figure 1.1

Stock Prices and Earnings, 1871–2000

Real (inflation-corrected) S&P Composite Stock Price Index, monthly, January 1871 through January 2000 (upper series), and real S&P Composite earnings (lower series), January 1871 to September 1999. *Source:* Author's calculations using data from S&P Statistical Service; U.S. Bureau of Labor

Some empirical facts. The pictures

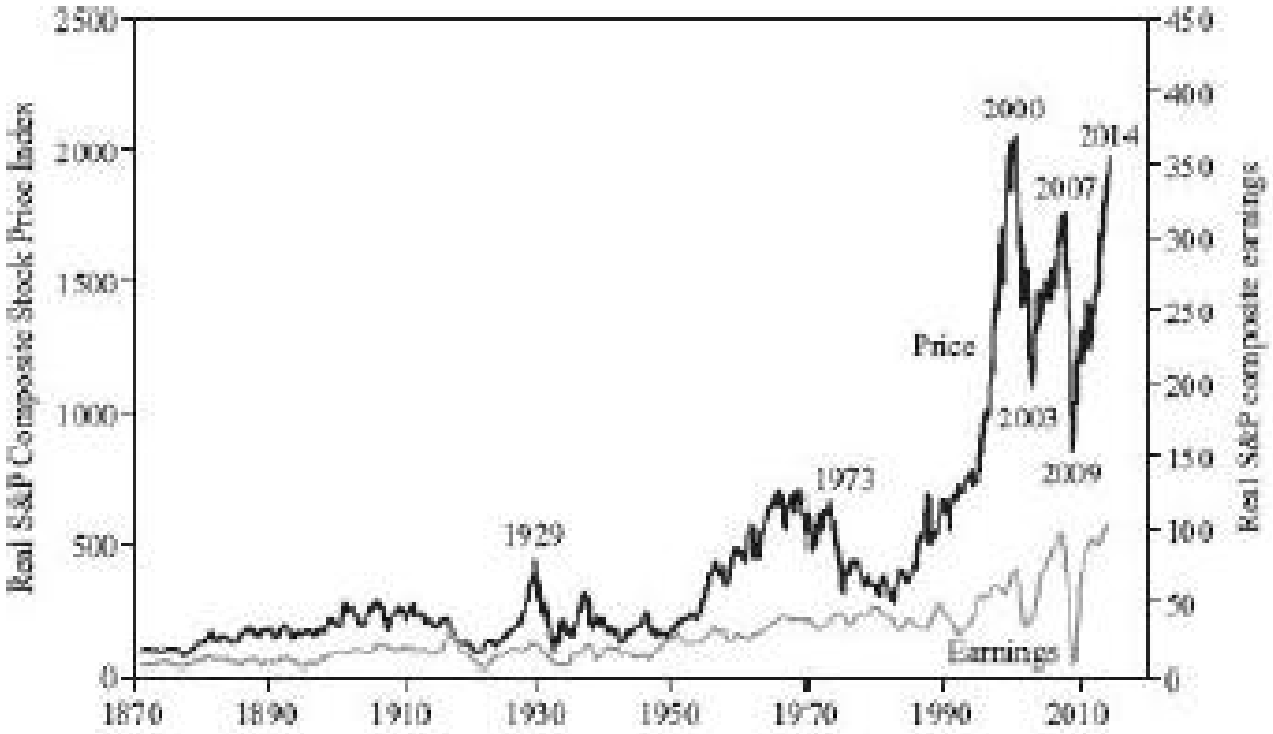


Figure 1.1

U.S. Stock Prices and Earnings, 1871–2014

Some empirical facts. The pictures

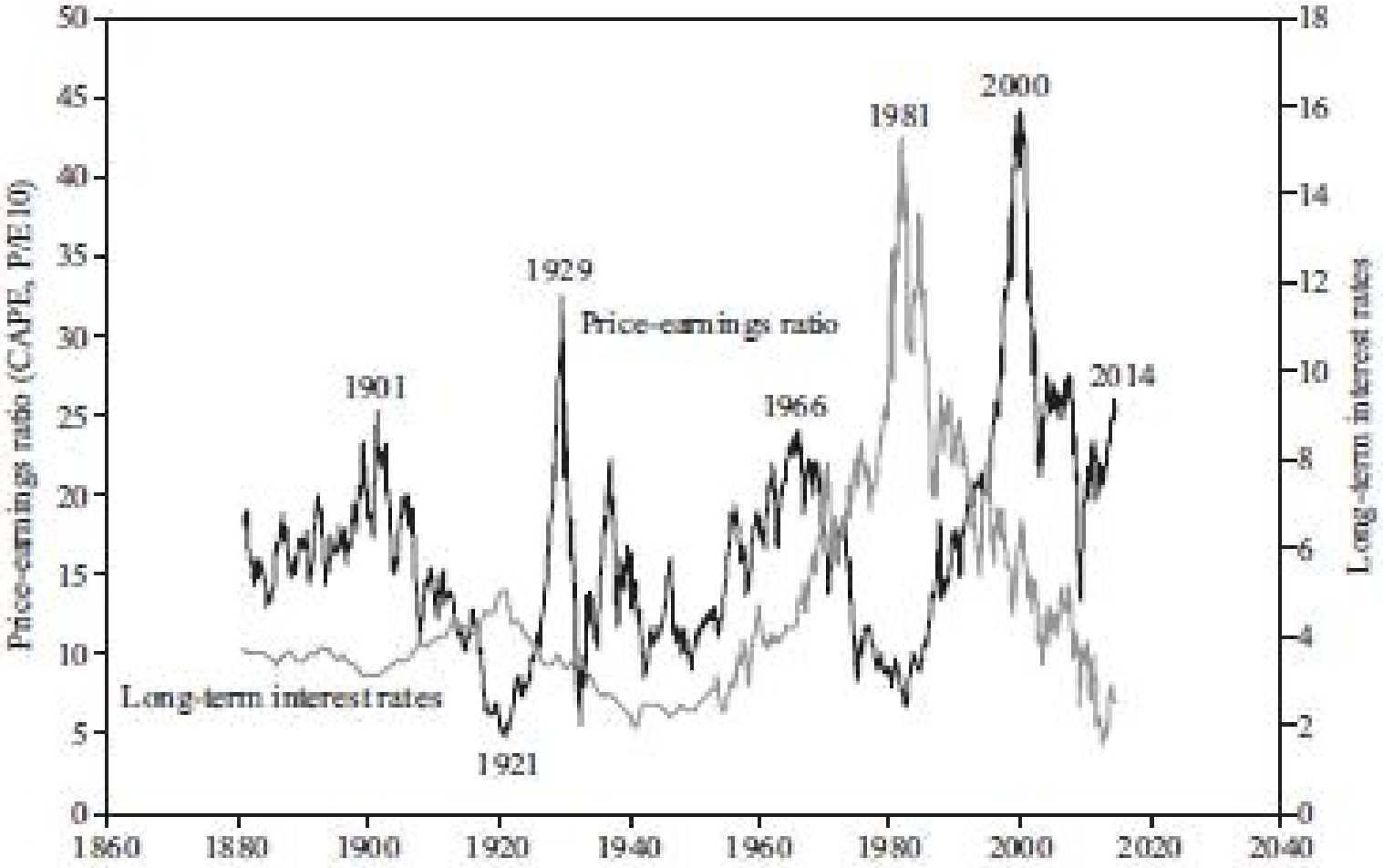


Figure 1.3

U.S. Cyclically Adjusted Price-Earnings Ratio (CAPE) and Interest Rates, 1881–2014

Price-earnings ratio, monthly, January 1881 to June 2014. Numerator: real (inflation-corrected) S&P Composite Stock Price Index. Denominator: moving

Some empirical facts. Real estate

- The housing market is a huge source of delusions, there is a widespread idea that to buy a house is a good long run investment
- Robert Shiller dedicated much care in dispelling such delusions (in doing this he created the Case-Shiller real estate index)

Some empirical facts. Real estate

- The following two slides are from the last edition of "Irrational Exuberance" (2015)
- In the first slide, you see the US home price index in real terms from 1890 compared with population, building costs and interest rates
- In the second slide, you see (from 1986) the Case Shiller index (based or resold property) and an index for new single family houses of similar kind
- In both indexes the 1998 2007 "bubble" is clear but in the second index the bubble is much smaller as in most cases it considers suburban homes

Some empirical facts. Real estate

- You should keep in mind that a house, contrary to bonds and stocks, in the long run has to be maintained and maybe rebuilt (in particular in suburban US)
- This implies costs that are not taken into account by the general index while they are partially taken into account by the Case Shiller index
- The index, in fact, implicitly considers the reduction in price due to necessary home improvement by the buyer
- On the other hand, improvements by the seller increase the price and the cost of these is not considered in the index

Some empirical facts. Real estate

- The overall trend is clear
- In 120 years the index increases of about 40% that is: about 0.3% per year (much less than population growth)
- (I repeat: this not considering maintaining costs)
- Moreover, most of the increase is concentrated in two episodes: post wwii and bubble to burst 1998-2015
- An update: in NOMINAL terms the Case Shiller index today is almost back to the bubble level of 2006
- Should we be worried?

Some empirical facts. Real estate

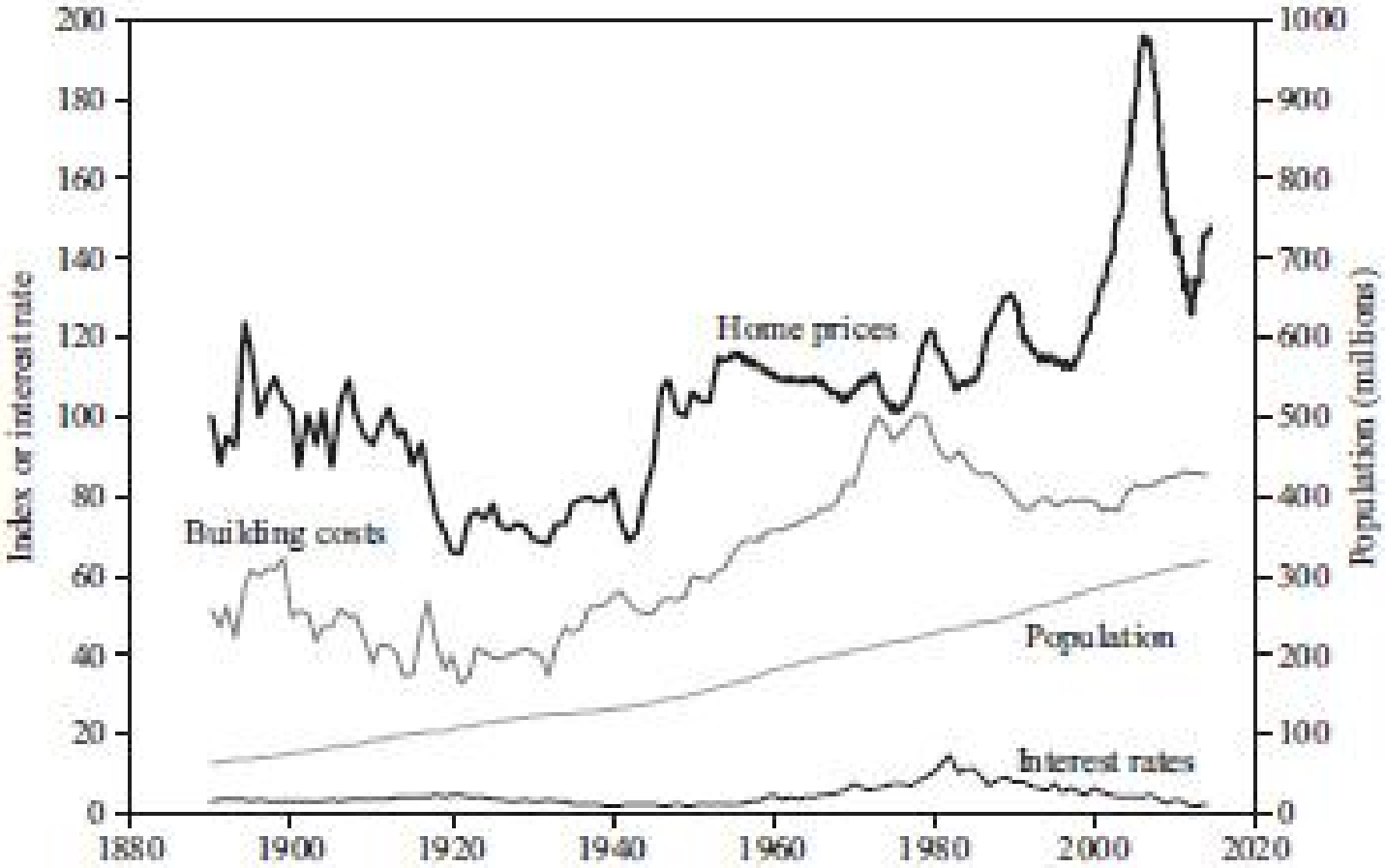


Figure 3.1
U.S. Home Prices, Building Costs, Population, and Interest Rates, 1890–2014

Some empirical facts. Real estate

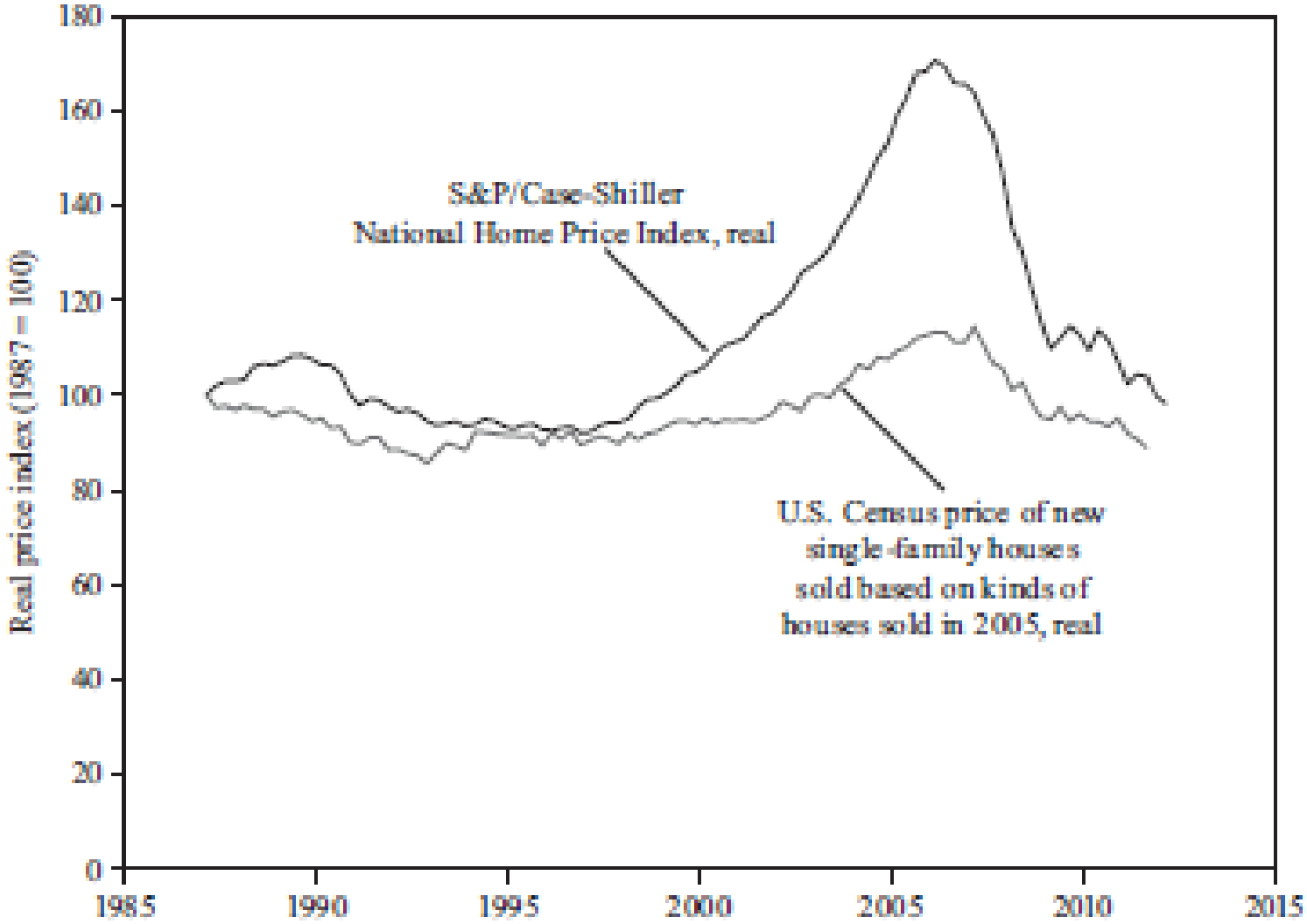


Figure A.4

Two Indices of U.S. Home Prices Divided by the Consumer Price Index (CPI-U), 1987–2013

Envoy: Very Big Numbers (2018/19) US Dollar Trillions

- Tot stock exch cap: USA 34/40 World 100
- Tot bond market USA 40/42 World 100 (about 1/2 of both are from public debt)
- Tot GDP Planet Earth 80/85 (USA 21)
- Tot debt Planet Earth 244 (USA 60/70). Households 46 (USA 14 Italy 1); non fin corp 73 (USA 10); gov 65 (USA 22 Italy 2.4); fin sect 60 (USA 14)
- Tot wealth Planet Earth 317 (1/3 USA 1/3 non fin)
- Tot derivatives (OTC) open interest 544 (value 10) (int rates 437 equity 6.5 exch rates 90)
- Wealth distrib: top 1% World 45% USA 35% ITA 25%. top 10% World 85% USA 75% ITA 55%

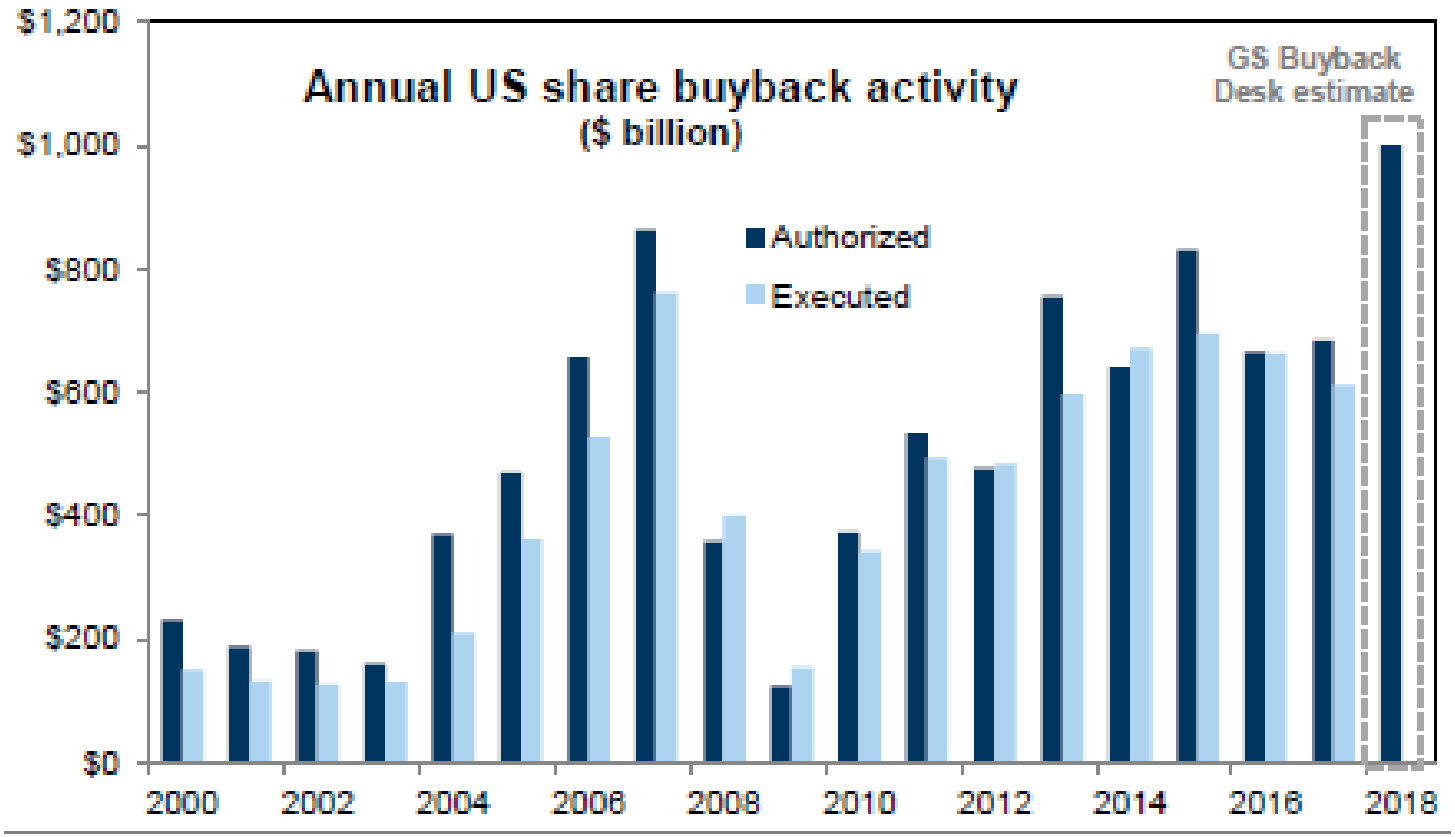
Some empirical facts. A recent worry

- Now, one example of the kind of themes you should be able to discuss (not for this course... for your job!)
- A worrisome evolution of US stock market in recent years is the fact that more and more demand for stocks comes from buybacks
- Buybacks are falling in the last few months
- These are financed by earnings (hence non reinvested) or, worse, by new debt (interest rates are low)
- The trend in US stock market is highly correlated with buybacks
- Sure, this is good for CEOs (increases earnings per share)
- Not sure for market stability

Some empirical facts. A recent worry

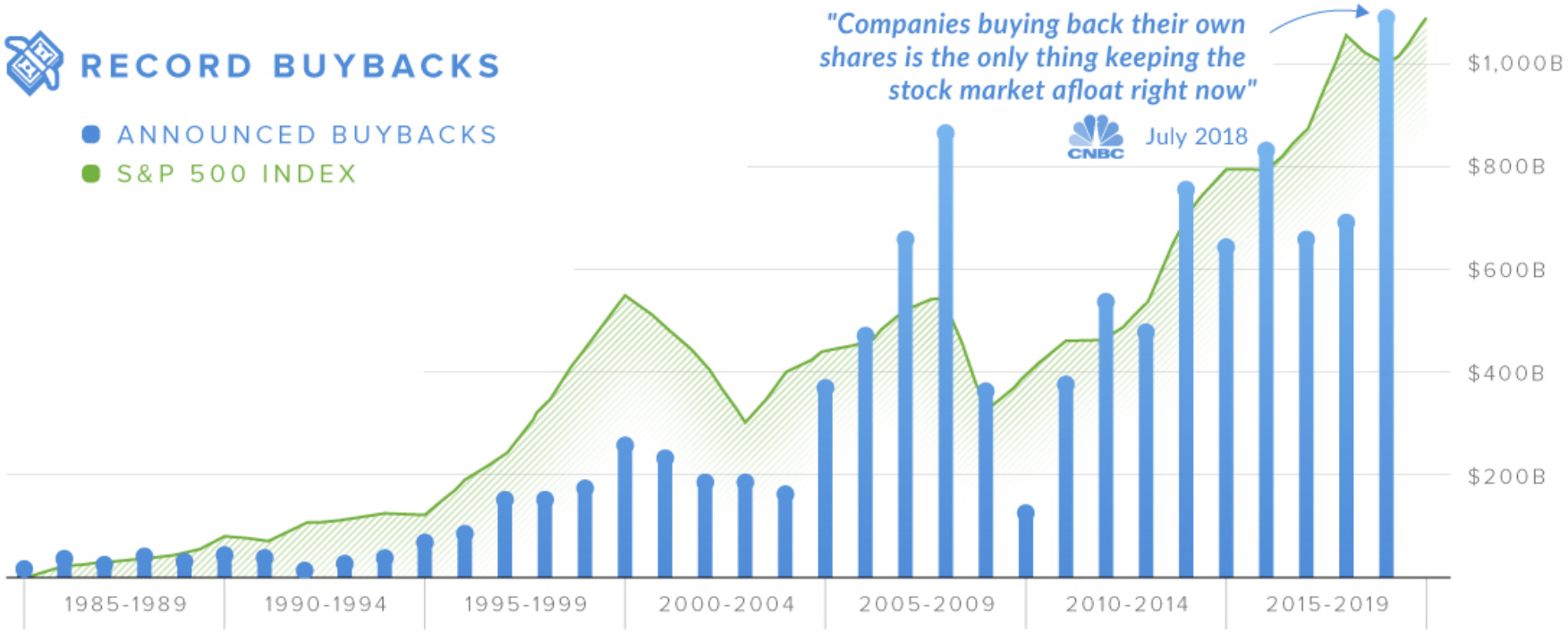
Exhibit 2: Buyback authorizations have surged in 2018

as of August 2, 2018



Source: Birinyi Associates, Goldman Sachs Global Investment Research

Some empirical facts. A recent worry



Some empirical facts. A recent worry

> Sometimes people perceive share repurchases as something that directly helps senior executives or people with high net worths [...] **But it also helps a lot of ordinary people who have their pension plans invested through mutual funds or institutional investors that invest in companies like ours.**



STEVEN KANDARIAN
METLIFE CEO

> Between 2008 and 2017, 466 of the S&P 500 companies spent around $4 trillion on stock buybacks, equal to 53% of profits. Another 30% of corporate profits went to dividends. **When more than 80% of corporate profits go to buybacks and dividends, there is reason to be concerned.**

CHUCK SCHUMER (D-N.Y.)
BERNIE SANDERS (D-VT.)





> In the investment world, discussions about share repurchases often become heated. But I'd suggest that participants in this debate take a deep breath: Assessing the desirability of repurchases isn't that complicated.

WARREN BUFFETT



Some empirical facts. A recent worry

Fall in share buybacks poses threat to US stocks

Increasingly wary companies spend less on repurchases for the second quarter in a row

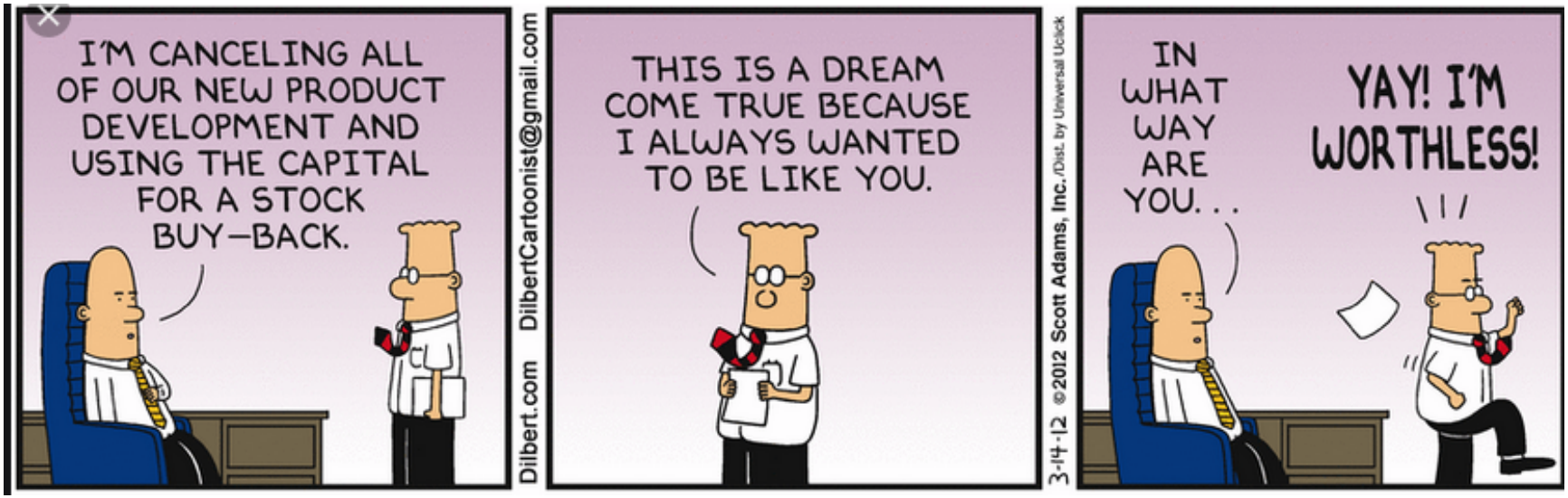
Richard Henderson in New York AUGUST 23, 2019 22

Big companies across America spent less on their own shares for a second quarter in a row, threatening a key pillar of support for the stock market that has helped to push indices to record highs.

Some empirical facts. A recent worry

Stock buybacks are neutral, in theory, for a company's value — every dollar handed back to shareholders is a dollar less in equity — but a reduction in the number of a company's shares outstanding increases earnings per share and boosts pay for managers with EPS-linked plans. Many companies also prefer the flexibility of a buyback programme, which can be cut or scrapped altogether, over a commitment to keep raising a dividend. Investors, too, often find buybacks more tax-efficient than dividends.

Some empirical facts. A recent worry



Lecture: Logarithmic Random Walk

Summary

- Log random walk: a simple (probability) model for (stock) prices over time written in log returns terms
- Precise specification of the model
- Time aggregation properties
- Empirical implications
- Monte Carlo simulations and caveat for long run implications of the model

Logarithmic random walk

- The (simple) log random walk (LRW) hypothesis on the evolution of prices states that, if we abstract from dividends and accruals, prices evolve approximately according to the stochastic difference equation:

$$\ln P_t = \ln P_{t-\Delta} + \varepsilon_t$$

- Here the "innovations" ε_t are assumed to be uncorrelated across time ($cov(\varepsilon_t; \varepsilon_{t'}) = 0 \quad \forall t \neq t'$), with constant expected value $\mu\Delta$ and constant variance $\sigma^2\Delta$.
- Sometimes, a further hypothesis is added and the ε_t are assumed to be jointly normally distributed.
- In this case the assumption of non correlation becomes equivalent to the assumption of independence.
- Since $\ln P_t - \ln P_{t-\Delta} = r_t^*$ the LRW is obviously equivalent to the assumption that log returns are uncorrelated random variables with constant expected value and variance both proportional to Δ.

Logarithmic random walk

- Notice that from the model assumptions we have:

$$P_t = P_{t-\Delta} e^{r_t^*} = P_{t-\Delta} e^{\varepsilon_t}$$

- So, if ε_t is assumed normal, P_t shall be lognormally distributed.

Logarithmic random walk

- A linear (that is: without logs) random walk in prices was sometimes considered in the earliest times of quantitative financial research (Louis Bachelier 1900), but it does not seem a good model for prices since a sequence of negative innovations may result in negative prices.
- Moreover, while the hypothesis of constant variance for (log) returns may be a good first order approximation of what we observe in markets, the same hypothesis for prices is not empirically sound:
- In general price changes tend to have a variance which is an increasing function of the price level.

Logarithmic random walk

- A couple of points to stress.
- First: Δ is the "fraction of time" over which the return is defined.
- This may be expressed in any unit of time measurement: $\Delta = 1$ may mean one year, one month, one day, at the choice of the user.
- However, care must be taken so that μ and σ^2 are assigned consistently with the choice of the unit of measurement of Δ.
- In fact μ and σ^2 represent expected value and variance of log return over an horizon of length $\Delta = 1$ and they shall be completely different if 1 means, say, one year or one day
- (See below for a particular convention in translating the values of μ and σ^2 between different units of measurement of time which is one of the consequences of the log random walk model).

Logarithmic random walk

- Second: suppose the model is valid for a time interval of length Δ and consider what happens over a time span of, say, 2Δ.
- By simply composing the model twice we have:

$$\ln P_t = \ln P_{t-2\Delta} + \varepsilon_t + \varepsilon_{t-\Delta} = \ln P_{t-2\Delta} + u_t$$

- Where $u_t = \varepsilon_t + \varepsilon_{t-\Delta}$.

Logarithmic random walk

- The model appears similar to the single Δ one and in fact it is, but it must be noticed that the u_t while uncorrelated (due to the hypothesis on the ε_t) on a time span of 2Δ shall indeed be correlated on a time span of Δ
- In fact: $Cov(u_t; u_{t-\Delta}) = Cov((\varepsilon_t + \varepsilon_{t-\Delta}); (\varepsilon_{t-\Delta} + \varepsilon_{t-2\Delta})) = Cov(\varepsilon_t; \varepsilon_{t-\Delta}) + Cov(\varepsilon_t; \varepsilon_{t-2\Delta}) + Cov(\varepsilon_{t-\Delta}; \varepsilon_{t-\Delta}) + Cov(\varepsilon_{t-\Delta}; \varepsilon_{t-2\Delta}) = Cov(\varepsilon_{t-\Delta}; \varepsilon_{t-\Delta}) = V(\varepsilon_{t-\Delta}) = \sigma^2\Delta$
- But: $Cov(u_t; u_{t-2\Delta}) = Cov((\varepsilon_t + \varepsilon_{t-\Delta}); (\varepsilon_{t-2\Delta} + \varepsilon_{t-3\Delta})) = Cov(\varepsilon_t; \varepsilon_{t-2\Delta}) + Cov(\varepsilon_t; \varepsilon_{t-3\Delta}) + Cov(\varepsilon_{t-\Delta}; \varepsilon_{t-2\Delta}) + Cov(\varepsilon_{t-\Delta}; \varepsilon_{t-3\Delta}) = 0$
- This means that the log random walk model can be aggregated over time, and maintain its properties, IF we "drop" the observations in between each aggregated interval
- (In this example if we keep every other observation)
- Notice that $E(u_t) = E(\varepsilon_t + \varepsilon_{t-\Delta}) = \mu 2\Delta$ and $V(u_t) = V(\varepsilon_t + \varepsilon_{t-\Delta}) = \sigma^2 2\Delta$

Logarithmic random walk

- The LRW is the traditional "beginners" model for the evolution of stock prices
- It is obviously a wrong model if we interpret it literally: prices are not dictated by "chance"
- Its main justification is based, as in other fields, on being it a probabilistic description of what macroscopically emerges from the aggregation of a very large number of individually negligible and uncohordinated (independent) shocks
- However, if we consider its probabilistic implications and compare these with the statistics we observe, It can be considered as a good first approximation descriptive model ("effective" would be the right word)

Logarithmic random walk

- In other words: its success does not depend on its interpretation of the actual process of price creation (it fails miserably)
- What it does, is to reproduce some "large scale" (i.e. Δ non too small) statistical properties of prices (average return, return standard deviation, return autocorrelation)
- These large scale properties only depend on a price being "moved" by a miriad of non cohordinated and each negligible shocks
- Such "ability to reproduce" is checked by comparing probabilities of events as given by the model with observed frequencies of such events.
- The model, even in this simple version, is then useful for introductory and simple purposes
- If we do go into details, the weight of empirical analysis has led most researchers to modify it in many ways

Logarithmic random walk

- While no consensus has been reached on an alternative standard model, there is a general agreement about several weak spots of the model:
- The fact that some sort of (very weak) dependence exists for today's returns on the full, or at least recent, history of returns.
- The constancy of the expected value and variance of the innovation term has been (strongly!) questioned.
- As hinted above, the idea of adding the Gaussian distribution hypothesis can be considered as useful just for teaching applications (this point shall be discussed in detail)

Logarithmic random walk

- In any case, the LRW still underlies many conventions regarding the presentation of market statistics.
- Moreover, the LRW is perhaps the most important justification for the commonly held equivalence between the intuitive term "volatility" and the statistical entity "standard deviation".

Logarithmic random walk

- An important example of the conventions influenced by the log random walk hypothesis, is the "annualization" of expected value and variance.
- We are used to the fact that, often, the rate of return of an investment over a given time period be reported in an "annualized" way.
- The precise conversion from a period rate to a yearly rate depends on accrual conventions.
- For instance, for an investment of less that one year length, the most frequent convention is to multiply the period rate times the ratio between the (properly measured according to the relevant accrual conventions) length of one year and the length of the investment.
- So, for instance, if we have an investments which lasts three months and yields a rate of 1% in these three months, the rate on an yearly basis shall be something like 4%.

Logarithmic random walk

- It is clear that this is just a convention: the rate for an investment of one year in length shall NOT, in general, be equal to 4%, this is just the annualized rate for our three months investment.
- This ceases to be just a convention, for instance, if the term structure of interest rates is constant.
- However, such a convention can be useful for comparison across investment horizons.
- (Beware such comparisons: they are dangerous!)

Logarithmic random walk

- In a similar way, when we speak of the expected return or the standard deviation/variance of an investment, it is common to report these in an annualized way, even if we speak of returns for periods of less or of more than one year.
- The actual annualization procedure is based on a convention which is very similar to the one used in the case of interest rates.
- As in this case the convention is "true", that is: annualized values of expected value and variance correspond to per annum expected values and variances, only in particular cases.
- The specific particular case on which the convention used in practice is the LRW hypothesis with constant expected value and variance.

Logarithmic random walk

- If we assume the LRW and consider a sequence of n log returns r_t^* at times $t, t-1, t-2, ..., t-n+1$ (just for the sake of simplicity in notation we suppose each time interval Δ to be of length 1 and drop the generic Δ) we have that:

$$E(r^*_{t-n,t}) = E(\sum_{i=0,...,n-1} r^*_{t-i}) = \sum_{i=0,...,n-1} E(r^*_{t-i}) = n\mu$$

$$Var(r^*_{t-n,t}) = Var(\sum_{i=0,...,n-1} r^*_{t-i}) = \sum_{i=0,...,n-1} Var(r^*_{t-i}) = n\sigma^2$$

Logarithmic random walk

- This obvious result, which is a direct consequence of the assumption of constant expected value and variance and of non correlation of innovations at different times, is typically applied, for annualization purposes, also when the LRW is not considered to be valid.
- So, for instance, given an evaluation of σ^2 on daily data, this evaluation is annualized multiplying it, say, by 256 (or any number representing open market days, different ones exist), it is put on a monthly basis by multiplying it by, say, 25 and on a weekly basis by multiplying it by, usually, 5.
- (Comment on the implications of these conventions: volatility and open market days)

Logarithmic random walk

- As we stressed before this is not a convention, but the correct procedure, IF the LRW model holds (provided we suppose it acts on open market days only).
- In this case, in fact, the variance over n time periods is equal to n times the variance over one time period.
- If the LRW model is not believed to hold, for instance, if the expected value and-or the variance of return is not constant over time or if we have correlation among the ε_t, this procedure shall still be applied but just as a convention.

Logarithmic random walk

- The fact that, under the LRW, the expected value grows linearly with the length of the time period while the standard deviation (square root of the variance) grows with the square root of the number of observations, has created a lot of discussion about the existence of some time horizon beyond which it is always proper to hold a stock portfolio.
- This problem, conventionally called "time diversification", and more popularly "stocks for the long run" (not in Siegel's sense), has been discussed at length both on the positive (commonly sustained by fund managers) and the negative side (more rooted in academia: Paul Samuelson was a non negligible opponent of the idea) we shall consider it in the next section.

Logarithmic random walk. What does it imply empirically?

- To have an idea of the empirical implications of the LRW hypothesis for returns (plus that of Gaussian distribution), we plot in the following figures an aggregated, total return, index of the US stock market in the 20th century together with 100 simulations describing possible alternate histories of the US market in the same period
- We suppose that the index evolution follows a (Gaussian) LRW
- The yearly log return expected value and standard deviation are identical to the historical average and standard deviation (20th century only)
- These are: resp. 5.36% (we consider just the excess return) and 18,1% (the use of %, as usual with log returns, is quite improper if common).

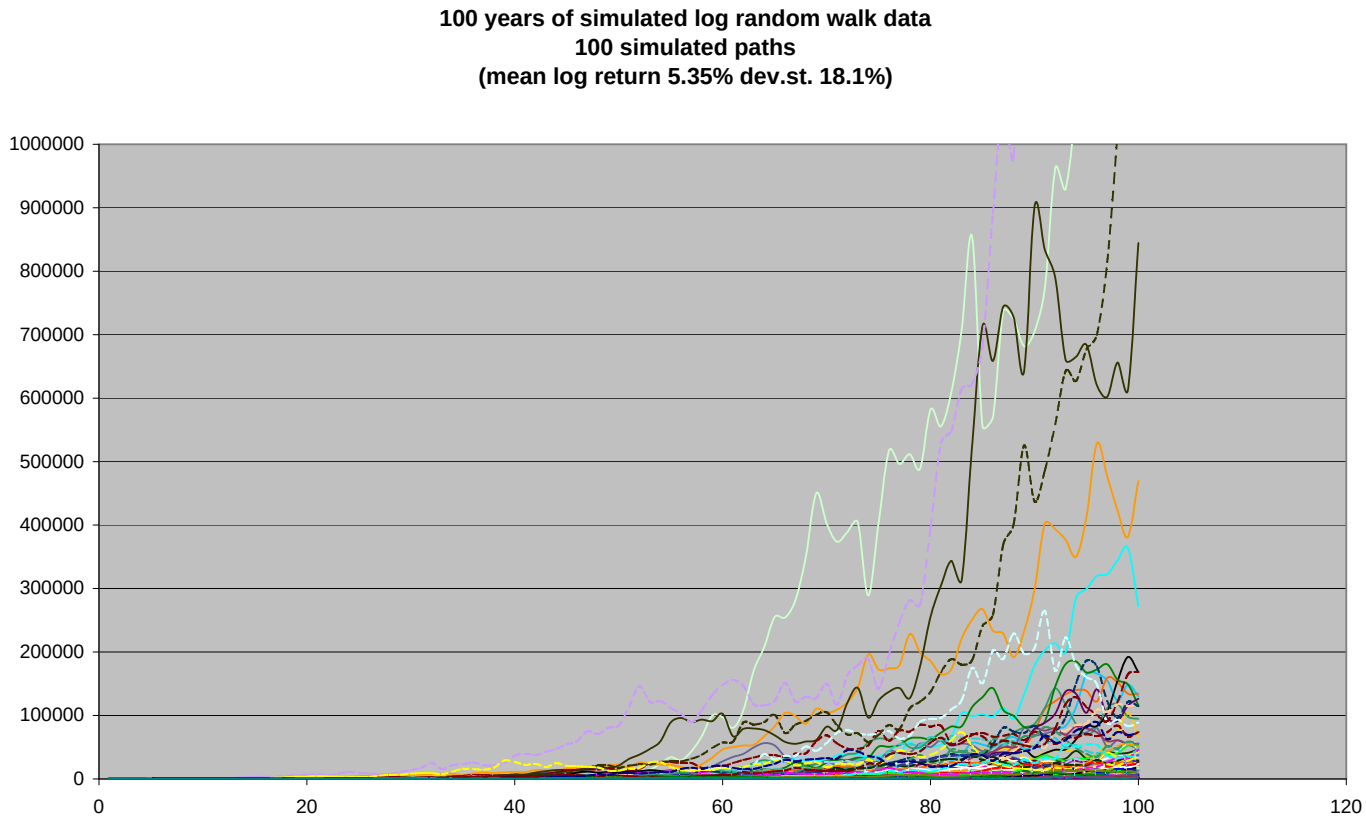
Logarithmic random walk. What does it imply empirically?

- Notice (and once noted, forget) that, according to the results of the previous lecture, the idea of putting together growth due to prices and due to dividents is not fully justified
- The initial value of the index is 100.
- The black line is the true evolution of the index in the 20th century.

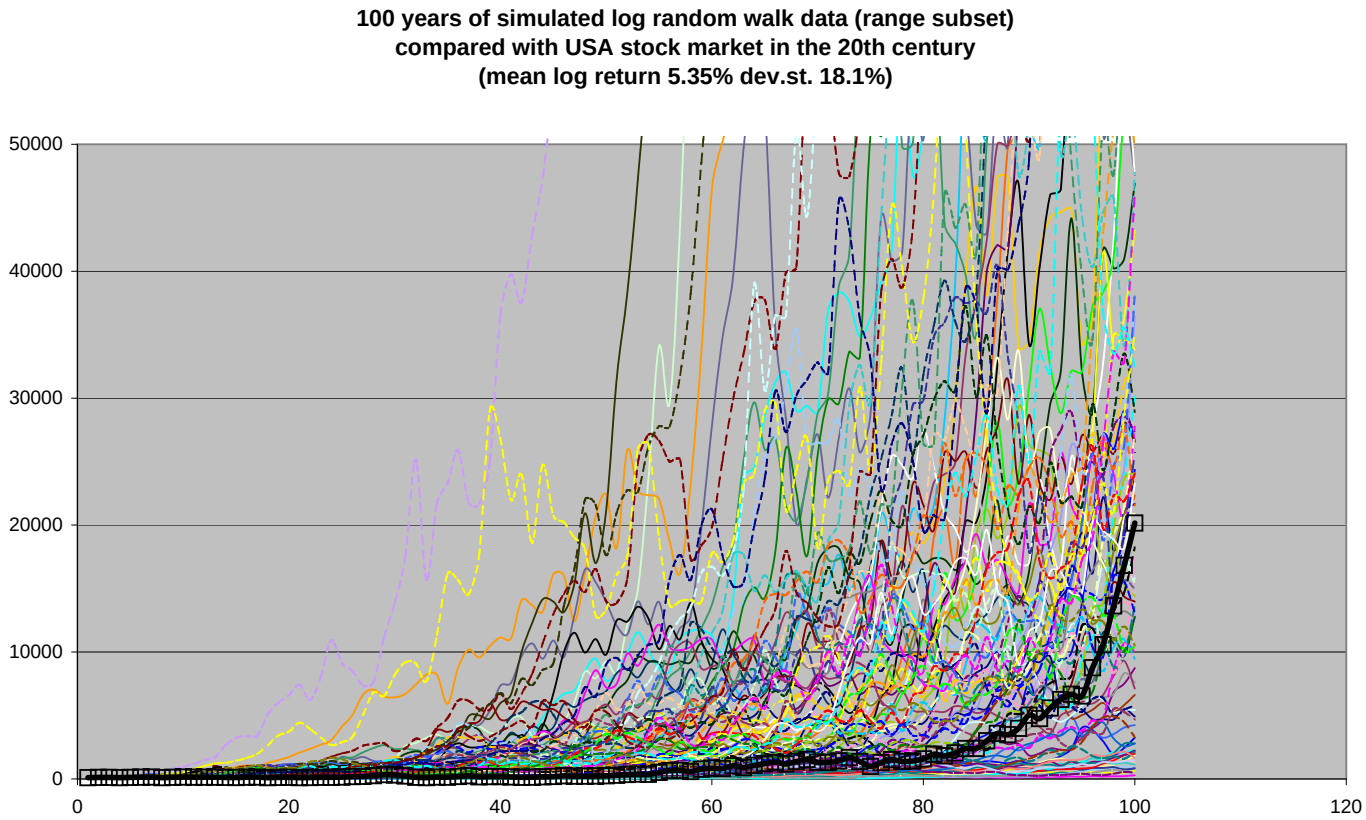
Logarithmic random walk. What does it imply empirically?

- Here to "simulate" means this:
- For each trajectory we (pseudo) randomly draw (using Excel) 100 "Gaussian returns", one per year
- This are built so that they are realizations of uncorrelated (hence independent) Gaussian random variables of expected value .0536 and standard deviation .181
- We then plot 100 times the exponential of the first of this, then 100 times the exponentiatial of the sum of the first and the second... and so on, up to the last one
- This is a trajectory
- We build 100 of these trajectories

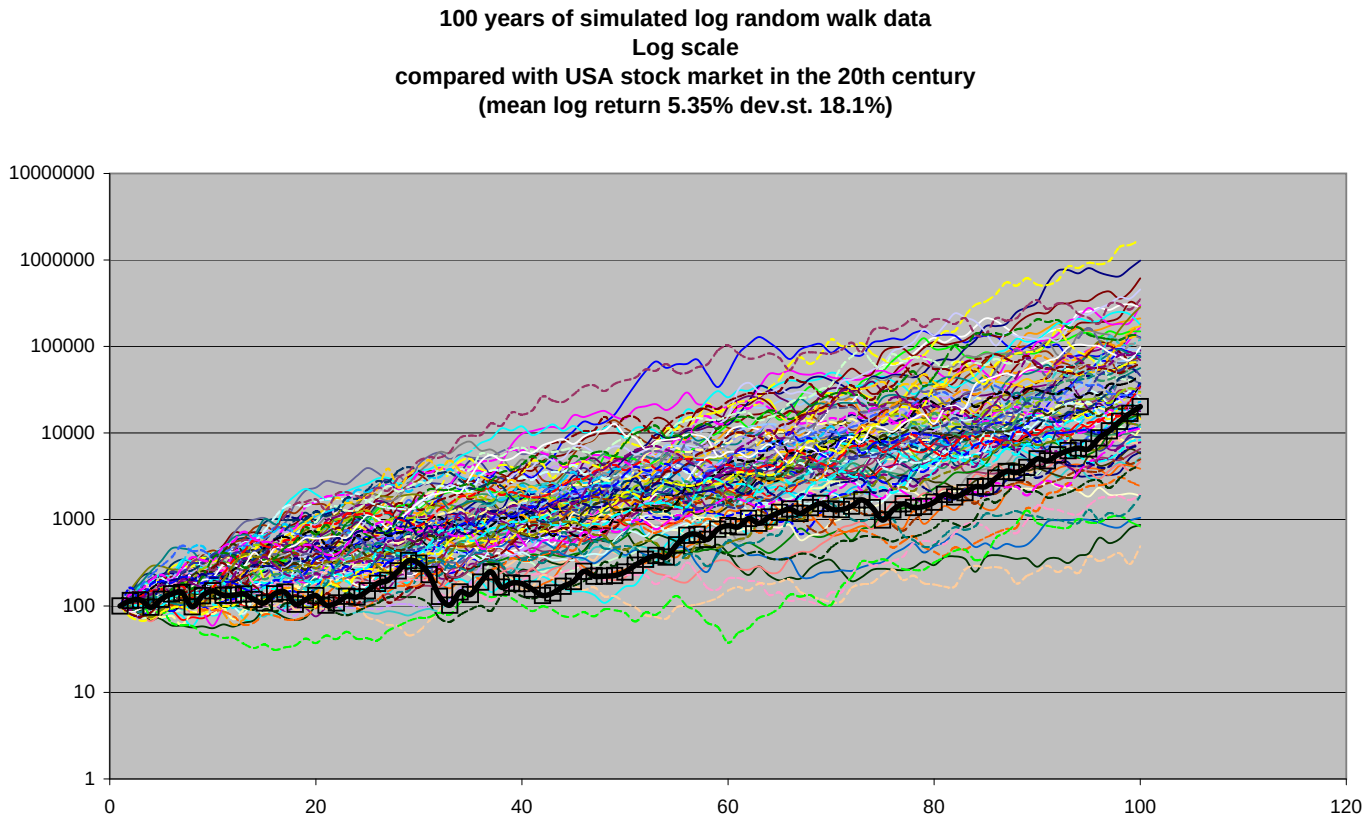
Logarithmic random walk. Index evolution from 100



Logarithmic random walk. Focus below 50,000



Logarithmic random walk. Index in log scale



Logarithmic random walk

- Data is presented both in price scale (starting value 100) and in log price scale.
- The reason is simple.
- Consider the distribution of log return after 100 year under our hypothesis.
- This is going to be the distribution of the sum of 100 iid Gaussian RV each with expected value of 5.36% (again a wrong %!) and standard deviation 18.1%
- The distribution of the "percentage" change between the initial and the terminal price, shall be Gaussian with expected value 5.36% times 100 equal to 536% and standard deviation18.1% times 10 equal to 181%.

Logarithmic random walk

- So, a standard $\pm 2\sigma$ interval (recall Tchebiceff) for the terminal value of this sum is $536\% \pm 362\%$
- In price terms, if we start from 100, we get $100e^{5.36 \pm 3.62}$ that is, an interval with lower extreme 569 ($= 100e^{5.36-3.62} = 100e^{1.74} = 569.7$) and upper extreme 794263 ($= 100e^{5.36+3.62} = 100e^{8.98} = 794263.2$).
- Notice that the expected log return (for a century) that is: $5.36 = .0536 * 100$ (which is also the median of the log return distribution, due to the Gaussian hypothesis) corresponds to a terminal price of $100e^{5.36} = 21270$
- This is the median of the price distribution (due to the monotonic and continuous transform from log return to prices)
- This is NOT the expected value of the price distribution as the transform from log returns to prices is not linear

Logarithmic random walk

- The expected value of the exponential of a Gaussian random variable, with expected value μ and variance σ^2, (this is a Lognormal), is $e^{\mu+\frac{1}{2}\sigma^2}$
- This is bigger than the median: e^{μ} of an amount which increases with the variance
- using the above values for μ and σ^2, he expected value of the price after one century of a stock of initial value 100 is 109449 that is $100e^{5.36+\frac{1}{2}1.81^2}$ (roughly five times the median)

Logarithmic random walk

- The striking fact in the plots is that under our hypotheses the possible histories can be wildly different (and wildly differ from the actual one).
- No problem in this if we recall the unconditional nature of the model.
- I should say: no problem if we understand the meaning of the sentence: "No problem in this if we recall the unconditional nature of the model"
- So: let's try and understand this very relevant point

Logarithmic random walk

- We only simulated 100 possible histories.
- Even with such a limited sample we have a top terminal price of more than 2.000.000, and increase of more that 2.000.000/100=20.000 times
- (This in a very lucky, for long investors, world. We wonder what studying Finance would be in such a world...or what such world would be)
- and a bottom terminal price frequently below 100
- (again: in a world so unlucky that, had we lived in it, probably we would not talk about the stock market)

Logarithmic random walk

- Both these extremes are possible, a priori, in a Gaussian log random walk world with expected log return .0536 and standard deviation .181
- Notice: we know everything, model and parameters, this is market uncertainty only!
- However, had such extremes been observed, our evaluation of expected log return and oflog return standard deviation, probably, would be quite different
- Think about this! Go back one century, the trajectory of "our world" would be one of the extremes, at that time?

Logarithmic random walk

- Compare this with the Siegel-Shiller data we discussed in section, then think about the result of our simulation in such extreme worlds.
- For instance, suppose the "true" pat were the worst path.
- Suppose we live in such a world. The simulations I would show you in this possible world (in a cold, candle lighted room, computing by hand and chalk on a blackboard), provided you and I were still interested in this topic, and not in making a life out of catching squirrels, would be quite different that what you see here.
- In other words: this possible story is a result totally compatible (under Gaussian LRW) with what we did actually see in our real history.
- Again: spend a little time thinking about this point.
- It could be "illuminating".

Logarithmic random walk

- Think now to the economic sustainability of such extreme worlds: such extreme market behaviours of the stock market cannot happen by themselves
- This is not the plot of winnings by some lucky or unlucky casino guy, it is the market value of an economy, which should sustain such values, provided investors are not totally bum
- While trajectories far from the actual one are well possible, these would correspond to a complete different history, not only a different market history.
- This is the meaning of "unconditional" or "marginal"
- If we conditioned the evolution of the stock market to happen in the "same given historical background" the result would not show such variability, but this is not the "random walk"
- More on this in a moment.

Logarithmic random walk

- Before going on, remember again that all this comes from the analysis of the stock market in a very, up to now, successful country: the USA.
- We analyze this so much, also because it was successful (and so, for instance, most Finance schools, journals and researchers are USA based).
- It could well be that what we observe now would be an extreme positive possible history, as seen in 1900
- This biases our conclusions, if we wish to apply such conclusions to the rest of the world or, even, to the future of USA.
- Maybe, a more balanced view could be gained by comparing this result with the evolution of stock markets all around the world
- (Robert J. Barro did this in: "Rare Disasters and Asset Markets in the Twentieth Century." (2006) Quarterly Journal of Economics, 121(3): 823–66.)

Logarithmic random walk

- Let's go back to the origins of the puzzling heterogeneity of "possible histories".
- This is an immediate consequence of the log random walk hypothesis.
- If we estimate μ and σ out of a long series of data (one century) we are using data from a very heterogeneous set of economic and historical conditions.
- Then we use this number in order to simulate "possible histories" without conditioning to any particular evolution of historical or economical variables which could and shall influence the stock price.
- In other words we are using the log random walk model as a "unconditional" model.

Logarithmic random walk

- That is: it is unconditional to everything you may know or suppose about the evolution of other variables connected with the evolution of the modeled stock price.
- This point is quite relevant if we wish to understand the sometimes surprising implications of this simple model.
- In the above example, according to the model and the historically estimated parameters we get the mean $\pm$ 2σ interval 5.36 ± 3.62, or, in price terms, $100e^{5.36 \pm 3.62}$ that is: an interval with lower extreme 569 and upper extreme 794263.

Logarithmic random walk

- It must be clear that such a wide set of histories are possible, with non negligible probability, only because we did assume nothing on the (century long) evolution of all the variables that shall influence prices.
- Only under this "ignorance" assumption such an heterogeneous set of trajectories can have non negligible probability.
- If we are puzzled by the result this is because, while the model describes the possible evolution of prices "in whatever conditions", unconditional to anything...
- (in fact, we estimate expected return and standard deviation using a long history, during which many different things happened)
- ...when we see the implications of the model we, almost invariably, shall be conditioned by our recent memories and our recall of recent events

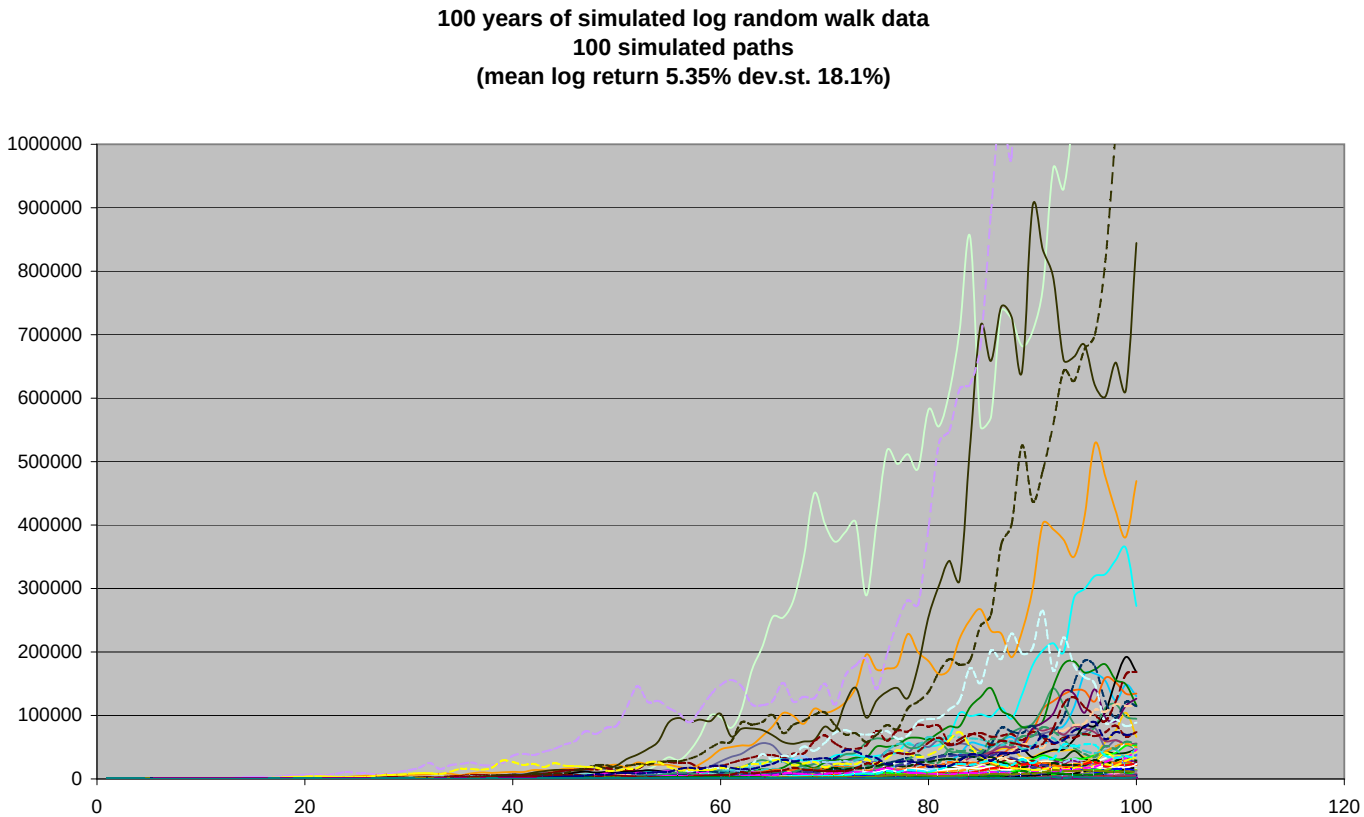
Logarithmic random walk

- Unconsciously, we shall make some hypothesis on the future as, for instance, the fact that economic growth shall be, on average, similar to what recently seen
- This misunderstanding is what makes the result "surprising" and maybe "preposterous"
- The estimates of $\mu$and σ we use (or even our assumption of zero correlation of log returns and, more in general, the structure of the model itself which contains no other variables but a single price) are NOT conditional to such (implicit) hypotheses
- In fact, if we understand the model unconditional nature, the wide variation bounds isn by no menas surprising

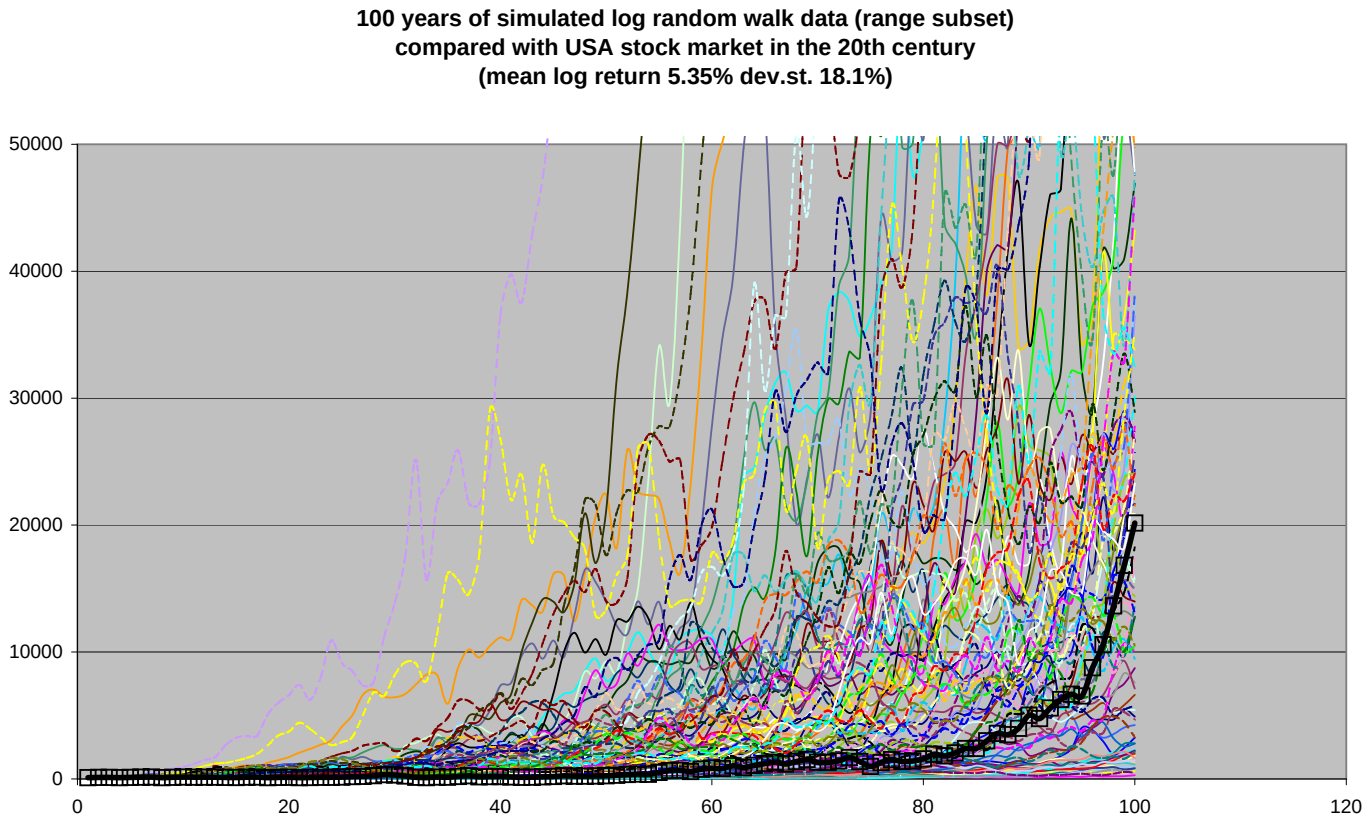
Logarithmic random walk

- This misunderstanding is quite common and it is to be always kept in mind when discussing results of the applications of the log random walk model.
- There exists a wide body of literature, both from the applied (macro funds) and the academic sides, that suggests ways for "conditioning" the model.
- This shall not be considered in this course.
- This range in terminal prices is so wide that, in order to make our real price history visible, we had to cut a slice from the range of prices.
- Let us conclude by plotting the paths again

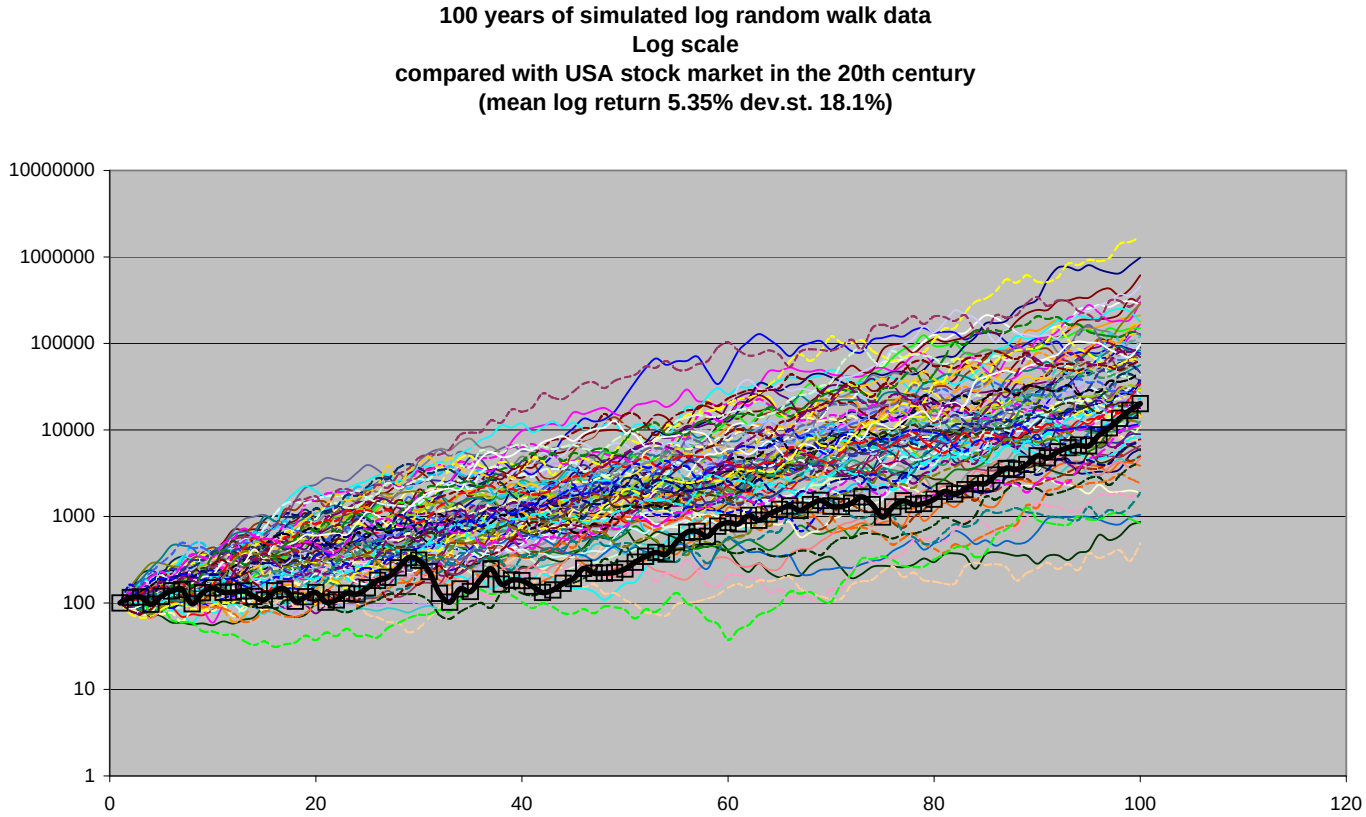
Logarithmic random walk



Logarithmic random walk



Logarithmic random walk



Lecture: Stocks for the Long Run

Summary

- Arguments about the profitability of stocks in the long run
- Three different arguments: not to be confused
- Two implications of log random walk
- Probability of loss going to zero; "time diversification"
- Siegel's historical argument (and its "ex post" stance)

Stocks for the long run

- This is a very interesting and popular topic, part of the lore of the financial milieu.
- A short discussion shall be useful to clarify some issues connected with the LRW hypothesis together with some implicit assumption underlying much financial advertising.

Stocks for the long run

- We have three flavors of the "stocks for the long run" argument.
- The first and the second are a priori arguments, depending on the log random walk hypothesis or something equivalent to it.
- The third is an a posteriori argument based on historical data (basically: Siegel's analysis).
- It is quite important to have a clear idea of the different weight and meaning of these arguments.
- In fact, most of the "puzzling" statements you may find about advertising "stock for the long run" by the investment industry, depend on a wrong "mix" of the arguments

Stocks for the long run

- First version:
- The basic idea of the first version of the argument can be sketched as follows.
- Suppose single period (log) (excess) returns have (positive) expected value μ and variance σ^2 (both constant over time).
- Moreover, suppose for simplicity that the investor requires a "Sharpe ratio" of say S out of his-her investment.
- Under the above hypotheses, plus the log random walk hypothesis, the Sharpe ratio over n time periods is given by

$$\frac{n\mu}{\sqrt{n}\sigma} = \sqrt{n}\frac{\mu}{\sigma}$$

- In words: if μ is positive, any positive value of the Sharpe ratio can be reached changing n

Stocks for the long run

- Another way of phrasing the same argument, equivalent to the first if we add the hypothesis of Gaussian returns is:
- The probability of the investment to yield a n periods return greater than

$$n\mu - \sqrt{n} z_{1-\alpha} \sigma$$

- is equal to $1 - \alpha$ (whatever the α you choose). That is

$$Pr(r_n^* > n\mu - \sqrt{n} z_{1-\alpha} \sigma) = 1 - \alpha$$

Stocks for the long run

- Proof: (if necessary).
- According to our hypotheses the n period log return r_n^* is distributed according to a Gaussian with $E(r_n^*) = n\mu$ and standard deviation $\sqrt{V(r_n^*)} = \sqrt{n}\sigma$
- If we standardize r_n^* we get

$$Pr(\frac{r_n^* - n\mu}{\sqrt{n}\sigma} > -z_{1-\alpha}) = 1 - \alpha$$

- Where $z_{1-\alpha}$ is the quantile of the standard Gaussian which leaves on its left a probability of $1-\alpha$ (hence, $-z_{1-\alpha}$ leaves the same probability on its right)
- But, rearranging the terms, this is equal to

$$Pr(r_n^* > n\mu - \sqrt{n}z_{1-\alpha}\sigma) = 1 - \alpha$$

Stocks for the long run

- The first derivative of $n\mu - \sqrt{n}z_{1-\alpha}\sigma$ w.r.t. $\sqrt{n}$ is $2\sqrt{n}\mu - z_{1-\alpha}\sigma$ and this is positive for

$$\sqrt{n} > \frac{1}{2}\frac{z_{1-\alpha}}{\mu}\sigma$$

- (In yearly terms we expect, from historical data, σ/μ between 3 for an index and and 5 to 6 for a single stock. So, if α is 2.5%, $z_{1-\alpha} = 1.96$, the rhs is from 3 to 6 and n should be bigger than 9 to 36 years)
- When this inequality is satisfied, the boundary $n\mu - \sqrt{n}z_{1-\alpha}\sigma$ is an increasing and unbounded function of n
- Hence, for any α and any chosen value C, there exists a n^* such that, for $n > n^*$, the probability for an n period return less than C is less than α
- In formulas: $\forall C$, $\exists n^*$ such that $\forall n > n^*$ we have:

$$n\mu - \sqrt{n}z_{1-\alpha}\sigma > C$$

Stocks for the long run

- The suggestion could be: suppose there is an investment with high expected (excess) return which you may fear because its volatility is very high
- In this case, if your time horizon is of an undetermined number n of years, make this investment over a (long) time orizon which gives you a Sharpe ratio high enough to quell your fear
- Even if this investment may seem too risky in the "short run" there is always a time horizon so that for that horizon, the probability of any given loss is as small as you like or, that is the same, the Sharpe ratio as big as you like.
- Typically, such high return (and high volatility) investment are stocks, so: "stocks for the long run".

Stocks for the long run

- The value of n for which this lower bound crosses a given C level (n^*) is the smallest integer n such that

$$n\mu - \sqrt{n} z_{1-\alpha}\sigma \geq C$$

- In particular, for $C = 0$ (meaning: no loss in the investment) the solution is

$$\sqrt{n} \geq \frac{z_{1-\alpha}\sigma}{\mu}$$

- With the above numbers, this requires a n greater than 36, for an index, to 144 years, for a single stock, quite a long wait!
- Notice that, even in the long run, and not surprisingly, a well diversified portfolio is better

Stocks for the long run

- This argument is mathematically correct
- Is it also relevant?

Stocks for the long run

- This depends on the investor's criterion for choosing investments
- This, for instance, could be the full period expected return given some probability of a given loss
- The Sharpe ratio for the full n periods
- The per period Sharpe ratio (which obviously is a constant)
- The total volatility over the full period of investment (which obviously increases without bounds), and so on
- In any case we should keep in mind that it is always difficult to compare investments over different time periods

Stocks for the long run

- We can, however, critique such suggestion on the basis of its implications
- For instance, a typical critique to the suggestion is phrased like this (this is a version of a critique by Paul Samuelsons):
- "Why should we consider as proper (in terms of risk) a given investment for n time periods if we do not consider it proper (in terms of risk) for each single one of those periods?"
- This is not a critique to the mathematical result, it express doubts about the picture of an investor who takes into account the per period Sharpe ratio or some measure of probable loss and expected return per period.
- In other words the critique is a statement about the fact that, very reasonably, we believe the investor does not consider as equivalent investments with identical sharp ratios but extending over different time spans.

Stocks for the long run

- Another frequent critique is: "It is true: the expected value of the investment increases without bounds but so does its volatility"
- So: "In the end, over the long run I am, in absolute terms, more and more uncertain on my investment result"
- (The mean-standard deviation ratio goes up only because the numerator grows faster than the denominator but the denominator grows in unbounded fashion).
- Again: this is reasonable as a critique if we believe the investor to decide on the basis of the absolute volatility of the investment over the full time period.

Stocks for the long run

- In a nutshell: acceptance or refusal, on an a priori basis, of this first version of the argument depend on how we model the investor's decision making
- However, this version of the argument, in general, cannot be labeled as "wrong": there may be a point in it, at the least if you are a very peculiar and, maybe, "strange" kind of investor.

Stocks for the long run

- Second version
- The second version of the argument, again based on the log random walk hypothesis, is formally correct but its standard interpretation is a real fallacy
- That is: it is impossible to justify it in any reasonable way
- This is the so called "time diversification" argument
- This is very popular in financial advertising and related literature
- The error, as usual, is not in the maths but in its interpretation

Stocks for the long run

- There is an enticing similitude, under the log random walk hypothesis, between:

1. The return of an investment for one year in, say 10, UCORRELATED securities with identical expected returns and volatilities (this last hypothesis is just for simplicity: the argument can be extended to different expected returns and volatilities)
2. The AVERAGE per YEAR return of a 10 years investment in a single security with the same expected value and volatility.

- (Notice the caps)

Stocks for the long run

- To be precise, in order for the result to hold we must forget the difference between additive properties of linear and log returns.
- Moreover the comparison implicitly requires zero interest rates.
- But let's accept these approximations.
- It is a sound way to proceed provided the user is able to understand the cases where his-her "approximations" do not work.

- Keeping this into account and using (as an approximation) log returns overall:
- The expected return and standard deviation for the return corresponding to the first strategy are:

$$E(r_\pi^*) \cong E(\sum_{i=1}^{10} \frac{1}{10} r_i^*) = \mu$$

and

$$\sqrt{V(r_\pi^*)} \cong \sqrt{V(\sum_{i=1}^{10} \frac{1}{10} r_i^*)} = \frac{\sigma}{\sqrt{10}}$$

- Note: the $\cong$ sign is there because r_π^* is only approximatively equal to $\sum_{i=1}^{10} \frac{1}{10} r_i^*$(the equality would require linear returns)

Stocks for the long run

- These are (if we abstract from the approximation) the same as the expected value and standard deviation for the "average per year" return of the second strategy:

$$E(\sum_{t=1}^{10} r_t^*/10) = \mu$$

and

$$\sqrt{V(\sum_{t=1}^{10} r_i^*/10)} = \frac{\sigma}{\sqrt{10}}$$

Stocks for the long run

- Beware!
- We should be wary from the beginning in accepting such comparisons:
- The two investments cannot be directly compared, since they are investments of the same amount but on different time periods.
- (Recall that this is also true for the previous version of the "stocks for the long run" argument, when we are asked to choose the time horizon of the investment)
- What am I doing with the money invested for a single year during the other nine years?

Stocks for the long run

- The central point, however, is another
- Sure: the mathematics are exactly the same and are, forgetting the approximation, correct
- However, in the portfolio case we are computing properties of the one year portfolio return
- That is: actual properties of the return we get out of the investment
- In the 10 periods case, instead, we are computing the properties of the AVERAGE per YEAR return
- These are NOT properties of the return I can get out of the second investment

Stocks for the long run

- The financial result I get from the second investment, in return terms, is not the average per year return, but the sum of the 10 log returns that is: the 10 periods return
- In this case, as usual, we have: $E(\sum_{t=1}^{10} r_t^*) = 10\mu$ and $\sqrt{V(\sum_{t=1}^{10} r_i^*)} = \sqrt{10}\sigma$
- These are the properties of the return from the second investment: the average return is just a statistic NOT "what I put in my pocket"
- Hence, NO time diversification exists.
- The apparent diversification is due to the fact that we are computing the properties of an AVERAGE PER YEAR year measure and not of the actual measure of return
- It is a "diversification" of the return measure, not of the performance

Stocks for the long run

- We can now combine the two critiques: the average per year return does not allow us to compare investments over different time horizons AND the average per year return is not the return of an investment
- Missing this point is easy. It also is dangerous
- In fact, it is so common to compare average returns or "annualized returns" over different time intervals that such an error may easily creep in

Stocks for the long run

- Let us try and build comparable investments
- Comparable investments could be a ten year investment of one unit in the diversified portfolio and a ten year investment in the single security (this is not the sole possibility, obviously)
- A possible correct criterion could be the comparison between the full ten year expected return and return variance of the two investments.
- With the same approximations as before, the ten years investment in the ten securities portfolio shall have expected return (over ten years) equal to 10μ and standard deviation σ
- Now we can compare the two investments and we see that the ten year investment in the single security has the same expected value, but a $\sqrt{10}$ times bigger standard deviation

Stocks for the long run

- The two investments, now comparable (same amount same horizon), are by no means equivalent
- The single security investment is seen, in the mean variance sense, as an inferior investment
- Just as we could expect

Stocks for the long run

- In the end:
- The first version of the "stock for the long run" argument could be argued for, at least under some hypothetical, maybe unlikely but coherent, behavioural hypothesis
- The "time diversification" version of the argument is a true fallacy
- A simple riconsideration of the argument shows that no time diversification exists
- The misunderstanding fully comes from the use of a "wrong" measure of return (average per year)

Stocks for the long run

- Third version (Siegel's)
- The third version of the stocks for the long run argument is the soundest, as it can be argued for without being liable of unlikely assumptions or even blatant logical errors.
- This third version, unlikely the other two, is not an a-priori argument, based on assumptions concerning the stochastic behavior of prices and the decision model of agents (and, maybe some logical error).
- Instead, it is an "a posteriori" or "historical" argument.
- As such its acceptance or rejection entirely depends on the way we study and interpret historical data.

Stocks for the long run

- In short, this argument states that, based on the analysis of historical prices, stocks "were always", or at least quite frequently, a good long run investment.
- Being an historical argument, even if true (and here is not the place to argue for or against this point, Shiller and Siegel did this with ability and ingreat detail) this it does not imply that the past behavior should replicate in the future.

Stocks for the long run

- The argument is quite popular among financial journalists
- (Provided they do not weight too much, say, the last 40 years of prices in Japan or the 2009-2015 period for most countries (and, now , the COVID-19 time)
- It is obviously broadly popular in trouble free times
- (At least as popular as the, historically false, argument about real estate as the most sure, if not the best, investment)
- This means that the argument was frequently popular (in the USA, during the first thirty and the last fifty years of the past century and after 2010/12)
- This argument is quite controversial among researchers.

Stocks for the long run

- While not the place for discussing the point, we would suggest the reader, just for the sake of amusement, to consider a basic fault of such "it always used to be..." arguments.
- The following short discussion is not required for the exam but could offer you some interesting business lunch argument.

Stocks for the long run

- We have a typical example of the case where the fact itself of considering the argument, or even the phenomenon itself to which the argument applies, depends of the fact that the phenomenon itself happened, that is: something "was good in the long run".
- In fact, we could doubt about the possibility for an institution (the stock market) which survives in the modern form, at least in the USA, since, say, the second half of nineteenth century, to survive up to today, without at least giving a sustainable impression of offering some opportunities.

Stocks for the long run

- Such arguments, if not accompanied by something else to sustain them, become somewhat empty as could be the analogue to:
- being surprised observing that the dish I most frequently eat, is also among those I like the most
- being surprised by the fact that that old people did non die young
- being surprised by the fact that, when we are in one of many queues, we spend most time in the slowest queue.

Stocks for the long run

- Sometimes, however, the "opportunity" of some institution and how to connect this with its survival can manifest in strange, revealing ways.
- For instance, games of chance exist from unmemorable time with the only "long run" property of making the bank holder richer, together with the occasional random lucky player.
- The overall population of players is made, as a whole, poorer.
- So, while it is clear here what is the "opportunity" of this institution, the survival of such an institution based on such “opportunities” tells us something interesting about human mind.

Stocks for the long run. To conclude (if there is time)

- The main reason why we considered three different versions of the argument is the fact that they are often mixed
- This creates lots of problems in understanding the different points
- An example of this comes directly from Siegel's book

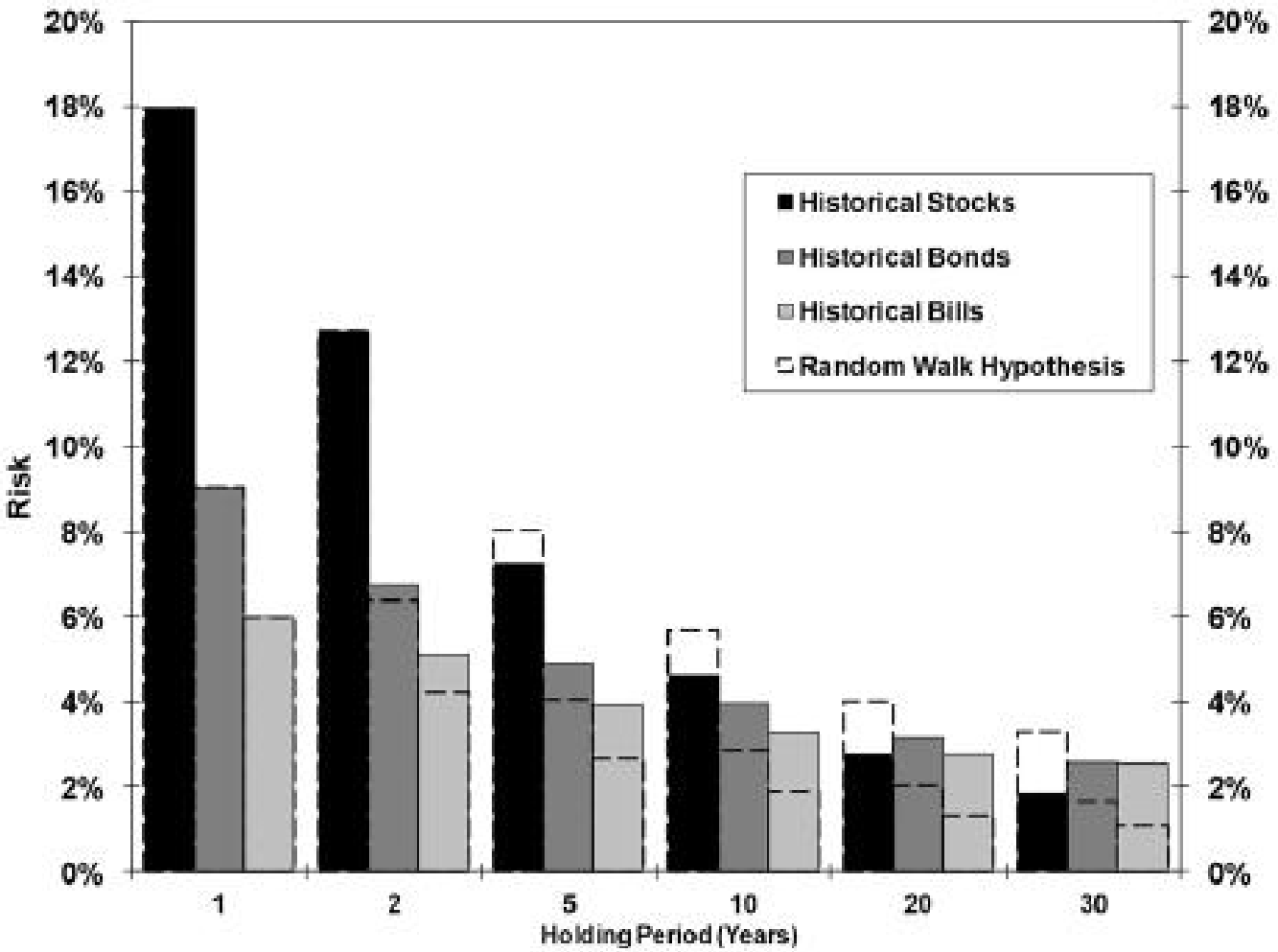
Volatility and Time, Siegel's version, BEWARE!!!

- The following slides from Siegel's book are both interesting and easily misunderstood
- In the first slide Siegel plots the standard deviations of t he AVERAGE returns over different time periods for different asset classes and compare these with what we should observe if the log random walk hypotheses were true
- What we should observe is that these standard deviations should decrease with speed $\sqrt{n}$ (log random walk, constant parameters)
- What we observe is that for stocks the decrease is faster than expected and for Bills and Bonds slower
- The plot can lead you to think that SIegel believes in the first stock for the long run argument.
- This is NOT the point of the plot. The point of the plot is the difference between the stock and bonds and bill behaviour wrt what implied by log random walk

Volatility and Time, Siegel's version, BEWARE!!!

FIGURE 6–2

Standard Deviation of Average Real Stock, Bond, and Bill Returns Over Various Holding Periods: Historical Data and Random Walk Hypothesis 1802–2012



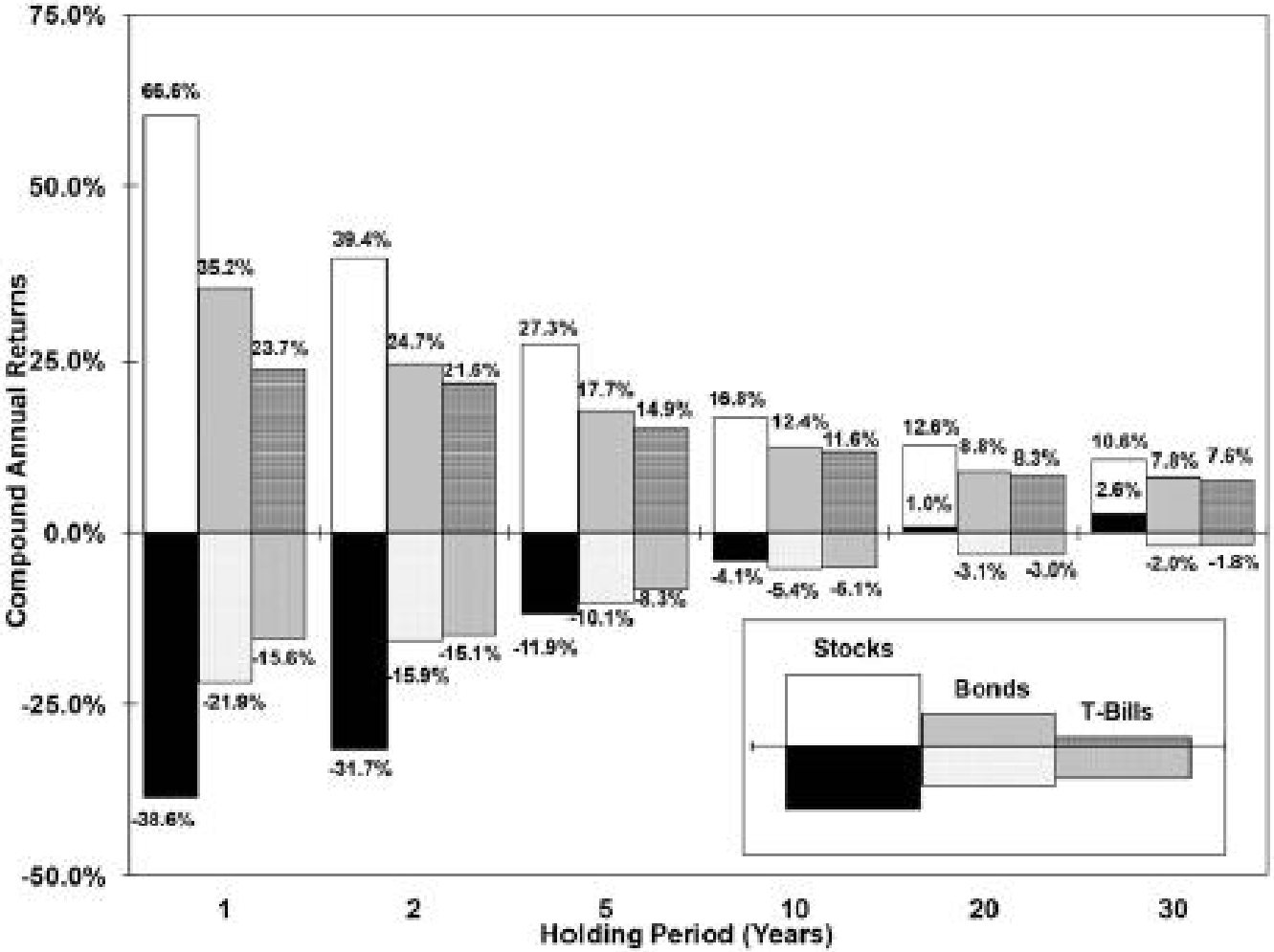
Volatility and Time, Siegel's version, BEWARE!!!

- The second plot shows you that the AVERAGE PER YEAR returns on different time intervals for stocks, bonds and bills have differen properties
- Obviously, the size of maxima and minima decrese to 0 (this is implicit in the use of the average return)
- However what we see is that for stock the longer is the time interval the higher is the unbalance in favour of positive returns
- To understand the plot without errors notice that, in order to get the AVERAGE return for the (e.g.) 30 years interval you need to run the "risk" of (e.g.) single year returns which are very negative or positive
- Again, the plot is correct but could induce you in error, in fact it is much less informative than the first plot

Volatility and Time, Siegel's version, BEWARE!!!

FIGURE 6–1

Highest and Lowest Real Returns on Stocks, Bonds, and Bills over 1-, 2-, 5-, 10-, 20-, and 30-Year Holding Periods 1802–2012



Volatility and Time, Siegel's version, BEWARE!!!

- What Siegel shows here is not that the standard deviation of the AVERAGE returns goes down. This is obvious and, as we did see, quite irrelevant
- What he shows is that, empirically, the standard deviation of the AVERAGE return, in the sample period he considers, did go down FASTER than the $1/\sqrt{n}$ rule
- There are many possible explanations for this (e.g. Siegel does not correct for the correlation among the different results due to the use of a single dataset)
- Some of these shall be considered in future lecture
- The problem in both pictures is the use of the average per year return
- This mixes the obvious fact that the average on longer periods has smaller variance and the interesting fact that, actually, this is smaller than expected

Volatility and Time, Siegel's version, BEWARE!!!

- Siegel is very clear on this point and, if you understand his arguments, you cannot make this mistake
- Alas! similar plots are available in "investor's literature" and comments are not quite so precise
- The following example is not particularly bad but should be useful to understand the problem

Volatility and Time, Siegel's version, BEWARE!!!

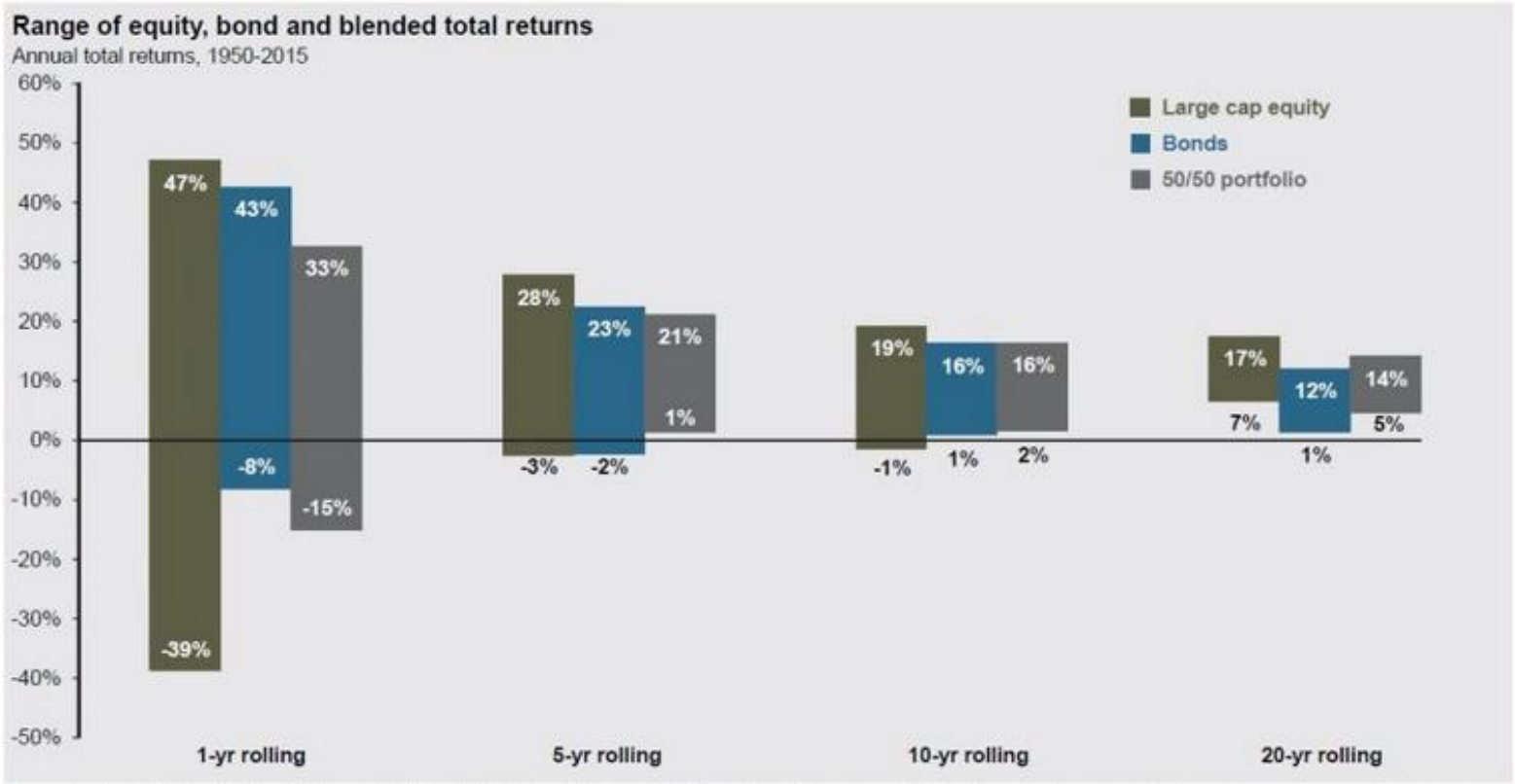
Time

The chart below illustrates just how time is your investment buddy. The bars show us the range of annual returns for an all stock portfolio (green), all bonds (blue), and an 50/50 stocks and bonds portfolio (grey) from 1950 to 2015 over various time frames.

Let's take a look at 1-yr rolling bars. The range of annual returns we saw during this time were 33% (at best), and -15% (at worst) for a 50/50 portfolio. Stretch your investment horizon out to 5 years, and your best and worst are 21% and 1% respectively. Push your horizon out even further, 10 years, you're look at 16% and 2%.

The point that I'm trying to make is, the longer your time horizon, the range of returns are compressed, and the lower the probability of loss. The shorter your time frames, the greater the probability for loss (click for larger image).

Volatility and Time, Siegel's version, BEWARE!!!



Source: Barclays, FactSet, U.S. Federal Reserve, Robert Shiller, Strategas/Ibbotson, J.P. Morgan Asset Management. Returns shown are based on calendar year returns from 1950 to 2016. Large cap equity represents the S&P 500 Shiller Composite and bonds represents the Strategas/Ibbotson for periods from 1950 to 1980 and the Barclays Aggregate after index inception in 1980.

https://www.baharianwealth.com.au/time-diversification-volatility-returns/

Volatility and Time, Siegel's version, BEWARE!!!

- What is wrong here is the statement "the range of returns are compressed"
- The range of AVERAGE returns are compressed but, in order to get a return over several period, you get all return of each single period
- At the end of this lecture you should be able to discuss the pros and cons (and in particular the approximate nature) of similar statements
- You should posses what is needed in order to point out which problems the average investor could have in understanding the true importance and consequences of such suggestions
- You should also be able to rewrite the suggestion in a more correct way

Lecture: Volatility and its Estimate

Summary

- Volatility: from intuition to statistical definition
- Connection with log returns
- The classic estimate and RiskMetrics estimate
- Implied volatility
- More on volatility and time

Volatility

- Show the RiskMetrics manual
- https://www.msci.com/documents/10199/5915b101-4206-4ba0-aee2-3449d5c7e95a
- In applied Finance the term "volatility" has many connected meanings.
- We mention here just the main three:

Volatility

- Volatility may simply mean the attitude of market prices, rates, returns etc. to change in an unpredictable and unjustified manner.
- This without connection to any formal definition of "change", "unpredictable" or "unjustified".
- Here volatility is tantamount chance, luck, destiny, etc.
- Usually the term has a negative undertone and is mainly used in bear markets. In bullish markets the term is not frequently used and it is typically changed in more "positive" synonima.
- A volatile bull market is "exhuberant", "tonic" or "lively".

Volatility

- More formally, and mostly for risk managers, volatility has something to do with the standard deviation of returns and, sometimes, is estimated using historical data (hence the name "Historical Volatility").

Volatility

- For derivative traders and frequently for risk managers "volatility" is the name of one (or more) parameters in derivative models which, under the hypotheses that make "true" the models, are connected with the standard deviation of underlying variables.
- However, in the understanding that these hypotheses are never valid in practice, such parameters are not estimated from historical data on the underlying variables (say, using time series of stock returns) but directly backwarded from quoted prices of derivatives, using the pricing model as fitting formula.
- This is in accord to the strange, but widely held and, in fact, formally justifiable, notion that models may be useful even if the hypotheses underlying them are false.
- This is "Implied Volatility".

Volatility

- Implied volatility is so important for the trading business that implied volatility indexes are traded.
- The VIX index (in particular the VIX index on the S&P500) is the most famous of these.

Volatility

- In what follows we shall introduce a standard and widely applied method for estimating volatility on the basis of historical data on returns, that is, we consider the second meaning of volatility.
- We shall compare it to the "classic" estimate of the variance (which should directly follow from the LRW hypothesis+independence), that is:

$$S_t^2 = \sum_{i=0,\ldots,n} \left(r_{t-i}^* - \overline{r}^*\right)^2 / n$$

- Where $\overline{r}^*$is the sample mean.

Volatility

- This is the standard unbiased estimate for the variance of independent random variables with identical expected values and variances
- (Note: the simple empirical variance of the data, where the denominator its taken as the actual number of observations $n+1$, could be used without problems as in standard applications the sample size is quite big).
- In this statistic, each data point is given the same weight: the idea is that each observation is "as important as" any other
- This comes from the i.i.d. hypothesis
- The log random walk only requires zero correlation of log returns, to get independence we must strengthen the standard hypothesis

Volatility

- In practice, this is not the estimate most used in practice.
- The most common choice is the smoothed estimate or "RiskMetrics" estimate

$$V_t = \frac{\sum_{i=0,\ldots,n} \lambda^i r^{*2}_{t-i}}{\sum_{i=0,\ldots,n} \lambda^i}$$

- From a statistician's point of view this is an exponentially smoothed estimate with λ a smoothing parameter: $0 < \lambda < 1$
- This smoothed estimate is not new. However, due to the fact that it was suggested by J. P. Morgan in the RiskMetrics context, it is commonly known in the field as the RiskMetrics estimate
- Common values of the smoothing parameter used to be around 0.95 (more on this in what follows).

Volatility

- Users of such an estimate do not consider each data point equally relevant.
- Old observations are less relevant than new ones.
- This is not reasonable if we truly believe volatility to be a constant w.r.t. time.
- Implicitly, then, while we "believe" the log random walk (with constant volatility) when "annualizing" volatility, we do not believe it when estimating volatility.

Volatility

- Moreover, it shall be noticed that, in this estimate, the sampling mean of returns does not appear.
- This is a (RiskMetrics) choice which can be justified in two ways:
- First we can assume the expected return μ over a small time interval to be very small.
- With a non negligible variance, it is quite likely that an estimate of the expected value of returns could show a sampling variability greater than its (probable)
- Use an imprecise estimate of the expected value inside the estimate of the variance could create problems to the statistical stability of the variance estimate.

Volatility

- Second, an estimate of the variance where the expected value is set to 0 tends to overestimate, not to underestimate, the variance
- Recall that: $V(X) = E(X^2) - E(X)^2 \leq E(X^2)$.
- For institutional investors, traditionally long the market, this could be seen as a conservative estimate.
- Obviously, this is not a reasonable choice for hedged investors and derivative traders.

Volatility

- The apparent truncation at n should be briefly commented.
- As we have just seen, the standard estimate should be based on the full set of available observations.
- This could be applied as a convention also to the RiskMetrics estimate.
- However, typically, such estimates are computed using one year of daily data or periods of similar lenght
- Now: consider the fact that, e.g., a $\lambda = 0.95$ raised to the power of 256 (conventionally one year of daily data) is less than 0.000002 so, the corresponding weight shall be about this divided by 20 (see below)
- This means that, at least with daily data, to truncate n after one year of data (or even before) is substantially the same as considering the full data set.

- As it is well known ☺, for $0 < \lambda < 1$:

$$\sum_{i=0}^{n} \lambda^i = \frac{1 - \lambda^{n+1}}{1 - \lambda}$$

- For N going to ∞ (and for $0 < \lambda < 1$) this becomes

$$\sum_{i=0,\ldots,\infty} \lambda^i = 1/(1 - \lambda)$$

- (This is the reason for the 20 above: $1/(1 - .95) = 20$)

- If n is not too small (and λ not very, very near to1) we can use this to approximate the V_t estimate as:

$$V_t = \frac{\sum_{i=0,\ldots,n} \lambda^i r^{*2}_{t-i}}{\sum_{i=0,\ldots,n} \lambda^i} \cong (1-\lambda) \sum_{i=0,\ldots,n} \lambda^i r^{*2}_{t-i}$$

- Which is simpler to write down and to work with

- In order to better understand the meaning of this estimate, it is useful to write it in a recursive form (this is also useful for computational purposes).
- We can directly check (proof not required) that:

$$V_t = \lambda V_{t-1} + \frac{r_t^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} - \frac{\lambda^{n+1} r_{t-n-1}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i}$$

- (This is the non required proof) In fact, since

$$V_{t-1} = \frac{\sum_{i=0,\ldots,n} \lambda^i r_{t-1-i}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i}$$

- We have that $\lambda V_{t-1} + \frac{r_t^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} - \frac{\lambda^{n+1} r_{t-n-1}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i}$ is equal to:

$$\lambda \frac{\sum_{i=0,\ldots,n} \lambda^i r_{t-1-i}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} + \frac{r_t^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} - \frac{\lambda^{n+1} r_{t-n-1}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} =$$

$$= \frac{\sum_{i=0,\ldots,n} \lambda^{i+1} r_{t-1-i}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} + \frac{r_t^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} - \frac{\lambda^{n+1} r_{t-n-1}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} =$$

$$= \frac{r_t^{*2} + \sum_{i=0,\ldots,n-1} \lambda^{i+1} r_{t-1-i}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i} = \frac{\sum_{i=0,\ldots,n} \lambda^i r_{t-i}^{*2}}{\sum_{i=0,\ldots,n} \lambda^i}$$

- Which is the definition of V_t. (End of non required proof)

- Let us discuss this result.
- For the standard range of values of λ and n the last term in V_t, that is: $\lambda^{n+1} r^{*2}_{t-n-1} / \sum_{i=0,\ldots,n} \lambda^i$ can be approximated with 0 (basically, because, with big n, λ^{n+1} shall be negligible).
- With this approximation, and using the approximate value of the denominator: $(1/(\sum_{i=0,\ldots,n} \lambda^i) \cong 1-\lambda)$ we have:

$$V_t \cong \lambda V_{t-1} + (1-\lambda) r^{*2}_t$$

- For each t, the new value of the estimate: V_t is a weighted mean of the old estimate V_{t-1} (weight λ, usually big, very near to 1) and of the latest squared log return (weight $1-\lambda$, usually small).

Volatility

- A simple consequence of this (and of the fact that the estimate does not consider the mean return) is the following.
- Since the squared return is always non negative and λ is usually near one, this formula implies that V_t , if the new return is 0, is equal to λV_{t-1} so that the estimated variance at most can decrease of a percentage of $1-\lambda$.
- You cannot observe "big downward jumps" of the estimate.
- On the other hand, the estimate can increase, in principle, of any amount when abnormally big squared returns are observed.
- This implies an asymmetric behavior: following any shock, we have an abrupt jump UP in V_t, while a sequence of "normal" values for returns with small shocks shall reduce the estimated value in a smoothed (exponential) way, the faster the smaller is λ.
- The reader should remember that this behavior of volatility estimated with this method, is purely a feature of the formula used for the estimate.

- Compare this with the behaviour of the traditional (equally weighted) estimate (again, with $\mu = 0$)

$$S_t^2 = \frac{\sum_{i=0,\ldots,n} {r_{t-i}^*}^2}{n+1} = S_{t-1}^2 + \frac{1}{n+1} r_t^{*2} - \frac{1}{n+1} r_{t-n-1}^{*2}$$

- Note that, in this case that the weight of the dropped "oldest" observation i identical to that of the "new" observation
- For this reason, if the oldest observation represents a "big shock", its exit from the estimate, not being progressively discounted, shall imply a strong drop in the estimate
- Smoothing avoids this by discounting

- The use the smoothed estimate of σ^2 implies, as mentioned, a disagreement with the standard version LRW hypothesis. In fact, it implies a time evolution of the variance of returns. The recursive formula:

$$V_t = \lambda V_{t-1} + (1-\lambda) r_t^{*2}$$

- is the empirical analogue of an auto regressive model for the variance of returns we could write as:

$$\sigma_t^2 = \gamma \sigma_{t-1}^2 + \varepsilon_t^2$$

- This is a particular case of a class of dynamic models for conditional volatility (ARCH Auto Regressive Conditional Herteroschedastic) of considerable fortune in the econometric literature.

Volatility

- The choice of λ is so important that it was the object of recent regulations.
- Here I reproduce a paragraph of the EBA (European Banking Authority) paper EBA/CP/2015/27.

Article 38 Observation period

1. Where competent authorities verify that the VaR numbers are computed using an effective historical observation period of at least one year, in accordance with point (d) of Article 365(1) of Regulation (EU) No 575/2013, competent authorities shall verify that a minimum of 250 business days is used. Where institutions use a weighting scheme in calculating their VaR, competent authorities shall verify that the weighted average time lag of the individual observations is not less than 125 business days.

2. Where, according to point (d) of Article 365(1) of Regulation (EU) No 575/2013 the calculation of the VaR is subject to an effective historical observation period of less than one year, competent authorities shall verify that the institution has in place procedures to ensure that the application of a shorter period results in daily VaR numbers greater than daily VaR numbers computed using an effective historical observation period of at least one year.

- The quoted *Article 365(1) of Regulation (EU) No 575/2013 (On prudential requirements for credit institutions and investment firms)*, is as follows:

> *Article 365*
> *VaR and stressed VaR Calculation*
> *1. The calculation of the value-at-risk number referred to in Article 364 shall be subject to the following requirements:*
> *(a) daily calculation of the value-at-risk number;*
> *(b) a 99th percentile, one-tailed confidence interval;*

(c) a 10-day holding period;

(d) an effective historical observation period of at least one year except where a shorter observation period is justified by a significant upsurge in price volatility;

(e) at least monthly data set updates.

The institution may use value-at-risk numbers calculated according to shorter holding periods than 10 days scaled up to 10 days by an appropriate methodology that is reviewed periodically.

2. In addition, the institution shall at least weekly calculate a "stressed value-at-risk" of the current portfolio, in accordance with the requirements set out in the first paragraph, with value-at-risk model inputs calibrated to historical data from a continuous 12-month period of significant financial stress relevant to the institution's portfolio. The choice of such historical data shall be subject to at least annual review by the institution, which shall notify the outcome to the competent authorities. EBA shall monitor the range of practices for calculating stressed value at risk and shall, in accordance with Article 16 of Regulation (EU) No 1093/2010, issue guidelines on such practices.

- The meaning of this rule (which could be quite involuted for the non technical person, but not for you) is as follows:
- If you use the exponentially smoothed estimate truncated at n so that your (daily data) exact weights are (Note: $0 < \lambda < 1$)

$$w_{t-i} = \lambda^i / \sum_{j=0}^{n} \lambda^j = \lambda^i \frac{1-\lambda}{1-\lambda^{n+1}}$$

- The rule requires for the *"weighted average time lag of the individual observations"* , that is: $\sum_{i=0}^{n} i w_{t-i}$ to be at least 125 (days).
- This, for given n puts strong constraints on the value of λ.
- If $n = 250$ (about one year of data) the only possible (non increasing) choice is $\lambda = 1$, so that each observation has the same weight $w_{t-i} = 1/(n+1)$
- If you want $\lambda < 1$ you must increase n.

Volatility

- However, even for $n \to \infty$, since the weighted average time lag is $\sum_{i=0}^{\infty} i w_{t-i} = \sum_{i=0}^{\infty} i\lambda^i(1-\lambda) = \lambda/(1-\lambda)$ the requirement asks *in any case* (that is: whatever be n) for a value $\lambda/(1-\lambda) > 125$ that is: $\lambda > 125/126 = .992063$.
- The smaller the n the bigger the required λ, but in any case it must be bigger that 125/126.
- The rule, in practice, asks for λ to be 1 (or almost 1) and for n to be bigger that 250
- This λ is much bigger than what used to be the common case in the past.
- The examples in the "classic" edition of the Risk Metrics Technical Document (iv edition 1996) use $\lambda = .94$ which, even with very big n, corresponds to a weighted average time lag of $.94/.06 = 15.(6)$ by far too small according to the new rules.

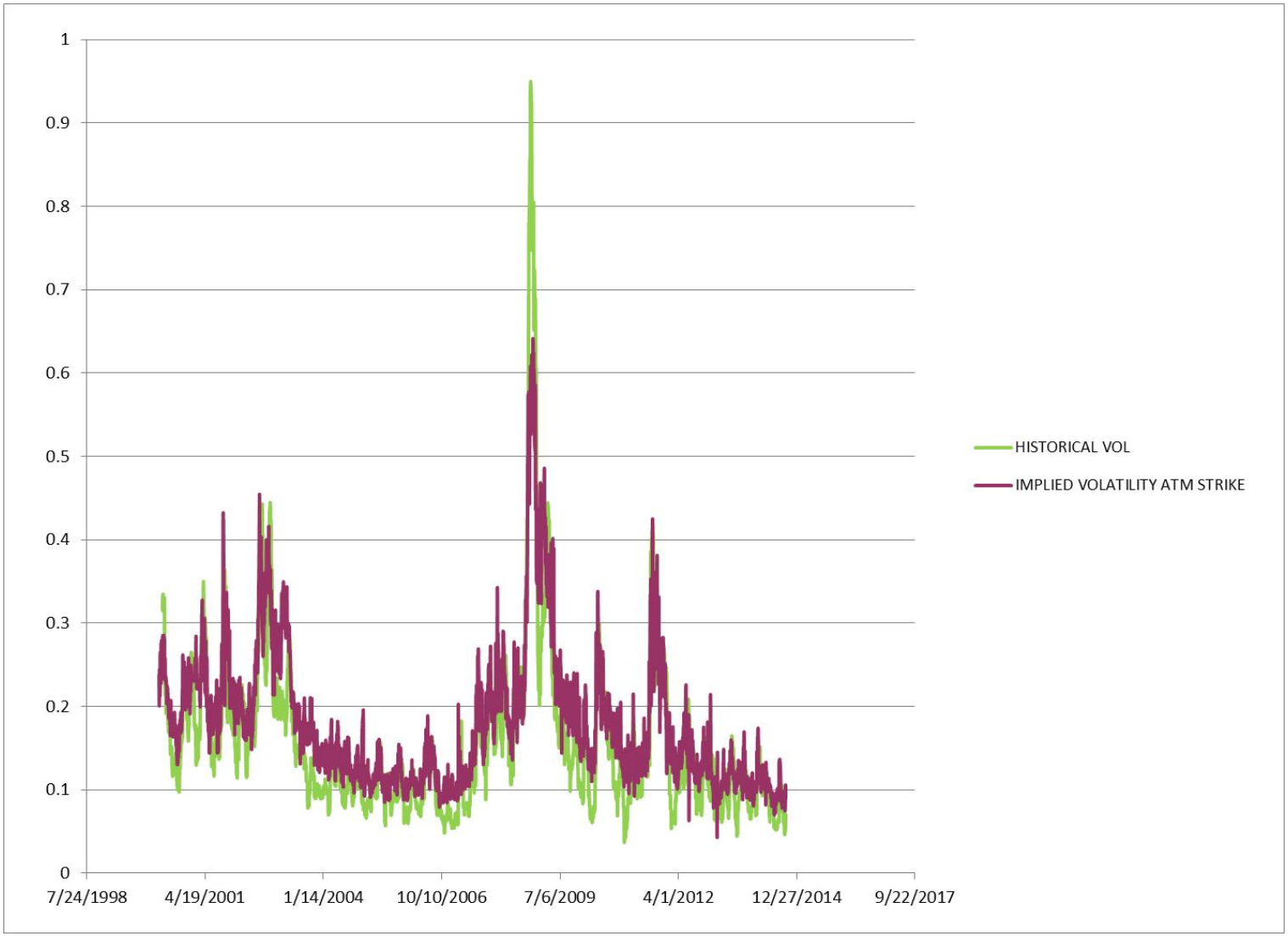
Volatility

- Those who wrote the rule seem to be very concentrated on avoiding risk managers "too fast discounting" past extreme observations
- This may be meaningful, however a, maybe not well considered, problem, creeps in
- As noted above, If each datapoint has, basically, the same weight so that there exist no discounting, the "effect" on the estimate of any extreme observation at time t shall be present, with the same force, in ALL future estimates up to time $t+n$
- Arrived at $t+n+1$ the observation drops out of the sample and the effect disappears with a suddent drop of the estimate
- This is exactly what was avoided by smoothing
- Do Regulators really want this?

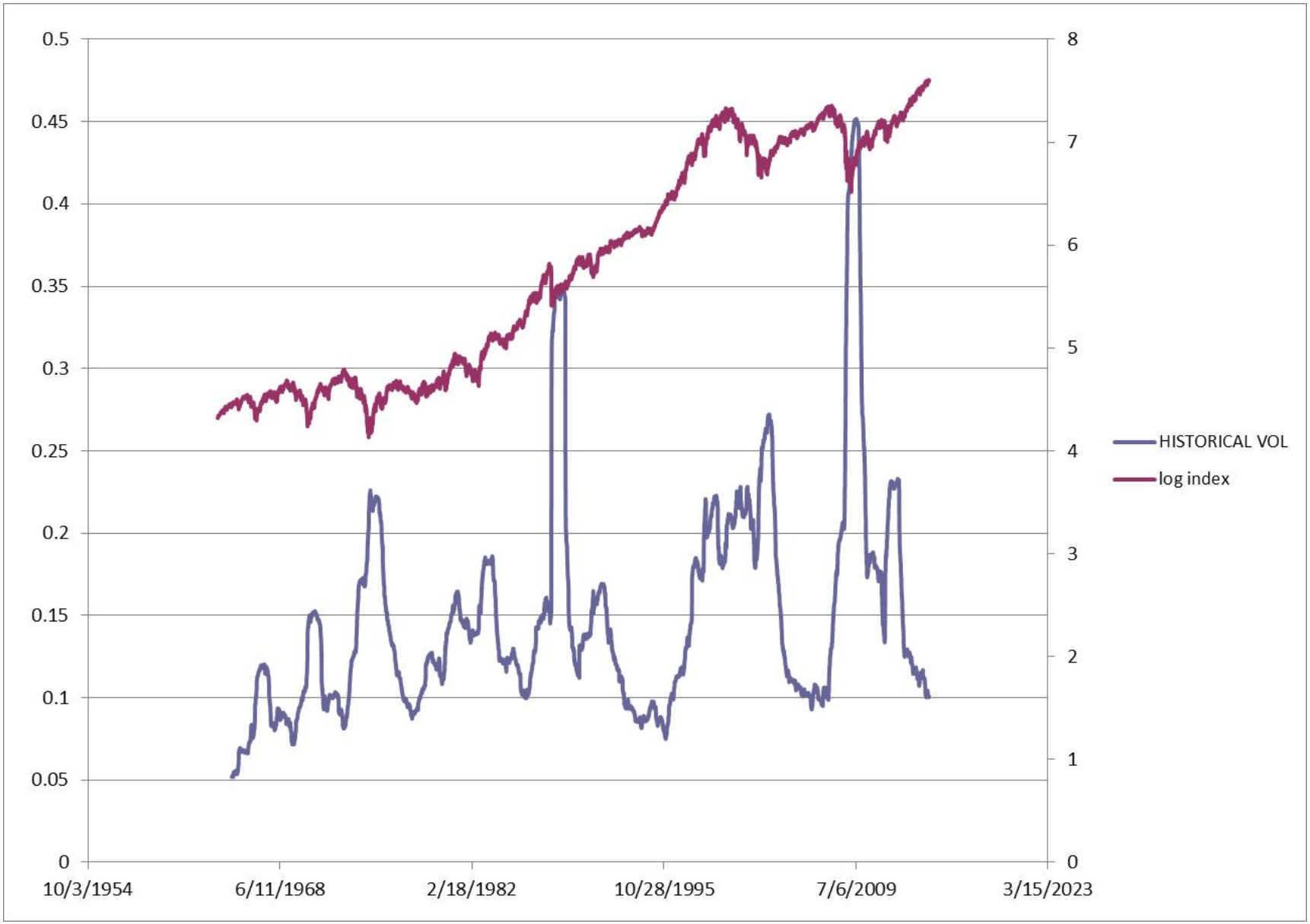
Volatility

- In the following plots we first compare "historical" and "implied" (atm forward) volatility (notice that, in most cases, implied volatility>historical volatility)
- We then see a comparison of historical volatility (classic estimate) and the evolution of the stock market
- The following two plots show you two smoothed estimates bused on the same data but different λ
- The last two plots show you the evolution of the Vix index (the second one, from Siegel book, poits out some relevant "event" connected with volatility)
- Notice the asymmetry in the evolution of VIX: these are similar to those in the evolution of the smoothed estimate (with not too big a λ)

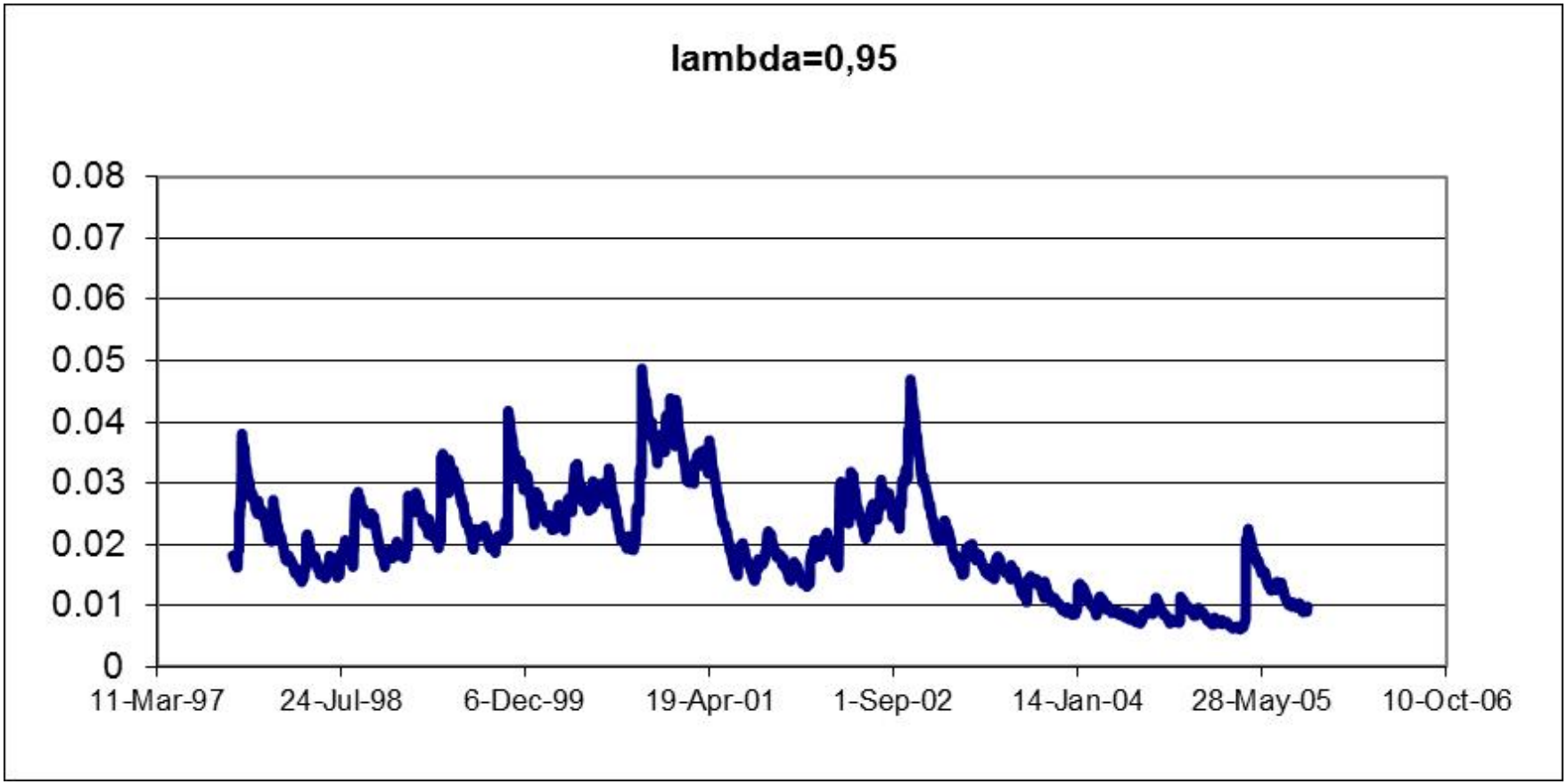
Historical VS Implied Volatility



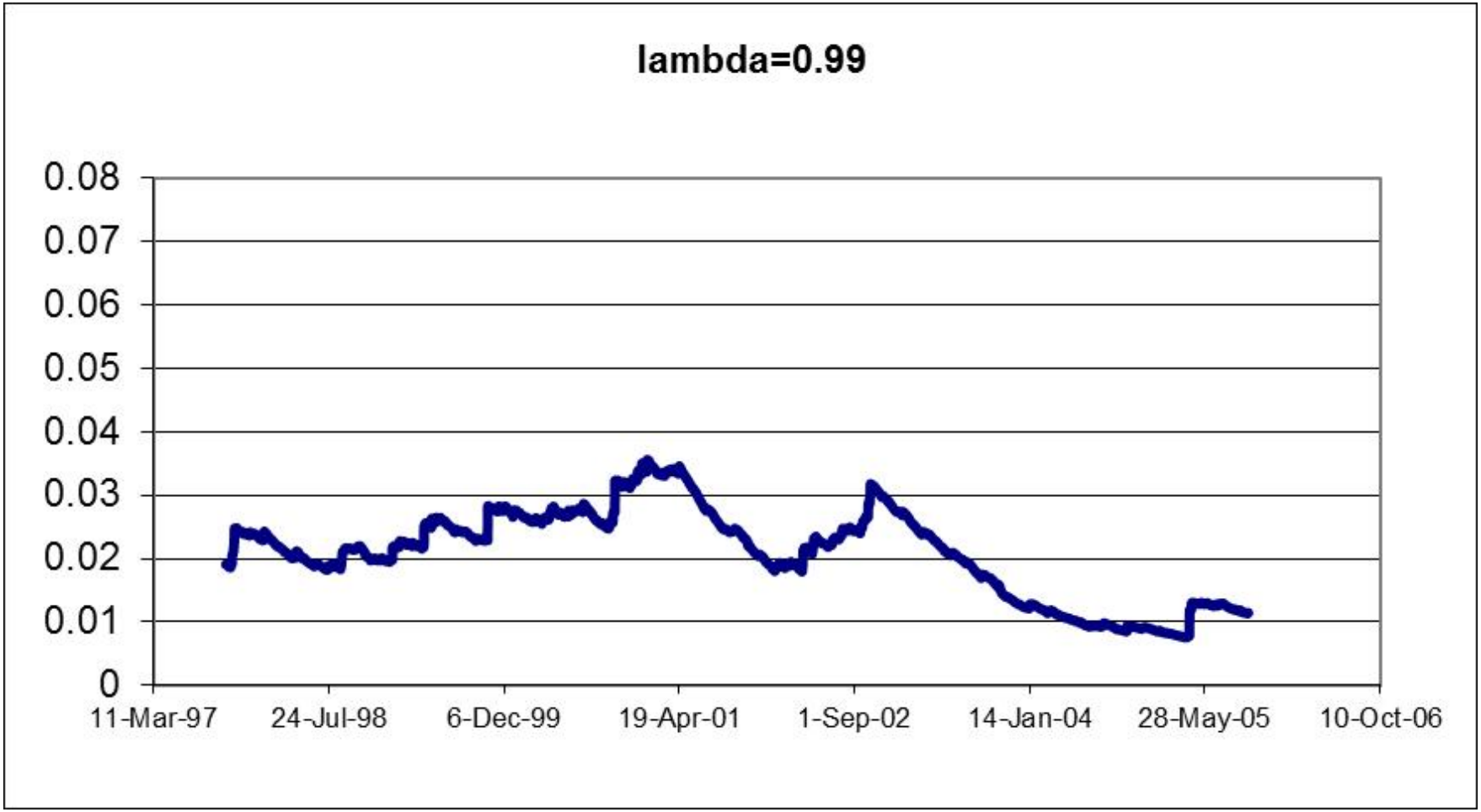
Index and Hist. Vola



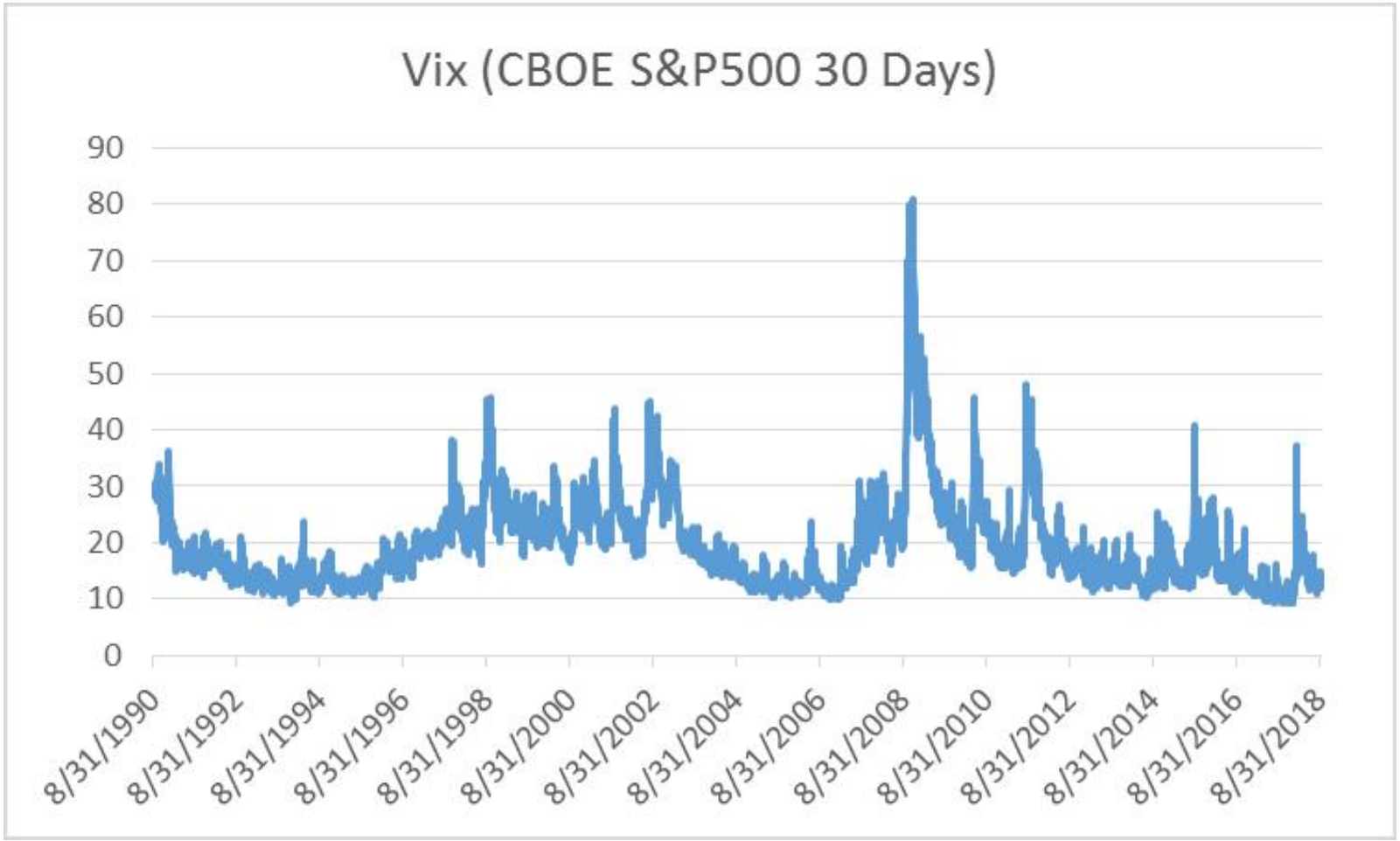
Smoothed Vola .95 n=256



Smoothed Vola .99 n=256



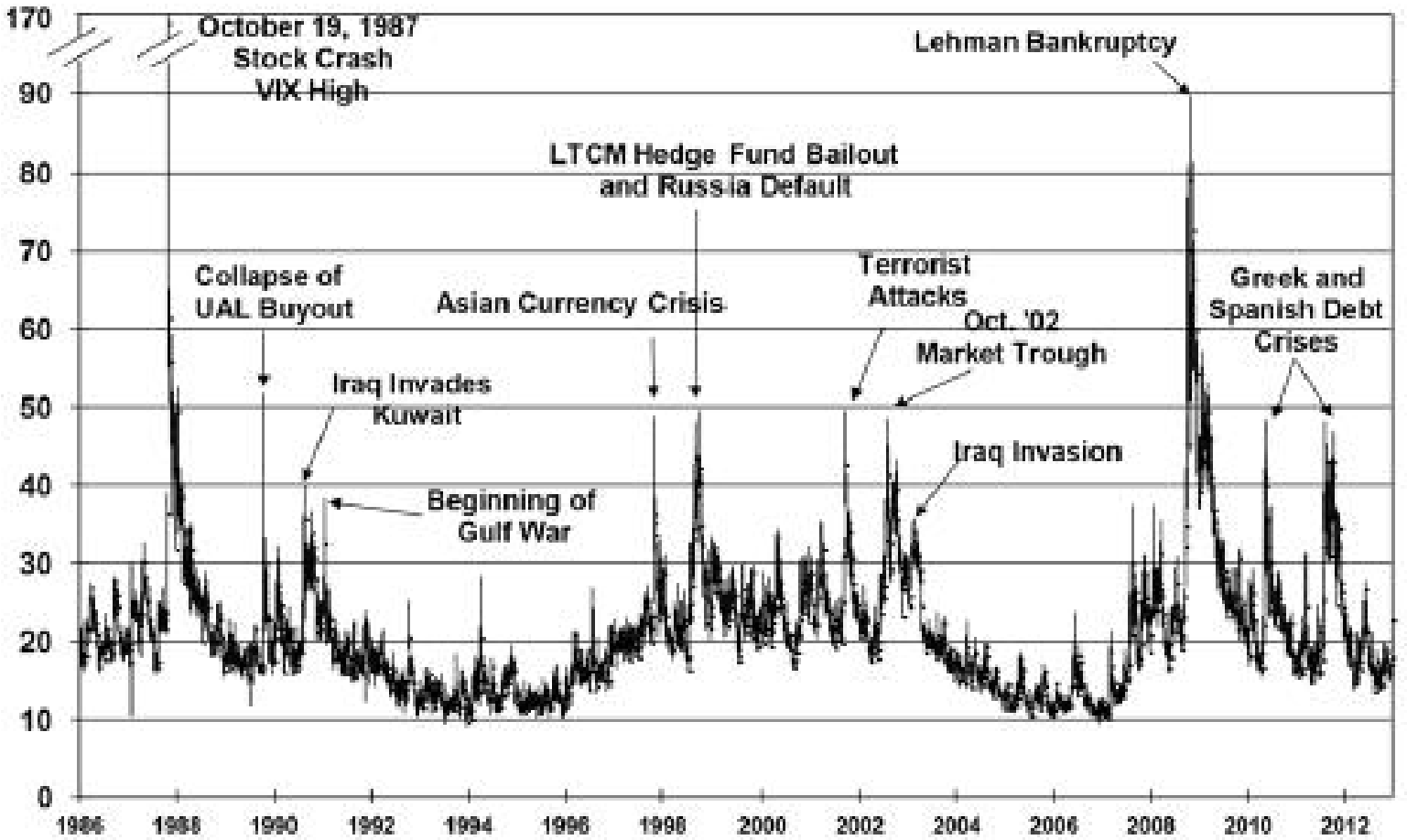
Vix


Vix (CBOE S&P500 30 Days)
90
80
70
60
50
40
30
20
10
0
8/31/1990
8/31/1992
8/31/1994
8/31/1996
8/31/1998
8/31/2000
8/31/2002
8/31/2004
8/31/2006
8/31/2008
8/31/2010
8/31/2012
8/31/2014
8/31/2016
8/31/2018

Vix from Siegel 2014

FIGURE 19–5

The Volatility Index (VIX) 1986–2012



Volatility and Time

- In the next slide, we compute three annualized estimate of the standard deviation of a typical stock (IBM) based on 5 years of data
- The three estimates are based on daily data, monthly data and yearly data
- Data are not overlapping so that we have only 5 observations for yearly data
- In a log random walk with constant variance we should observe very similar values of the estimates
- In practice, if we abstract (and should NOT), from samplig variability we se values which are smaller the longer the observation period
- Compare this with the first of Siegel's plots below

Exercise 3c: VOLATILITY

The following table presents a comparison between standard deviation computed using 5 years data expressed in 3 different basis: yearly, monthly and daily. These data are presented in the other sheets, where there are 100 simulation per each kind of series.

	Year	*Month*	*Day*
mean estimate	27.63	29.61	30.07
st.err.	10.02	2.95	0.60

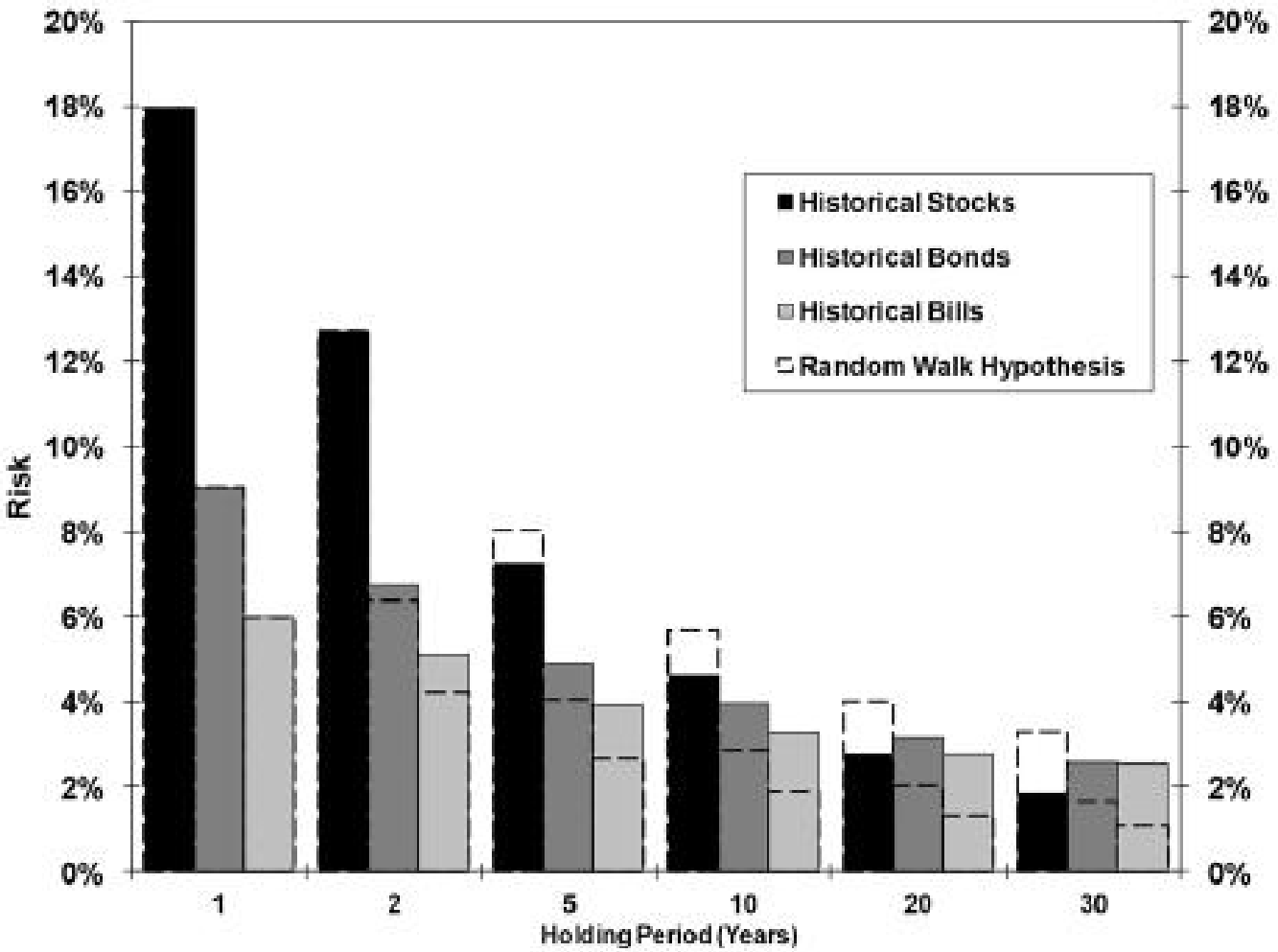
Volatility and Time, Siegel's version, BEWARE!!!

- Now back to Siegel's plots.
- Siegel compares the estimated volatilities of the average returns (BEWARE!) over different time intervals and for different asset classes
- He compares the empirical estimates with the values you should get if the log random walk hypothesis were true
- (Alas: Siegel uses overlapping data and this biases his results, and also the log random walk implication)
- Moreover, the use of average returns, contrary to the use of annualized volatility, could induce the reader into dangerous "time diversification" fancies
- In any case, from the qualitative point of view, the message is correct: there appears to be a small negative correlation of log returns whose effect you can only see if the time horizon is long enough

Volatility and Time, Siegel's version, BEWARE!!!

FIGURE 6–2

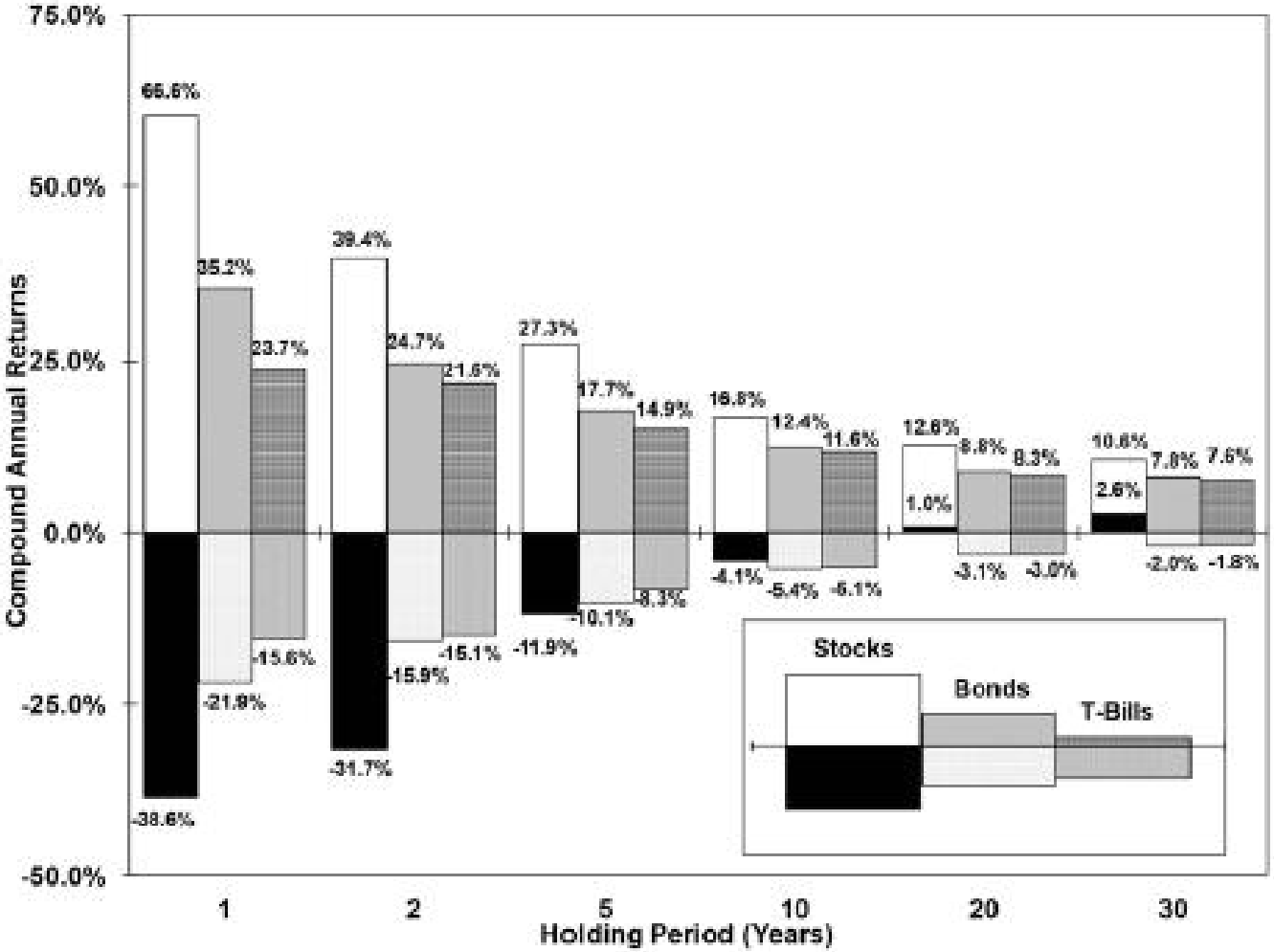
Standard Deviation of Average Real Stock, Bond, and Bill Returns Over Various Holding Periods: Historical Data and Random Walk Hypothesis 1802–2012



Volatility and Time, Siegel's version II, BEWARE!!!

FIGURE 6–1

Highest and Lowest Real Returns on Stocks, Bonds, and Bills over 1-, 2-, 5-, 10-, 20-, and 30-Year Holding Periods 1802–2012



Lecture: Is it easier to estimate μ or σ^2?

Summary

- Statistical properties of the estimates of expected return and volatility
- The setting: log random walk
- The sampling standard deviations of the estimates of expected return and return variance
- Surprisingly different behaviours of the two sampling standard deviations and the reason for these
- The role of higher frequency data in the two estimates

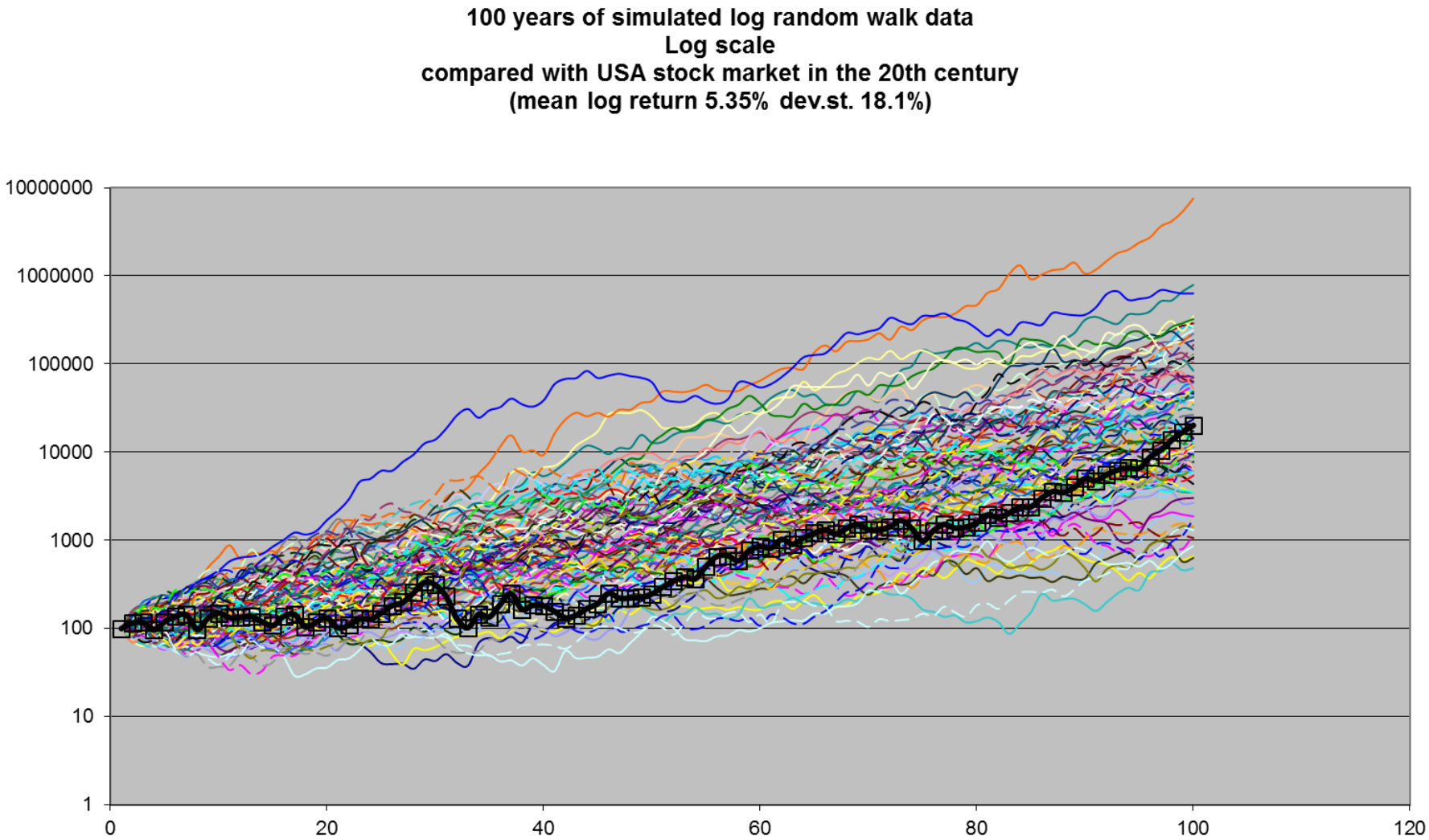
Is it easier to estimate μ or σ^2?

- From: Robert Merton (1980, J.F.E., "On estimating the expected return on the market: An exploratory investigation")
- ..."it is not for want of applications that expected return estimation has not been pursued. A more likely explanation is simply that estimating expected returns from time series of realized stock return data is very difficult. As it is shown in appendix A, the estimates of variances or covariances from the available time series will be much more accurate than the corresponding expected return estimates. Indeed, even if the expected return on the market were known to be a constant for all time, it would take a very long history of returns to obtain an accurate estimate. And, of course, if this expected return is believed to be changing through time, then estimating these changes is still more difficult"...

Is it easier to estimate μ or σ^2?

- Let us recall the graphs of the simulated log random walks (in logarithmic scale)

Is it easier to estimate μ or σ^2?



Is it easier to estimate μ or σ^2?

- Now, let us plot the corresponding log returns

Is it easier to estimate μ or σ^2?



Is it easier to estimate μ or σ^2?

- How is it possible that the series in log levels be so "different" while in log returns they seem, basically, to fill the same space (roughly between -30 and +40)?
- The answer is simple but it has deep implications
- Log level trajectories show very different terminal values
- When we translate these to returns, the differences in means are there but

1. We almost do not notice them, because they are much smaller that the standard deviations
2. We notice, instead, that the variability of log returns is very similar across different series

Is it easier to estimate μ or σ^2?

- Compute, for each trajectory, the average log return (Mean) and the standard deviation of the log returns (Std)
- Then, compute the mean and average of these (approximations of the expected value of the estimates and of their sample standard deviation)
- This is what you get

	sample mean	sample std	t ratio	lb -2sigma	ub+2sigma
Mean	5.1%	1.7%	3.02	1.7%	8.5%
Std	18.1%	1.2%	14.45	15.6%	20.5%

Is it easier to estimate μ or σ^2?

- Both the sample mean of the Means and Stds (over the 100 trajectories) are very similar to the true values (we know these true values because this is simulated data)
- However, the standard deviation of the Mean (across different trajectories) is, in percentage of the value of the Mean, much bigger than the standard deviation of the Std
- (Compare the t ratios and the confidence intervals)
- This means that the standard deviations measured along the different trajectories are much more similar than the averages
- For this reason the log price trajectories (where the Means dominate summing thru time) seem so different while the log return trajectories seem to be "equally messy"
- Is this telling us that it is easier to learn about "risk" than "return"?

Is it easier to estimate μ or σ^2?

- The answer is "YES". We can see it empirically and, if we assume log random walk, we can prove that this must be the case
- The implications of this, joint with the already discussed fact that we expect standard deviation quite bigger than expected values, are, in fact, amazing
- The behaviour of market agents is deeply influenced by this simple but, maybe, a priori surprising property of stock price series
- And, beware, many market agents do not realize, at least at a conscious level, that this is happening
- But there is more

Is it easier to estimate μ or σ^2?

- Consider log returns from a typical stock, let them be iid and, just to fix ideas, let the expected value (on an yearly basis) be .07 and the standard deviation .3.
- If, using yearly data we estimate $\hat{\mu} = \sum_{i=1,\ldots,n} r_i/n$ (forget, for the moment, the $t-i$ notation) we have, as is well known, a sampling variance equal to:

$$V(\hat{\mu}) = \sum_{i=1,\ldots,n} V(r_i/n) = \frac{n}{n^2}\sigma^2 = \sigma^2/n$$

- Under the above hp, the sampling standard deviation is $.3/\sqrt{n}$ where n is the number of years used in the estimation
- The ratio of the expected value and the standard error of its estimate, under these hypotheses, shall be $\sqrt{n}7/30$, roughly: $\sqrt{n}/4$.

Is it easier to estimate μ or σ^2?

- This is quite small, if n is not big
- For a t-ratio of 2 we need $\sqrt{n} = 8$ that is $n = 64$ (years!).
- For a t-ratio of 4 we need $\sqrt{n} = 16$ that is $n = 256$ (years!).

Is it easier to estimate μ or σ^2?

- We are using yearly data, what happens if we use higher frequency data?
- Let k be the number of observations per year, $\mu_k = \mu/k$ the expected value per subperiod and $\sigma_k^2 = \sigma^2/k$ the variance per period
- The sampling variance for the estimate of $\mu_k$shall then be

$$V(\hat{\mu}_k) = V(\sum_{i=1}^{kn} r_{ki}^*/kn) = \sigma_k^2/kn = \sigma^2/k^2 n$$

- This seems much better that before, but it is an illusion: we do not need an estimate of μ_k, we need an estimate of $\mu = k\mu_k$.

Is it easier to estimate μ or σ^2?

- This estimate is $k\hat{\mu}_k$ and we must then compute $V(k\hat{\mu}_k)$. We get
$$V(k\hat{\mu}_k) = k^2 V(\hat{\mu}_k) = \sigma^2/n$$
- Exactly the same as with "aggregated" data.
- This should not surprise us: in fact the arithmetic mean of log returns is simply given by the log of the ratio of the last to the first price divided by the required number of data points.
- In other words it only changes because of the denominator: n for a yearly mean and kn for a sub period of length k mean.
- No information useful for estimating μ is added by using sub period data, hence no improvement in the variance of the estimate.

Is it easier to estimate μ or σ^2?

- In summary: the expected return is difficult to estimate with good precision for two reasons.
- First: σ is expected to be much bigger than μ and the t-ratio depends on the ratio of these.
- Second: even if we increase the frequency of observations nothing changes for the estimate of the (yearly) μ so that its sampling variance stays the same.

Is it easier to estimate μ or σ^2?

- Now let us do a similar analysis for the variance.
- In order to make things simple we shall suppose that μ is known and data are Gaussian.
- This allows us to quickly derive some useful result.

Is it easier to estimate μ or σ^2?

- Let us compute the sampling variance of our estimate of the variance (classic formula, known μ).
- Let r_i^* be the yearly log return

$$V(\hat{\sigma}^2) = V(\sum_{i=1}^{n} \frac{(r_i^* - \mu)^2}{n}) = \frac{1}{n} V((r^* - \mu)^2) =$$

$$= \frac{1}{n}(E((r^* - \mu)^4) - E((r^* - \mu)^2)^2) = \frac{1}{n}(\mu_{c4} - \sigma^4)$$

- Where $\mu_{c4} = E((r^* - \mu)^4)$ is the fourth centered moment and without further hypothesis could be any non negative constant.

Is it easier to estimate μ or σ^2?

- If the r_i^* are Gaussian we have $\mu_{c4} = 3\sigma^4$ (take it as given) and the resulting variance of the sampling variance is

$$V(\sum_i \frac{(r_i^* - \mu)^2}{n}) = \frac{2}{n}\sigma^4$$

- So that the sampling *standard deviation* of the estimated variance shall be: $\sqrt{2}\sigma^2/\sqrt{n}$.
- The ratio of the variance with the sampling standard deviation of its estimate shall be, in this case

$$\frac{\sigma^2}{\sigma^2}\frac{\sqrt{n}}{\sqrt{2}} \approx \frac{\sqrt{n}}{\sqrt{2}} \approx .7\sqrt{n}$$

- In order to get a ratio equal to 2 we need $\sqrt{n} > 2/.7$ and we get there with just $n = 9$ instead of 64 as in the expected value case.

Is it easier to estimate μ or σ^2?

- For a ratio of 4 or greater we need now $\sqrt{n} > 4/.7$ and for this $n = 33$ suffices (instead of $n = 256$ for the above discussed case of the expected value).
- But there is much more: now using higher frequency data improves the estimate.
- Let our strategy be that of estimating yearly variance as k times the estimated variance for a sub period of length$1/k$ in yearly terms (the prime sign is to indicate that this is a new estimate)

$$\hat{\sigma}'^2 = k\hat{\sigma}_k^2$$

Is it easier to estimate μ or σ^2?

- Using the same notation and hypotheses as above we get

$$V(\hat{\sigma}_k^2) = V(\sum_{i=1}^{kn} \frac{(r_{ki}^* - \mu_k)^2}{kn}) = \frac{1}{kn} V((r_k^* - \mu_k)^2) =$$

$$= \frac{1}{kn}(E((r_k^* - \mu_k)^4) - E((r_k^* - \mu_k)^2)^2) =$$

$$= \frac{1}{kn}(\mu_{kc4} - \sigma_k^4) = \frac{2}{kn}\sigma_k^4 = \frac{2}{kn}\frac{\sigma}{k^2}^4$$

- Where we used the Gaussian hypothesis (last step).

Is it easier to estimate μ or σ^2?

- Then
$$V(\hat{\sigma'}^2) = k^2 V(\hat{\sigma}_k^2) = \frac{2}{kn}\sigma^4$$
- And we see that now k IS in the formula: using sub period data improves the estimate.
- Now the ratio between the true variance and the standard deviation of its new, sub period based, estimate is approximately equal to
$$\frac{\sigma^2}{\sigma^2}\frac{\sqrt{kn}}{\sqrt{2}} \approx .7\sqrt{kn}$$
- So that the use of k sub periods per year has an effect identical to that of multiplying the number of years by k.

Is it easier to estimate μ or σ^2?

- With, say, monthly data, we need less than one year (actually 9 months) for a ratio of 2 and slightly less that 3 years (just 33 months) so that ratio for the variance becomes greater that 4. (Instead of 33 years with yearly data).
- Estimating σ^2 is then easier than estimating μ for two reasons:
- First: the ratio between the estimate and its standard deviation, the relative precision of the estimate, is bigger than that for the expected value, whatever be the n.
- This comes from the empirical, and theoretical idea that expected returns of stocks are much smaller than volatilities.
- Second: even if the first reason was not true, we still have the fact that using higher frequency data improves (dramatically) the quality of the estimate for σ^2 while it is irrelevant for the estimate of μ.
- As mentioned above, all our formula hold for known μ, the standard estimate and Gaussian log returns.

Is it easier to estimate μ or σ^2?

- For the general case (you estimate μ) we have the following result:
- With i.i.d. log returns not necessarily Gaussian, for the estimate

$$S^2 = \sum_{i=1}^{n} (r_i^* - \bar{r}^*)^2/(n-1))$$

- We get (no proof required)

$$V(S^2) = \frac{\mu_{c4}}{n} - \frac{\sigma^4 (n-3)}{n(n-1)}$$

- This, for not too small n and a fourth centered moment not very different from the Gaussian case, gives us the same result as the above formula.

Is it easier to estimate μ or σ^2?

- Notice that in all these cases the sampling variance of the estimate of the variance (as that of the estimate of the expected value) goes to 0 with n going to infinity.
- Let us run a similar analysis for the smoothed estimate.
- We are going to use the approximation for the denominator given by:

$$\sum_{i=0,\ldots,n} \lambda^i \approx \sum_{i=0,\ldots,\infty} \lambda^i = 1/(1-\lambda)$$

Is it easier to estimate μ or σ^2?

- The variance of the smoothed estimate (μ set to 0) is (back to the $t-i$ notation)

$$V(V_t) = V((1-\lambda)\sum_{i=0,\ldots,n} \lambda^i r_{t-i}^{*2}) = (1-\lambda)^2 \sum_{i=0,\ldots,n} \lambda^{2i} V(r_{t-i}^{*2}) =$$

$$= \frac{(1-\lambda)^2}{1-\lambda^2}(\mu_4 - \mu_2^2) = \frac{1-\lambda}{(1+\lambda)} 2\sigma^4$$

- Here the last equality is true if the expected value is zero (as assumed in RiskMetrics) and log returns are Gaussian (and recall:$1-\lambda^2 = (1+\lambda)(1-\lambda)$).
- Notice that this does not go to 0 if we increase n.
- This is, actually, not so important: even in the case of the classic estimate we would truncate the sample (if you believe a parameter is changing you do not look for an estimate converging to a constant).

Is it easier to estimate μ or σ^2?

- Here it is meaningless to compare this with the quality of the estimate for μ because this is assumed equal to zero.
- It is, however, interesting to compare the result with a result based on sub period of length k.
- Everything depends on the choice of λ for the sub periods. The simplest choice is to set it as $\lambda^{1/k}$ In this case we have

$$V_{kt} = (1 - \lambda^{1/k}) \sum_{i=0,\ldots,kn} \lambda^{i/k} r^{*2}_{kt-i}$$

Is it easier to estimate μ or σ^2?

- So that, following the same steps as for $V(V_t)$

$$V(V_{kt}) = \frac{1-\lambda^{1/k}}{(1+\lambda^{1/k})} 2\frac{\sigma^4}{k^2}$$

- and we have, for the estimate $V'_t = kV_{kt}$ (we use the prime sign because this is different with respect to the estimate using aggregated data)

$$V(kV_{kt}) = k^2 \frac{1-\lambda^{1/k}}{1+\lambda^{1/k}} 2\frac{\sigma^4}{k^2} = \frac{1-\lambda^{1/k}}{1+\lambda^{1/k}} 2\sigma^4$$

- This, for $0 < \lambda < 1$ and $k > 1$, is always smaller than the variance computed using only full period data ($k = 1$)

Is it easier to estimate μ or σ^2?

- Notice, again, that for this smoothed estimate, contrary to the standard one, the sampling variance does not go to 0 for n going to infinity.
- No problem: we are not looking for consistent estimates of a (probably) time varying quantity
- On the other hand, the bigger k the smaller the sampling variance for a GIVEN n.
- Remember, however, that k cannot be taken as big as you wish as the log random walk hypothesis becomes untenable for very short time intervals between observations.

Summary: Is it easier to estimate μ or σ^2?

- Estimate for μ

$$\hat{\mu} = \sum_{i=1,\ldots,n} r_i^* / n$$

- Estimates of σ^2

$$\hat{\sigma^2} = \sum_{i=1}^{n} \frac{(r_i^* - \mu)^2}{n}$$

$$S^2 = \sum_{i=1}^{n} (r_i^* - \bar{r}^*)^2 / (n-1)$$

$$V_t = (1-\lambda) \sum_{i=0,\ldots,n} \lambda^i r_{t-i}^{*2}$$

Summary: Is it easier to estimate μ or σ^2?

- Under the log random walk hypothesis the first three estimates are unbiased, while the smoothed estimate is biased (upward) if $\mu \neq 0$

Summary: Is it easier to estimate μ or σ^2?

- The sampling variances are (Gaussian hp, $0 < \lambda < 1$):

$$V(\hat{\mu}) = \sum_{i=1,\ldots,n} V(r_i/n) = \frac{n}{n^2}\sigma^2 = \sigma^2/n$$

$$V(\hat{\sigma^2}) = \frac{2}{n}\sigma^4$$

$$V(S^2) = \frac{3\sigma^4}{n} - \frac{\sigma^4(n-3)}{n(n-1)} \cong \frac{2}{n}\sigma^4$$

$$V(V_t) = \frac{1-\lambda}{(1+\lambda)}2\sigma^4$$

Summary: Is it easier to estimate μ or σ^2?

- Let us now compute the ratios between the parameters and the standard deviations of their estimates
- For μ and considering μ/σ ratios for single stocks around .25 (optimistic)

$$\frac{\mu}{\sigma/\sqrt{n}} = \frac{\mu\sqrt{n}}{\sigma} \cong .25\sqrt{n}$$

- For σ^2 (only the non smoothed estimates)

$$\frac{\sigma^2}{\sqrt{\frac{2}{n}\sigma^4}} = \frac{\sigma^2\sqrt{n}}{\sigma^2\sqrt{2}} = \frac{\sqrt{n}}{\sqrt{2}} \cong .707\sqrt{n}$$

Summary: Is it easier to estimate μ or σ^2?

- This means that, with 9 years of data, we may expect a ratio for the estimate of the expected value around .75 while for the variance this shall be (obviously) almost three times bigger, that is around 2.12
- The "almost three times" rule, is clearly valid whatever n
- Moreover, if we divide each year in k subperiod of length $1/k$ and use the corresponding returns, the sampling variance of $\hat{\mu}$ does not change while the sampling variance of the estimate of the variance becomes

$$V(\hat{\sigma'}^2) = \frac{2}{kn}\sigma^4$$

- So that the ratio between the expected value and the standard deviation of the variance estimate becomes

$$\frac{\sigma^2}{\sigma^2}\frac{\sqrt{kn}}{\sqrt{2}} \cong .707\sqrt{kn}$$

- Where each of the n years of data is multiplied by k.

Summary: Is it easier to estimate μ or σ^2?

- With 9 years of monthly data as before this ratio becomes now $\cong 2.12\sqrt{12} = 7.34$.
- With 9 years of weekly data this is $\cong 2.12\sqrt{52} = 15.28$.
- But even with just one year of DAILY data this is $\cong .707\sqrt{256} = 11.312$.
- The corresponding ratios for the estimate of μ are, as before, about .75 in the first and the second case and .25 for a single year
- (This is, simply, the expected value / standard deviation ratio for a single yearly observation).
- Indeed: it is MUCH easier to get a good estimate of σ^2 than of μ

Summary: Is it easier to estimate μ or σ^2?

- For the smoothed estimate the sampling variance is

$$V(V_t) = \frac{1-\lambda}{(1+\lambda)} 2\sigma^4$$

- Which does not go to 0 with n increasing because we progressively "forget" (discount) the data
- The faster we forget (the smaller is λ) the bigger is the variance.
- If $\lambda = 0$ we are actually only just using one observation for the estimate (the most recent) so that the variance becomes $2\sigma^4$

Summary: Is it easier to estimate μ or σ^2?

- If I divide each year on k subperiods and interpolate the λ^i used for yearly data with $\lambda^{i/k}$ for each subperiod we improve the quality of the estimate.
- In this case the variance of the estimate shall be

$$V(V_t') = \frac{1-\lambda^{1/k}}{1+\lambda^{1/k}} 2\sigma^4$$

Summary: Is it easier to estimate μ or σ^2?

- The ratios between $\sigma^2$and the sampling standard deviation of its estimate are, in this case

$$\frac{\sigma^2}{\sigma^2\sqrt{2}}\sqrt{\frac{1+\lambda}{(1-\lambda)}} \cong .707\sqrt{\frac{1+\lambda}{(1-\lambda)}}$$

$$\frac{\sigma^2}{\sigma^2\sqrt{2}}\sqrt{\frac{1+\lambda^{1/k}}{(1-\lambda^{1/k})}} \cong .707\sqrt{\frac{1+\lambda^{1/k}}{(1-\lambda^{1/k})}}$$

- For $\lambda = .95$ the first ratio is $\cong 4.41$ and, if we use monthly data, that is: $k = 12$ the second becomes $\cong 15.29$.
- This seems even better than the classic (non smoothed) estimate.
- Do not be fooled!!!
- Since we used the limiting form of the denominator, we are implicitly assuming a very large n (infinite in the limit) and, under this assumption, the sampling variance of the "standard" estimate goes to 0.

Is it easier to estimate μ or σ? (not for the exam)

- In this lecture we discussed in terms of variance. Is what we said still valid with standard deviation?
- It is possible to show that, under hypotheses, $V(g(X)) \approx g'(E(X))^2 V(X)$
- If we apply this result to, say, $X = \hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(r_i^* - \mu)^2}{n}$, $E(X) = \sigma^2$ and $g(X) = \sqrt{X}$ with $g'(X) = \frac{1}{2} X^{-1/2}$ we get:
- $V(\sqrt{\hat{\sigma}^2}) = V(\hat{\sigma}) \approx (\frac{1}{2}(\sigma^2)^{-1/2})^2 \frac{2}{n}\sigma^4 = \frac{1}{4}\frac{1}{\sigma^2}\frac{2}{n}\sigma^4 = \frac{1}{2n}\sigma^2$
- The ratio between σ and the approximation of its estimate standard deviation is $\sigma / \sqrt{\frac{1}{2n}\sigma^2} = \sqrt{2n} = 1.4142\sqrt{n}$ twice better than the $.707\sqrt{n}$ valid for the variance
- The quality of the approximation $V(g(X)) \approx g'(E(X))^2 V(X)$ improves with n
- For the case of k observations per each year (or any unit of time), just put kn in the place of n

Lecture: Non Gaussian Returns

Summary

- Violations to the (Gaussian) log random walk hypothesis
- The meaning of "heavy tails"
- Rare events (and why the central limit theorem may not hold)
- Alternatives to the Gaussian distribution

Non Gaussian returns

- The history of the study of the distribution of "security prices differences" and, later "log security prices differences", is fascinating as the story of many enquiries in finance
- Many empirical and theoretical papers, during the fifties and the sixties (and we even have older references) did find that prices differences or log returns do not seem to be distributed according to the Gaussian distribution
- This was puzzling because the idea was, and still is, that log returns (or prices differences) are iid and very little statistically dependent
- The central limit theorem would imply that, at least over not very short term horizons, the distribution of sums of log returns should be Gaussian

Non Gaussian returns

- This requires log return to possess finite first and second moment AND the sums to comprise a non stochastic number of terms
- Be it as it may, empirically, and even at long enough scales, the Gaussian distribution was seen as only fitting historical return (in fact, sometimes, price differences) data "passably", NOT "well"
- In particular, compared to the Gaussian distribution, the actual distribution of log returns shows too many extreme events (we shall see in a moment the exact meaning of this)
- All of a sudden, during the seventies this discrepancy and its implications were put aside (at least, in the majority of papers on financial modelling) and, while still present in the data, did not come back to dominate the academic stage...
- ...until Monday, Oct 19th 1987 (https://ritholtz.com/2012/10/25th-anniversary-black-monday-1987-crash/)

Non Gaussian returns



The New York Times

NEW YORK, TUESDAY, OCTOBER 20, 1987

STOCKS PLUNGE 508 POINTS, A DROP OF 22.6%; 604 MILLION VOLUME NEARLY DOUBLES RECORD

U.S. Ships Shell Iran Installation In Gulf Reprisal

Offshore Target Termed a Base for Gunboats

A Huge Blow to the Five-Year Bull Market

Dow's Record Fall

WORLDWIDE IMPACT

Frenzied Trading Raises Fears of Recession — Tape 2 Hours Late

Does 1987 Equal 1929?

Who Gets Hurt?

Non Gaussian returns



THE WALL STREET JOURNAL.

© 1987 Dow Jones & Company, Inc. All Rights Reserved.

★★ EASTERN EDITION TUESDAY, OCTOBER 20, 1987 PRINCETON, NEW JERSEY ···· 50 CENTS

What's News—

Business and Finance

THE STOCK MARKET CRASHED as panic selling swept the Dow Jones industrials down 508.00 points, or 22.6%, to 1738.74. The record decline far exceeded the drop on Oct. 28, 1929, when the average slid 12.8%. Most other market indicators also skidded to record lows, as Big Board volume soared to 604.3 million shares, well above the previous record.

Economists don't expect the market's debacle to usher in a depression, mainly due to safeguards in the banking and financial systems. But many small investors remain worried.

(Stories in Columns 6, 3 and 4)

* * *

Treasury bond prices rallied from an early slide, winding up with huge gains in heavy trading. Some 30-year Treasury issues closed up more than three points. Gold surged to 4½-year highs, though a broad range of other commodity prices declined.

(Stories on Pages 39 and 42)

* * *

The plunge by London stocks could threaten a record $11.2 billion share offering by British Petroleum later this month. Meanwhile in France, the record drop on the Paris bourse could have political repercussions.

(Stories on Page 30)

* * *

Treasury Secretary Baker met with Bonn officials, and both sides agreed to support unspecified changes in the seven-nation accord on

World-Wide

THE U.S. BOMBARDED an Iranian oil platform in the Persian Gulf.

The strike by Navy warships, ordered by Reagan in retaliation for Friday's attack on a U.S.-flagged Kuwaiti oil tanker by Tehran's forces, destroyed the platform, which had been converted to a military command center in the gulf. The U.S. had broadcast warnings prior to the raid. American forces also destroyed some Iranian radar and communications equipment, according to the Pentagon. The attack was presented by the U.S. as a moderate military response, but one that demonstrated Washington's willingness to respond with force to Iranian provocation. (Story in Column 1)

Iran's president said that the U.S. attack was a "big mistake," and that Tehran would "take decisive retaliatory action." Moscow also faulted the action.

* * *

NAKASONE SELECTED Noboru Takeshita to succeed him as Japan's premier.

The choice by Nakasone came after the former finance minister and his two rivals were deadlocked over who should be named president of the ruling Liberal Democratic Party. Because the conservative party controls the parliament, the winner of the top party post also becomes prime minister. Although Japan's policies won't change radically under Takeshita, 63, his lack of foreign-policy experience could worsen Japan's international relations. (Story on Page 10)

The ruling party is expected to confirm Takeshita as party leader Oct. 31, when Nakasone steps down after five years as prime minister.

* * *

Indian troops captured a Tamil separatist stronghold in the town center of Jaffna after penetrating Sri Lankan rebel defenses, officials in New Delhi and Colombo said.

Market Fallout

The bears ran wild on Wall Street in the stock market's worst one-day plunge ever. The reaction by market professionals and traders was one of disbelief. Stories include:

- Trading on index futures at the Chicago Mercantile Exchange led a shaky market into a full-fledged panic, page 32.
- The lucrative boom in takeover stocks dried up, page 32.
- How some market bears got out of stocks before the plunge, page 32.
- Markets in Europe and Asia were battered. The Tokyo exchange was deluged with sell orders today, page 36.
- Wary investors are looking for a safe haven in Treasury bills and cash equivalents, page 42.
- Mutual funds struggled to keep up with redemptions, page 42.

The Market Debacle Rouses Worst Fears Of Little Investors

* * *

Many Are Beginning to Talk In Terms of Doomsday; Big Investors Are Braver

A WALL STREET JOURNAL News Roundup

The little investor doesn't know where to turn, although bigger ones are putting up a brave front.

"I'm scared," says Julie Ianotti, an executive secretary in Houston. "My stock is my nest egg, for a house or something. Should I sell? Tell me, should I sell?"

When Ms. Ianotti's mother suggested the price drop could portend a new depression, the young Ms. Ianotti at first scoffed at the idea. "I said, 'Yeah, sure, Mom. It'll go back up.' " But as the Dow Jones Industrial Average was taking a rec-

The Crash of '87

Stocks Plunge 508 Amid Panicky Selling

The Big Crash

DJIA at 15 minute intervals

NYSE volume on a half-hourly basis

(In millions of shares)

Source: Knight-Ridder Tradecenter

A Repeat of '29? Depression in '87 Is Not Expected

Percentage Decline Is Far Steeper Than '29; Bond Prices Surge

NEW YORK – The stock market crashed yesterday.

The Dow Jones Industrial Average plummeted an astonishing 508 points, or 22.6%, to 1738.74. The drop far exceeded the 12.8% decline on the notorious day of Oct. 28, 1929, which is generally considered the start of the Great Depression.

Panic-driven trading on the New York Stock Exchange reached 604.3 million shares, nearly double the price record volume of 338.5 million shares set last Friday, when the Dow plunged a then-record 108.35 points.

Commodities prices also skidded, except for precious metals and especially

This article was prepared by Wall Street Journal staff reporters Tim Metz, Alan Murray, Thomas E. Ricks and Beatrice E. Garcia.

gold, which surged $30.10 an ounce to $481.30, a 4½-year high. The bond market, a refuge for much of the capital being wrenched out of the stock market, recovered from steep early losses to end the day sharply higher.

Final Slide

The industrial average tumbled 130 points in the final 30 minutes of the session. The decline yesterday and last week

Non Gaussian returns

- This was quite an extreme event and had lasting implications for traders behaviour
- In this lecture we shall try to assess the meaning and relevance of this, and other, empirical characteristics of security prices

Non Gaussian returns

From Mandelbrot (1963) "The variation of certain speculative prices", Journal of Business.

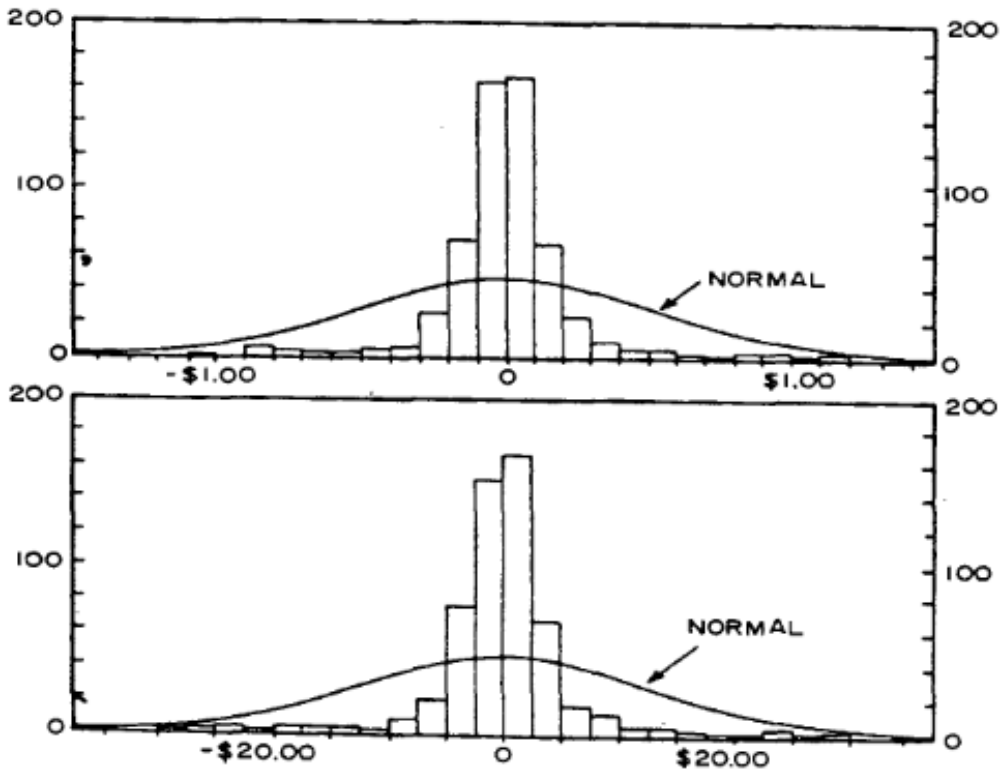


FIGURE E14-1. Two histograms illustrating departure from normality of the fifth and tenth difference of monthly wool prices, 1890-1937. In each case, the cohtinuous bell-shaped curve represents the Gaussian "interpolate" from -3σ to 3σ based upon the sample variance. Source: Tintner 1940.

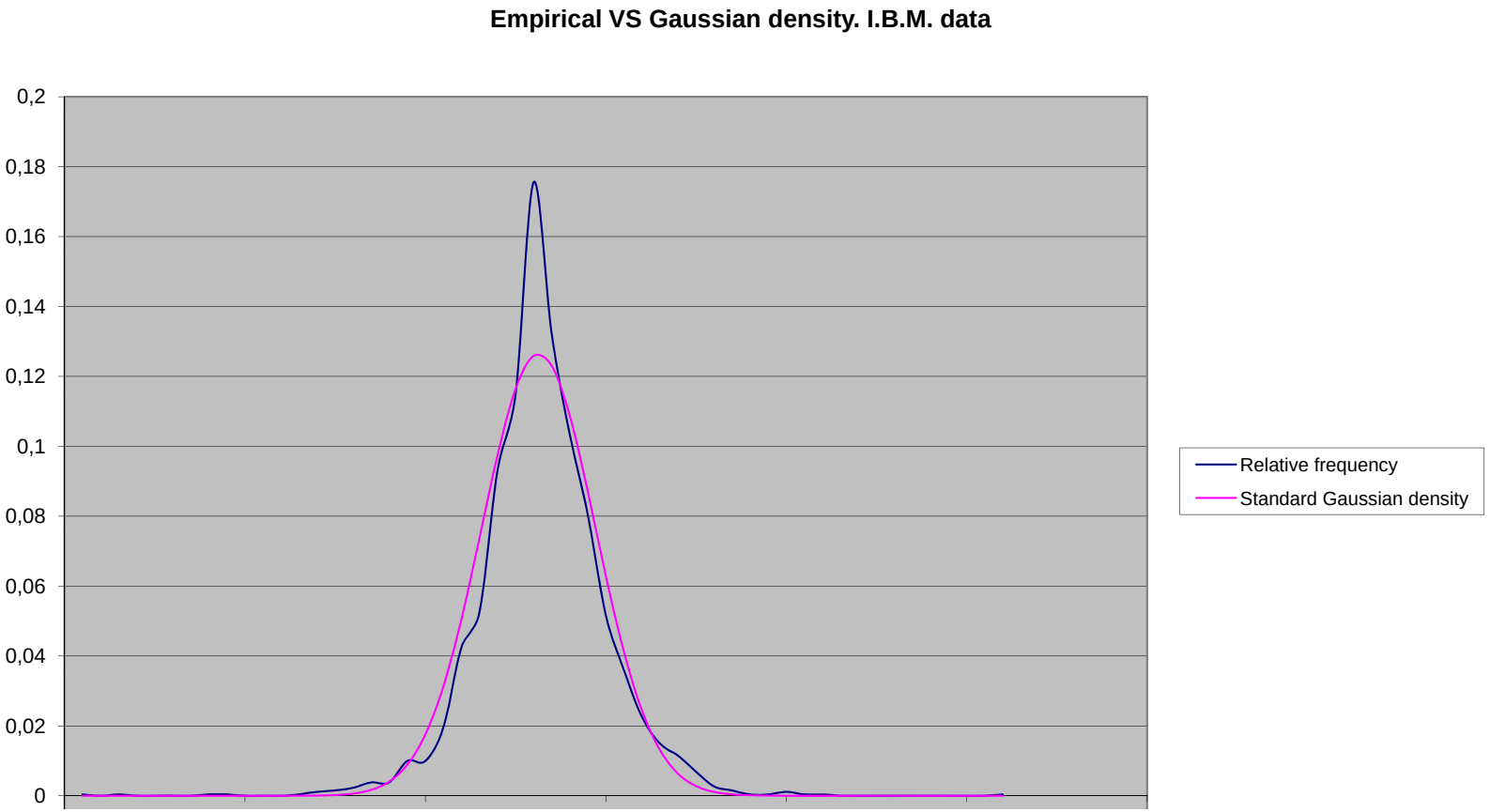
Non Gaussian returns

- When studying stock returns, we observe extreme events, more frequently negative, in the order of more that μ minus 5σ (whatever the estimate we choose for μ and σ provided is overall reasonable)
- The frequency of such events is incompatible with the probability of such or more negative events if this probability is computed under the hypothesis of Gaussianity.
- A different way to say this would be: in order to fit extreme event frequency with Gaussian probability, we should use estimates of σ far too big to correctly represent most return observations

Non Gaussian returns

- Let us consider an example based on a long series of I.B.M. daily returns.
- Let's take a smoothed frequency distribution and compare it with a Gaussian density (with parameters equal to the sample mean and std of data, see below for the details)

Non Gaussian returns



Non Gaussian returns

- At first sight not such a bad approximation, except in the CENTER of the distribution,... however:
- Between Jan 2nd 1962 and Dec 29th 2005 (10387 obs) the IBM daily returns show a standard deviation (classic estimate, in daily terms) of 0.0164 and an average of 0.000303
- (Note: the above plot is for the same series but between May 19th 1995 and Sep 28th 2005, see below)
- In the same time period, for 14 times the return was below $\mu - 5\sigma \cong -8.2\%$ in a single day .
- The number of observations is 10387 so the observed frequency of a $\mu - 5\sigma$ event (or more negative) is 0.00135=14/10387
- In a Gaussian distribution, the probability of a $\mu - 5\sigma$ or more negative observation is less than 0.00000028

Non Gaussian returns

- These extreme events are observed with a frequency which is, roughly, 4500 times the probability of such observations for a Gaussian with a reasonable estimate of expected value and variance
- This is true for a very "mature" and "conservative" stock the like of I.B.M.
- Obviously, a frequency of 0.00135 is very small, so the events are still infrequent
- However, these events (big crashes) are those which are remembered in the story of the market and may have huge consequences on market behaviour and regulation

Non Gaussian returns

- We should not concentrate on extreme events only, at first sight we see "too much frequency" in the middle of the distribution, if compared to the Gaussian
- Consider the frequency we observe between, say, $\mu - .5\sigma$ and $\mu + .5\sigma$
- In our IBM dataset about 46% of the observations lie in this interval
- In a Gaussian distribution this should be less that 40% (38.3%)
- Hence, we have an excess of observations clustered around μ, too

Non Gaussian returns

- Summary:
- While, at a first sight, the Gaussian is not an horrible approximation
- Empirical data show both "heavy tails" (in particular the left tail) and an excess of observations near the mean: "tall head"
- Since probability sums to one, we shall also observe "thin sides" (lack of observations for intermediate σ intervals)

Non Gaussian returns

- We can go in more detail
- There are other ways to see this
- In the Gaussian distribution, a probability of 0.00135 (the observed frequency of the more than -5σ event), corresponds to a distance from the expected value equal almost exactly to -3σ
- This would imply that a value of σ compatible with this frequency (for Gaussian data) is 0.0164*5/3=0.0273 that is a standard deviation which is 66% greater than the historically observed one
- However, we would never use such an estimate of σ as it would fit the extreme cases, but would imply an average volatility much bigger than the observed one

Non Gaussian returns

- It may also be that, just those "extreme" days, the expected value of the return became negative.
- On the other hand, to explain the excess of observations at the center of the distribution, we should have a Gaussian with a smaller variance
- And we can't have both!
- Either we suppose a sudden explosion of volatility, or a sudden drop of the expected return, or we must think that the distribution of returns is not Gaussian.
- Here we concentrate on the last possibility.

Non Gaussian returns

- The behaviour of the empirical distribution of returns can be summarized in the motto: fat tails, thin shoulders, tall head.

Non Gaussian returns

- What we observe is that, while overall the Gaussian interpolation of the histogram is not bad, if we zoom on the extreme tails: first and last two percent of data, we see that the tails of the histogram decrease at a slower rate than those of the Gaussian distribution.
- (This reminds us of a joke/principle by a famous statistician: Charles Winsor, as probably recounted by another great statistician: John Tukey: "all actual distributions are Gaussian in the MIDDLE!")
- Moreover, toward the center of the distribution, we see how the "shoulders" of the histogram are thinner than those of the Gaussian and, correspondingly, the histogram is more peaked around the mean.
- (hence, also the "MIDDLE" part of the joke does not hold here)

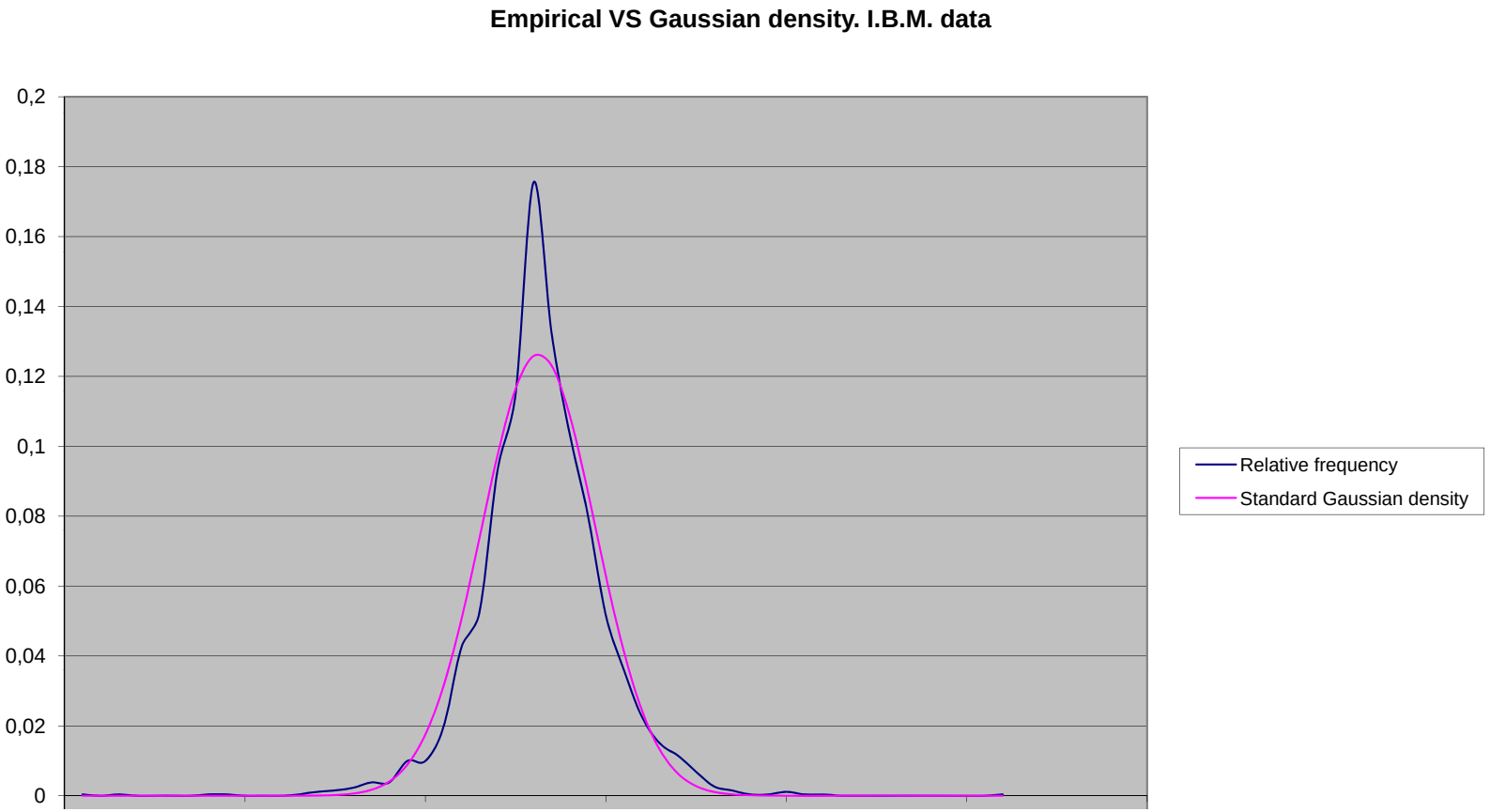
Non Gaussian returns

- Let us now look in more detail at the I.B.M. frequency data distribution we already considered, and some other useful plot
- The plots are from the excel file "Exercise 4 - Non Gaussian returns", in this worksheet we use a subset, from May 19th 1995 to Sep 28th 2005, of the same I.B.M. series as before.
- The plots are similar to Mandelbrot plots while more that 30 years pass between the end of his datase and the beginning of this
- There is strong persistence in this empirical observation

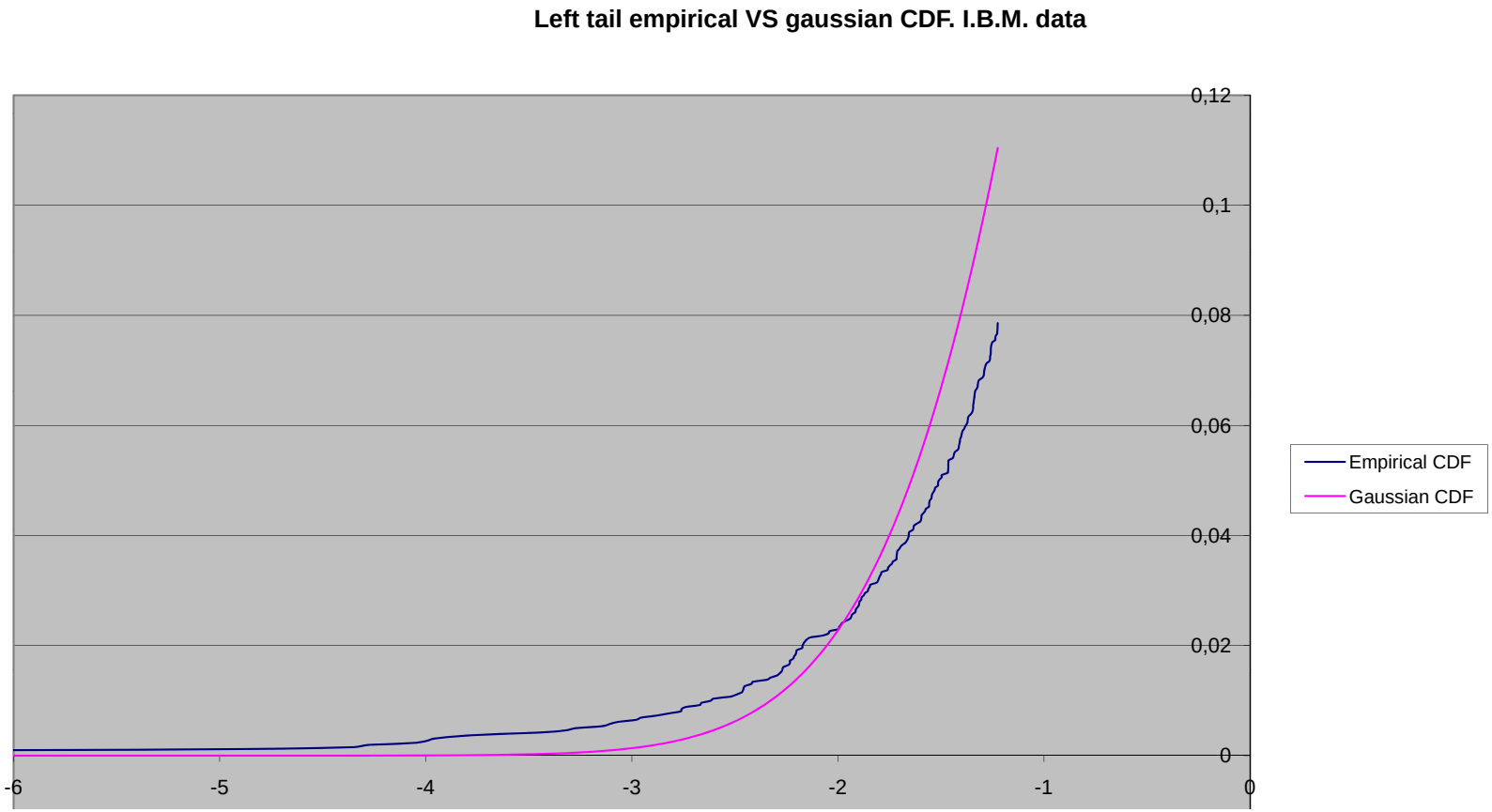
Non Gaussian returns

- The first plot, as above, compares the interpolated histogram of empirical data with a Gaussian density with the same mean and variance as the data.
- You can clearly see the mentioned "thin shoulders, tall head".
- Since tails, fat or not, are tails that is: they are thin, in the second plot we focus on the extreme left tail and at this scale the difference between the empirical and the Gaussian distribution.
- The x axis is in standard deviation units. At the left of, roughly, -2, the empirical tail is above the Gaussian tail:
- Extreme observations are more frequent then what we would expect in a Gaussian distribution with the same mean and variance as the data.

Non Gaussian returns



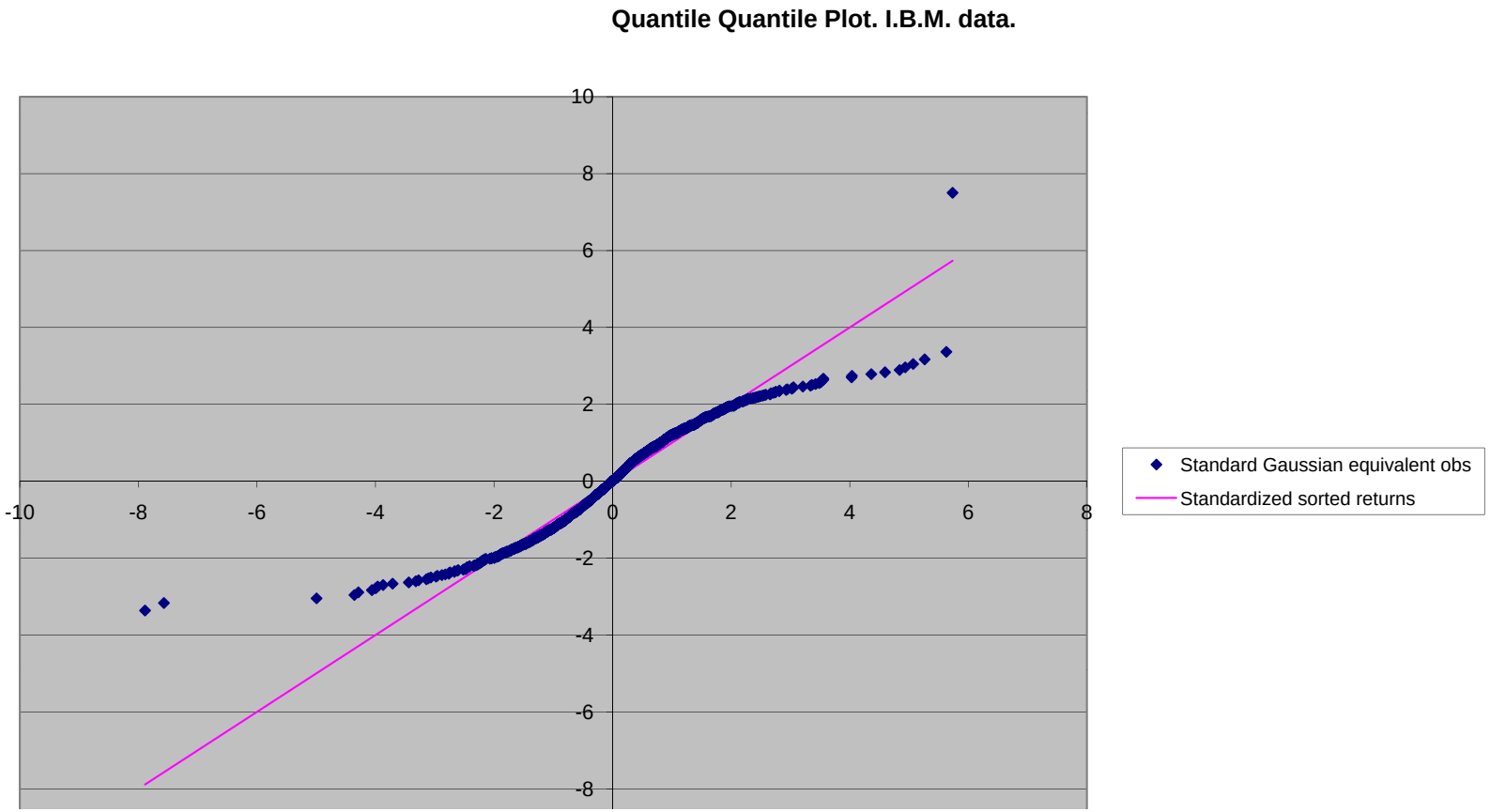
Non Gaussian returns



Non Gaussian returns

- Another way to compare the empirical distribution with a Gaussian model (or any model you may choose) is the Quantile-Quantile (QQ) plot.
- Here we plot the observed quantiles (diamonds) on the corresponding Gaussian quantiles (line)
- A good fit should correspond to diamonds overlapping the line

Non Gaussian returns



Non Gaussian returns

- For the I.B.M. data we see how, on the left tail, observed data are above the diagonal meaning a left tail heavier than the Gaussian.
- On the opposite side of the plot we see how the QQ plot lies below the bisecting line.
- Again: this means that we are observing data far from the mean and on the right side of it with an higher frequency than compatible with the Gaussian hypothesis.
- Since the data are standardized, the scale of the plot is in terms of number of standard deviations.
- We see that, on the left tail, we even observe data near and beyond -6 times the standard deviation.
- The tail from minus infinity to -6 times the standard deviation contains a probability of the order of 5 divided by one billion for the standard Gaussian distribution.

Non Gaussian returns

- In this dataset of 2610 obs we also observe 2 data points on the leftmost $\mu - 7\sigma$ tail.
- If we read our data as the result on independent extractions from the same Gaussian, these observations, while possible, are quite unlikely
- The probability of observing 2 times, in 2610 independent draws, something which has in each draw a probability of 0.000000000012798 to be observed, is virtually 0.
- It is a good exercise to compute this probability

Non Gaussian returns (exercise)

- Use the binomial distribution.
- Question: suppose the probability of observing a -7σ in each of 2610 independent "draws" is 0.000000000012798 .
- What is the probability of observing 2 such events?
- The answer, computed with Excel is:

$$\binom{2610}{2} 0.000000000012798^2 (1-0.000000000012798)^{2608} =$$

$$= 0.00000000000000055767$$

- or $55767 * 10^{-22}$. At the precision level of Excel, we have a bona fide 0

Non Gaussian returns

- We can also follow a different, if strongly related, line of thought.
- We see that in this dataset made of 2610 daily observations we observe a extreme negative return of around -8σ.
- This is the most negative (standardized) return (the exact value is -7.88σ) hence the minimum observed value.
- Now let us ask the following question: what is the probability of observing such negative a minimum if data come from a Gaussian?
- Suppose data are iid and distributed according to a (standardized) Gaussian.
- In this case the probability of observing data below the minus 8 sigma level is $\Phi(-8)$ and this for each of the 2610 observations.

Non Gaussian returns

- The probability of observing AT LEAST a value less than or equal this is 1 minus the probability of never observing such value, that is, due to iid: $1-(1-\Phi(-8))^{2610}$.
- It is clear that $1-\Phi(-8)$ is almost 1 but $(1-\Phi(-8))^{2610}$ is smaller.
- Is it small enough to make $1-(1-\Phi(-8))^{2610}$ big enough so that a minimum value of -8σ over 2610 iid observation from a standard Gaussian be not termed "anomalous"?
- The computation is numerically not too simple as $\Phi(-8)$ is a VERY small number, Excel gives for this 6.22096E-16, but the precision of the Excel routine for its computation cannot be guaranteed.

Non Gaussian returns

- However using the Excel value we get $(1-\Phi(-8))^{2610} = 0.99999999998261$ so that even if we take into account the 2610 observations the probability of observing as minimum of the sample a -8σ data point (the complement to 1 of 0.99999999998261) is still not really different from 0.
- I checked the result using Matlab (whose numerical routines should be more precise than Excel's) getting a very similar result.
- In order to get $1-(1-\Phi(-8))^n$ in the range of .01 (still very unlikely) we would need n =15,000,000,000,000.
- These are open market days and would correspond to roughly to 59 billions of years.
- This is a span more than 4 times (almost 5 times, actually) the current estimate of the age of our universe. (Again: beware of roundings!).

Non Gaussian returns

- In any sense, we must conclude, a -8σ value, while possible, is quite unlikely ...
- ...IF data come from a standard Gaussian, with constant mean and variance equal to the estimated ones.
- The point is that, as we have seen, given the variance if you want more probability on the tails you must allso put more probability near the center of the distribution
- I practice, if you want heavy tails with given variance, you need a probability distribution which tells you that "most of times" what you observe is very near the expected value
- This is called a "low entropy" state
- It happens that the Gaussian distribution, at the opposite, maximizes entropy among all continuous distributions on $\mathbb{R}$ with a given vraiance
- In this sense, the Gaussian is a very peculiar and absolutely non "normal" distribution (but this fact is common knowledge among statisticians)

Which distribution? A statistical dilemma (not for the exam)

- The fact that the Gaussian distribution does not fit well the empirical distribution of log returns may be a big problem
- In particular, we have problems for risk management and for option pricing
- As we have seen the Gaussian distribution, if fitted to mean and variance of the empirical data, strongly undervalues the probability of extreme (negative but also positive) observations

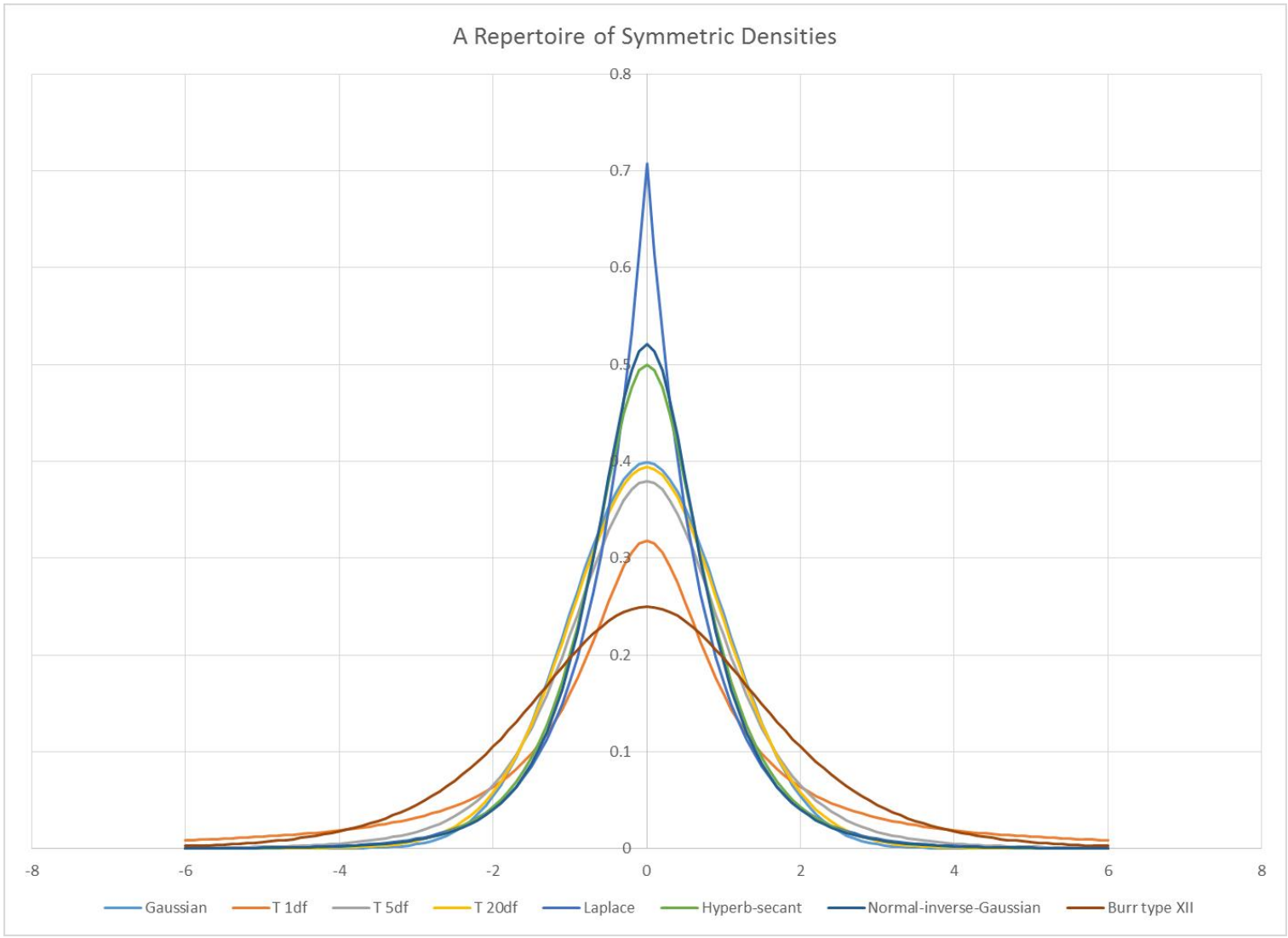
Which distribution? A statistical dilemma (not for the exam)

- Any risk management system requires a correct estimation of such probabilities
- Option pricing requires a careful evaluation of the tails of the log return distribution as on these mostly depends the value of very out of the money options
- It may be that the Gaussian, due to some Central Limit Theorem effect, works decently for well diversified portfolios or for log returns over longer time span (more about this in what follows)
- However, since the beginning of RiskMetrics, need was felt for alternative models (or for some model independen "non parametric" estimate)

Which distribution? A statistical dilemma (not for the exam)

- Statistics is by no means at need of models
- Other distributions exist (quite a lot, in fact a full vocabulary we should learn, as it would be difficult to speak "statistical" without words)
- In the following slide you can see the densities of a small choice of these is the subset of symmetric densities
- (warning: expected values and variances are not the same in these plots and, for some of the densities, the expected value and the variance may not exist)

Which distribution? A statistical dilemma (not for the exam)



Which distribution? A statistical dilemma

- In what follows we concen
- trate on the properties of a simple alternative model
- This is a simple example of a general class of "heavy tailed densities" which we shall further consider when dealing with semiparametric Value at Risk estimates

Which distribution? A statistical dilemma

- The alternative we discuss here is the T distribution.
- In its non standardized form, this has density (not for the exam)

$$f(x,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{s\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-a)^2}{s^2\nu}\right)^{-\frac{\nu+1}{2}}$$

- Where ν, a and s are parameters such that: ν is a positive integer (common name "degrees of freedom"), a is a real number and s is a positive real number

Which distribution? A statistical dilemma (not for the exam)

- The density is, obviously, symmetric around a
- The moments of this distribution exist only up to order $\nu - 1$
- The first moment (expected value), which exists if ν is at least equal to 2 is a
- The second centered moment (variance) exists if ν is at least 3 and is equal to $s^2\nu/(\nu-2)$

Which distribution? A statistical dilemma

- The, maybe daunting, term $\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{s\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}$ is there just so that the integral of the density be 1
- The "live" part of the density (technically, we call this "kernel" of the distribution) is $\left(1+\frac{(x-a)^2}{s^2\nu}\right)^{-\frac{\nu+1}{2}}$
- This is just the reciprocal of a polynomial in x (a case of "rational function")
- When $|x-a|$ increases, this function monotonically decreases at "speed" proportional to $|x-a|^{-(\nu+1)}$ this means that if I multiply it by, say $|x-a|^q$ with $q \geq \nu$ it becomes of the order $|x-a|^{-1}$so that it is not integrable over the real line
- For this reason moments exist only up to order $\nu-1$

Which distribution? A statistical dilemma

- It is an example of "polynomial tails" density meaning that the tails go to zero as fast as the reciprocal of a polynomial
- Notice that the Gaussian always possess all moments because its tails decrease at speed proportional to $e^{-\frac{x^2}{2}}$ (faster than any rational function)
- We shall use distributions with polynomial tails for computing the VaR (semiparametric method, very popular today)

Which distribution? A statistical dilemma (not for the exam)

- The origin of the T distribution is both suggenstive, if we want to apply it in Finance, and interesting:
- It is a "Inverse Gamma mixture of Gaussians"
- This may seem cryptic but think about the following log return generating mechanism

Which distribution? A statistical dilemma (not for the exam)

- In the interval between the observation of two prices (and suppose this interval is given and short, say: one day) many different things may happen
- Information may be very precise or very uncertain (changing, contradictory, conflicting, obscure etc.) and this may have consequences on log return volatility
- We may suppose that, the greater the uncertainty of the implications of new information the bigger the volatility of the return distribution (uncertainty on new information may be seen as a risk)
- Given this, if we knew the exact implications of information on log return volatility, we would model the distribution of log return as a Gaussian
- But its volatility is random, due to info uncertainty, and we do not observe it
- What we observe, and must model, are log returns so that we need a model for the marginal distribution of such returns

Which distribution? A statistical dilemma (not for the exam)

- As usual, the marginal distribution is the expected value of the conditional distribution where we use as dinsity for computing the expected value the density of the variable we want to "marginalize out"

$$h(r^*) = \int_0^\infty f(r^*|\sigma^2) g(\sigma^2) d\sigma^2$$

- Here: $f(r^*|\sigma^2)$ is the Gaussian density, which describes log returns conditional to their σ^2 and $g(\sigma^2)$is the distrubution of the random variance

Which distribution? A statistical dilemma (not for the exam)

- Let g be chosen as an "inverted Gamma"
- That is: the distribution of $1/\sigma^2$ is a Gamma
- This is a generalized version of the negative exponential with expected value $\nu s^2/(\nu-2)$ and standard deviation $\nu s^2/((\nu-2)\sqrt{\nu-4})$ (they exist if $\nu > 4$ but do not bother with this now)
- With this choice we get as $h(r^*)$ (the marginal or "mixture" distribution of log returns) the required T

Which distribution? A statistical dilemma (not for the exam)

- Notice that, for small ν, the expected value of $g(\sigma^2)$ is bigger and, for big ν, it goes down to s^2
- The standard deviation of $g(\sigma^2)$ behaves in the same way: is big if ν is small and goes to 0 if ν is big
- In other words: if ν is small we have a bigger uncertainty on the value of σ^2 and its expected value is bigger
- When ν grows the expected value of $g(\sigma^2)$ converges to s^2 and most of $g(\sigma^2)$ tightly concentrates around this value.
- In the limit, for $\nu \to \infty$, $g(\sigma^2)$ concentrates all probability on s^2 and $h(r^*)$ becomes a Gaussian with variance s^2
- This is exactly what happens to a T distribution of the usual T ratio if ν grows

Which distribution? A statistical dilemma (not for the exam)

- This is just a starting point:
- A clear limit of the model is that s^2 is a constant: the volatility of log return with "very precise" information, is always the same
- We can make it dependent on other variables which may express the "average uncertainty level" of the market
- We could also enrich the model by assuming, for instance, that the choice of ν depends on observable variables, presumably linked to the "quantity of new information"
- These could be, for instance, the number of trades in the time interval between the two prices or the volume of traded shares and so on

Which distribution? A statistical dilemma

- There is no time for discussing the fitting of this distribution to our data.
- However, try $\nu = 5$ and let us see how our T fits the empirical frequencies of extreme tails and of observations near the centre of the distribution
- Recall that, for the gaussian, the CDF at -8σ is 6.22096E-16
- For the T with $\nu = 5$ the same probability is 0.000246453

Which distribution? A statistical dilemma

- This means that the probability of never observing a return of -8σ over 2610 days, if data come from this T, is 0.826 (for the Gaussian was: 0.999999999998261)
- The probability of observing, over 2610 days, one or more -8σ value (or more negative) is, then, 0.174 that is: much much much... more likely that in the Gaussian case (0.000000000001739)
- The tails seem "heavy" enough

Which distribution? A statistical dilemma

- Head and shoulders:
- In the big IBM dataset, the frequency of observations between $-.5\sigma$ and $+.5\sigma$ around the mean was 46% while it should have been less that 40% (38.3%) for a Gaussian
- You'll be amazed to hear that, for a 5 degrees of freedom T distribution, the probability observing data in this interval is about 45%
- In doing all these computations, recall that the Student's T distribution variance is $\nu/(\nu-2)$ where ν is the number of degrees of freedom so that, for instance, the quantile corresponding to -8 in a standard Gaussian is, now, $-8\sqrt{\nu/(\nu-2)}$

Which distribution? A statistical dilemma

- Probably, we are on the right track, there is, however, a problem (which we shall declare but not discuss)
- If we believe this model works for daily data and that daily log return are independent, it should be that the central limit theorem works across days
- Considering log returns over a span of many days, the Gaussian distribution should fit increasingly well
- There is no agreement on the fact that this actually does happen and, most likely, it doesn't
- At least not as fast as expected

Which distribution? A statistical dilemma

- In our (big) IBM dataset (more that 10000 obs) if we study log returns for 10 (market) days (2 weeks) we get a better agreement with the Gaussian distribution
- The agreement, however, is not satisfactory
- We still observe 5 datapoints below $\mu - 3\sigma$ for a frequency of .004541 (one of these is a -5\sigma point)
- The Gaussian probability should be about .00135 which is quite smaller but, at least ,comparable
- In the $\mu \pm .5\sigma$ interval the empirical frequency is 47.5% against the Gaussian probability of 38.3%

Which distribution? A statistical dilemma

- If we consider returns over 20 trading days (slightly less that one month), the Gaussian works better but still not quite well
- We still observe 2 datapoints out of 554 below $\mu - 3\sigma$ for a frequency of .00361 (one of these is a -5σ point)
- In the $\mu \pm .5\sigma$ interval the empirical frequency goes down but is still 45.3% against the Gaussian probability of 38.3%

Which distribution? A statistical dilemma

- It is not really meaningful to try for returns for 30 or 40 market days (we quickly exhaust even this very long series)
- Such a scarce success of the CLT is the more puzzling as the starting point was a good daily fit to T with $\nu = 5$
- Beware: the distribution of the sum of independent T distributed random variables is not T distributed (in fact is quite complicated)
- This not withstanding the CLT should begin to hold and we should observe something more similar to a T with $\nu = 50$ or 100 (bona fide Gaussians) that what we actually observe
- In fact, if you simulate 10000 iid log returns distributed according to a T with $\nu = 5$ and then sum them 10 and 20 at a time (non overlapping) the Gaussian fit is already quite good for $n = 10$ and almost perfect for $n = 20$

Which distribution? A statistical dilemma

- While this evaluation is quite approximate, we can say that, with true data, while things are moving in the right direction, they do this quite slower than expected
- In fact, with simulated data, the CLT "effect" is strong for $n = 10$ and almost perfect for $n = 20$
- If CLT does not fully work, maybe something more is at work in real world data

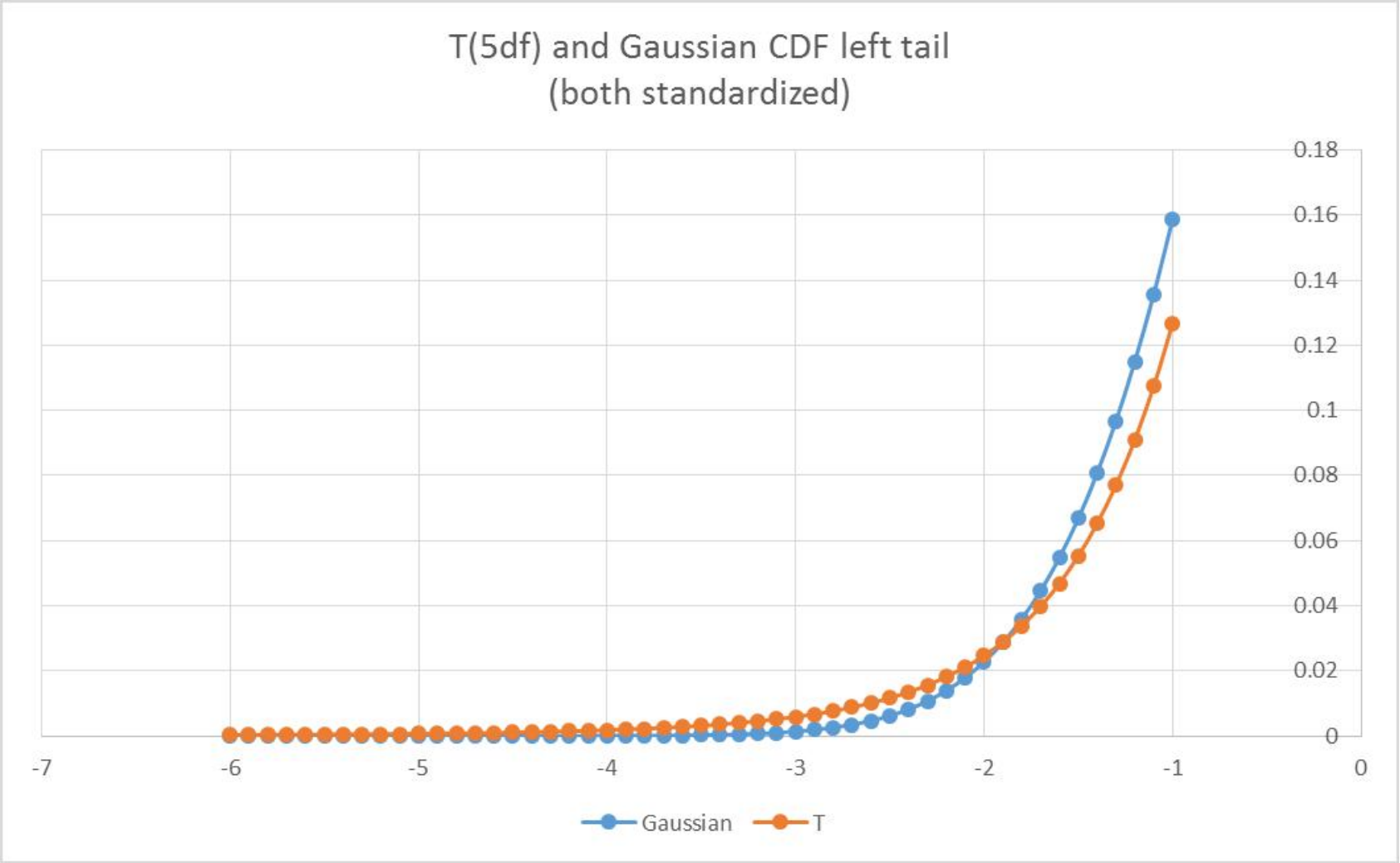
Which distribution? A statistical dilemma

- It may be that the process of information arrival is not stationary (we suppose an unchanging density for σ^2), it may be that there is correlation in information arrival
- It may be something completely different: correlated extreme events, jumps in return distribution etc.
- In any case, as you see, we are discussing about "usable" empirical models which already give a good performance and which we can try and improve
- The CLT effect is surprisingly slow, the T approximation of the daily log return distribution is quite good
- Moreover, this is not a useless data fitting exercise: by improving our models, we get a better understanding of stock returns behaviour

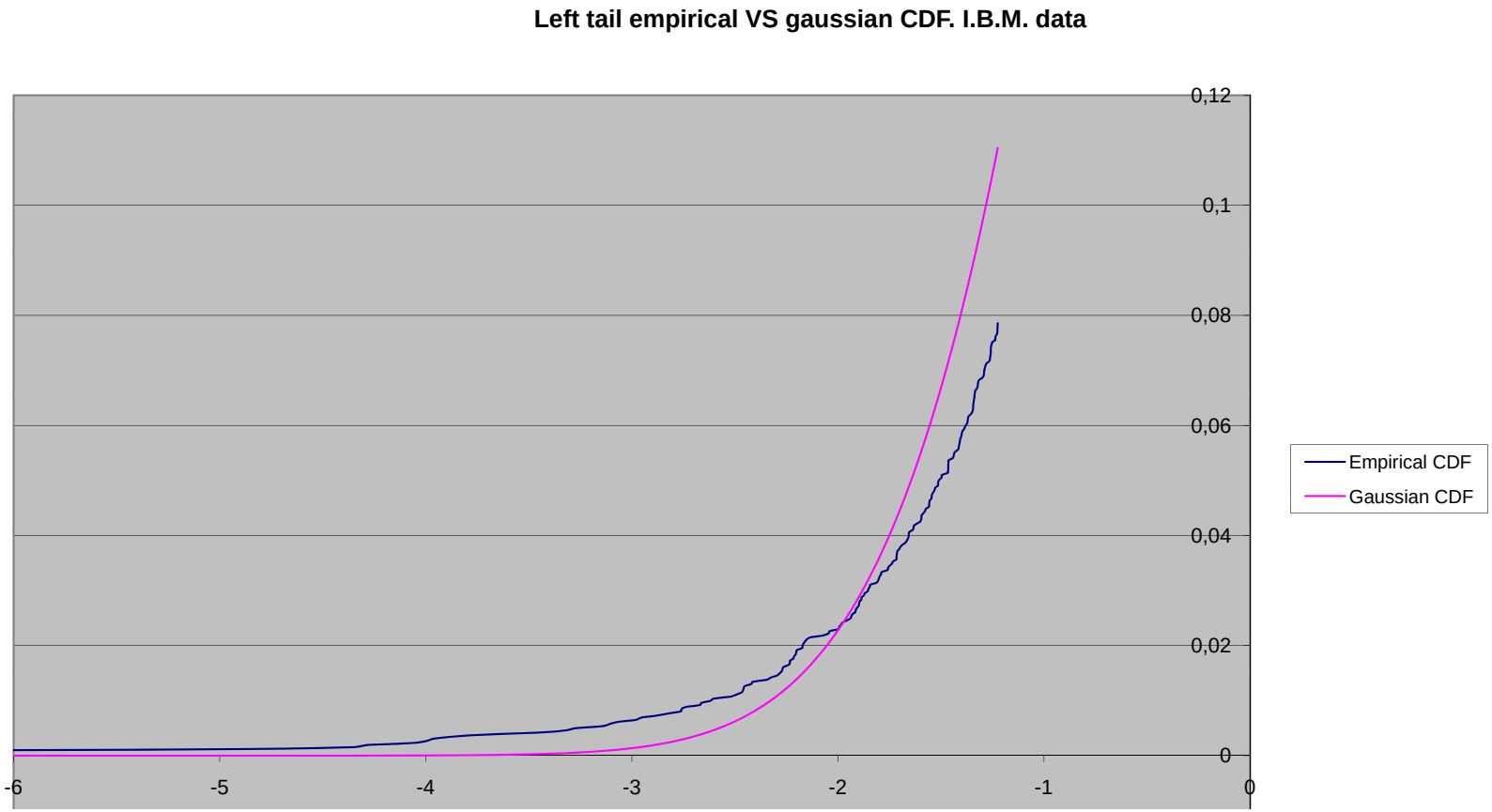
Which distribution? A statistical dilemma

- Now some plot
- We compare the empirical plots with the Gaussian VS T distribution with 5df plots
- (Note that, in the density plots, the vertical units are different)
- We see that the T 5df behaves, w.r.t. the Gaussian, in a very similar way as true return data
- While a low degrees of freedom T is not a full solution to the heavy tail problem, it shall surely be useful as an alternative to Gaussian scenario to test pricing models and risk management systems

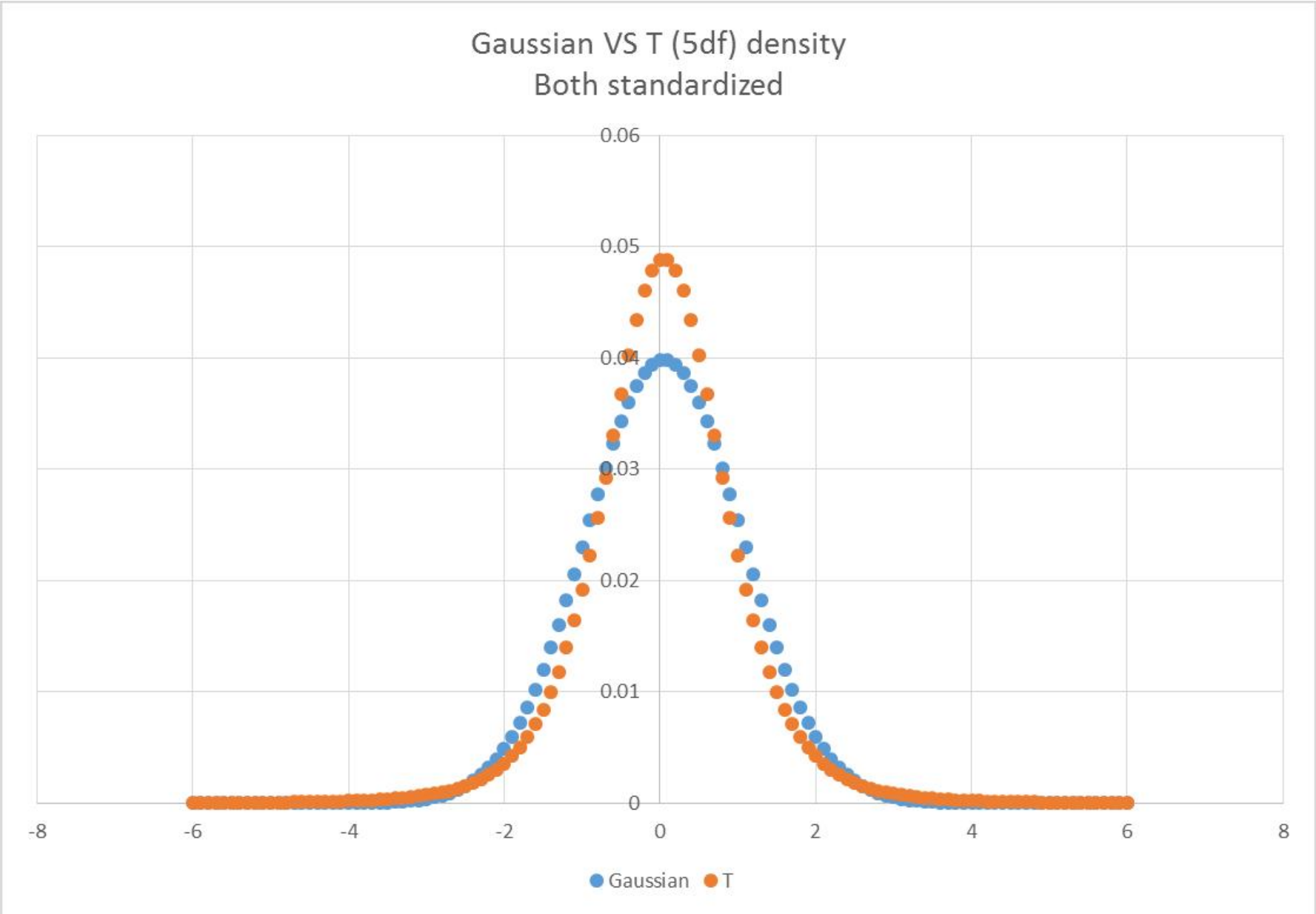
Which distribution? A statistical dilemma



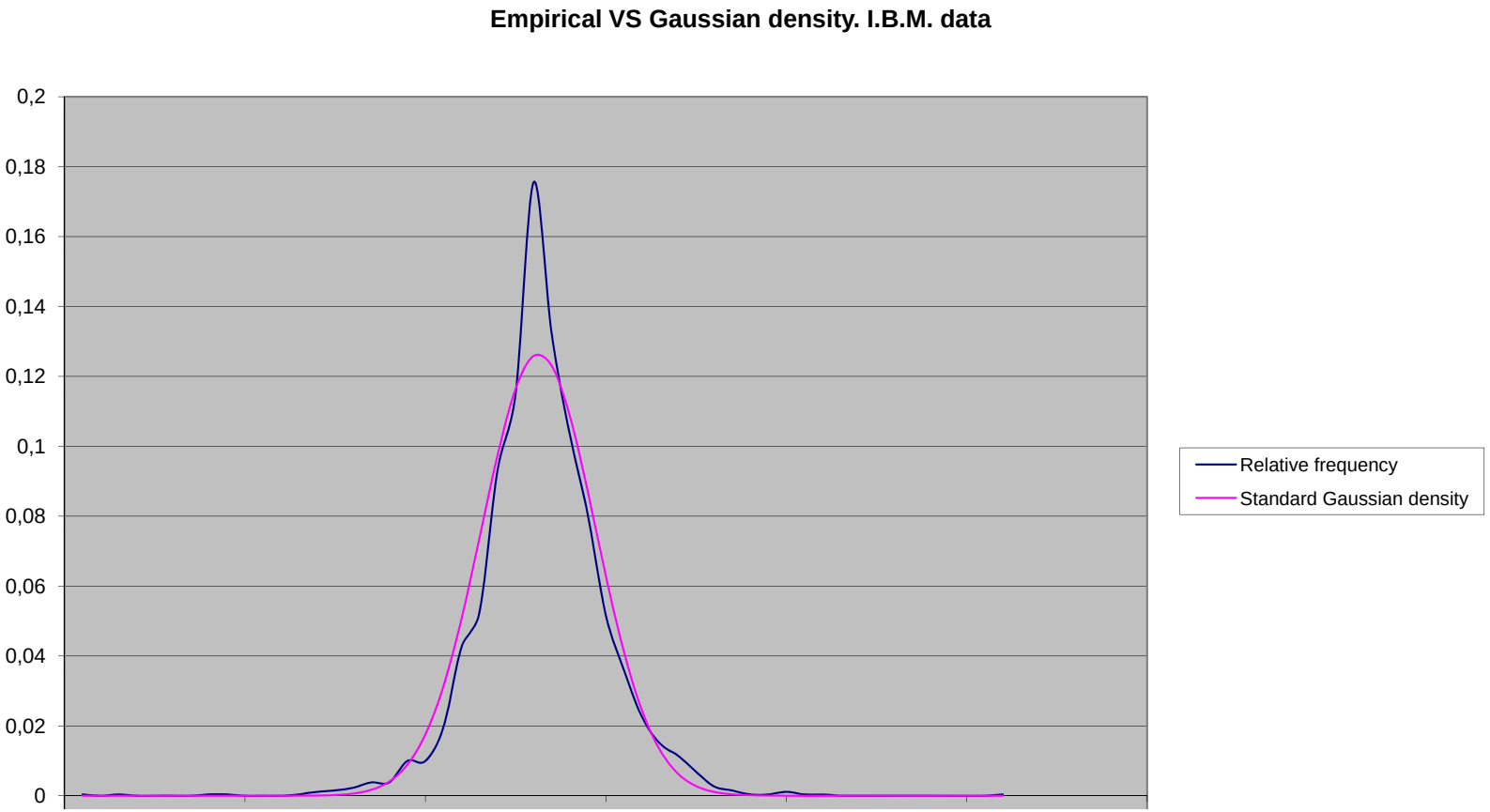
Which distribution? A statistical dilemma



Which distribution? A statistical dilemma



Which distribution? A statistical dilemma



Lecture: VaR, Definition and Parametric

Summary

- A measure of market risk VaR
- Its origin with RiskMetrics
- The classic version: parametric VaR
- Point estimate and confidence lower bound

VaR

- Three different ways for computing the VaR.
- First it is necessary to define the VaR.
- A sum W is invested in a portfolio at time t_0 and we are interested in the p&l (profit and losses) between t_0 and $t_1$that is: $W_{t1} - W_{t0}$.
- In all the (for us) relevant cases, that is when some "risk" is involved: this p&l shall be stochastic due to the fact that W_{t1}as seen at $t0$ is stochastic.
- Our purpose is to give a simple summary of such stochastic behaviour of the p&l aimed at quantifying our "risk" in a possibly immediate way.

VaR

- Many such measures can be (and have been) suggested.
- The RiskMetrics procedure chose as its basis the so called VaR "Value at Risk".
- Given a level α (usually very small: 1% to 5% as a rule) of probability, the VaR is defined as the α-quantile of the distribution of $W_{t1} - W_{t0}$.
- As you see VaR has three dimensions: invested sum (W_{t0}), time interval or "horizon" ($t0 \rightarrow t1$), probability (α)

- The definition of a α-quantile $x_\alpha$for the distribution of a random variable X is easy to write down and understand when the distribution of X : $F_X(x)$ is continuous a strictly increasing at least in an interval x_l, x_usuch that $F_X(x_l) < \alpha < F_X(x_u)$,
- In this case, we simply have

$$x_\alpha \equiv x : \quad P(X \leq x) = F_X(x) = \alpha$$

- and

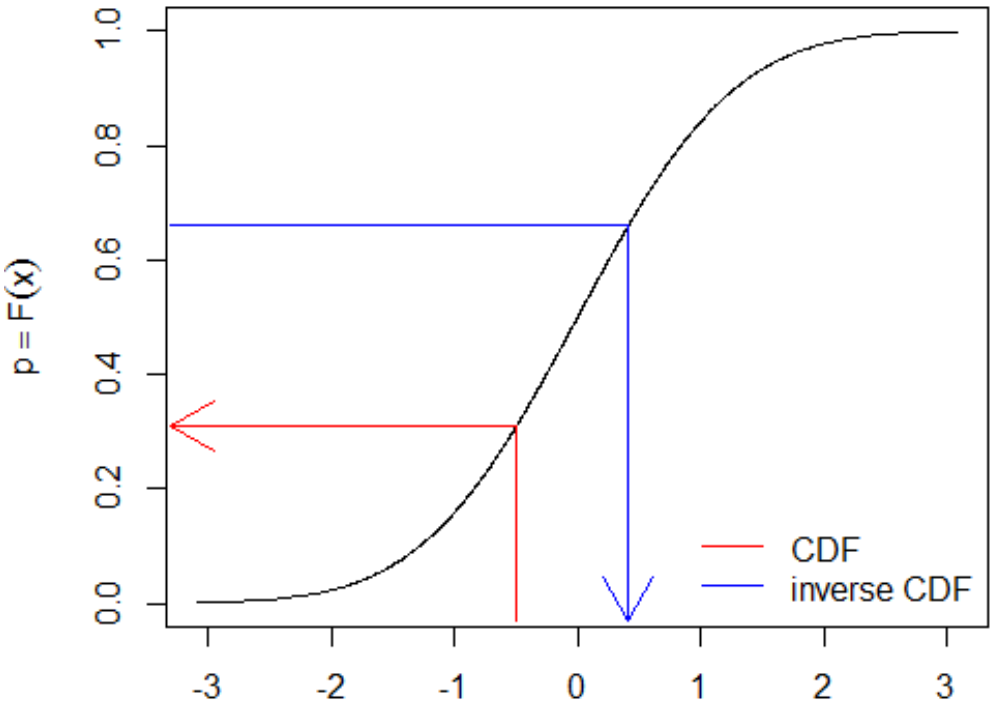
$$x_\alpha = F_X^{-1}(\alpha)$$

- Where the inverse of $F_X(x)$, that is:$F_X^{-1}(\alpha)$ is defined in a unique way, continuous and strictly increasing at least for $F_X(x_l) < \alpha < F_X(x_u)$.

VaR

- In this picture we see how a CDF can be used in two ways:

1. Given a value x find the value of the probability p which x leaves on its left: $p = F(x)$
2. Given a probability p, find the x which leaves on its left that probability: $x = F^{-1}(p)$ (x is the p-quantile)



- In the case F_X is continuous and increasing, the α-quantile is nothing but the value of X which corresponds to a cumulated probability exactly equal to α and we indicate such value with x_α.
- In the case of a cumulative distribution function with jumps (corresponding to probability masses concentrated in specific values of x) there may be no x such that $F_X(x) = \alpha$ for a given α.
- In this case the convention we use here is that of setting $x_\alpha$equal to the maximum of the values x of X such that x is of positive probability and $F_X(x) \leq \alpha$.
- Barring this possibility, it is correct to say that the VaR at level α for a time horizon between t_0 and $t_1$of your investment is that profit and loss such that the probability of observing a worse one is equal to α.

- This definition seems to imply that we are required to directly compute a quantile of the p&l.
- This is not the case.
- In fact what is required is a quantile of the return distribution.
- Indeed we have

$$W_{t1} - W_{t0} = W_{t0} r_{t0,t1}$$

$$W_{t1} - W_{t0} = W_{t0}(e^{r^*_{t0,t1}} - 1)$$

- Where $r_{t0,t1}$ and $r^*_{t0,t1}$ are, respectively, the linear and the log return in the time interval from t_0 to t_1.

VaR

- The problem of finding the required quantile of the p&l is equivalent to the problem of finding the same in the distribution of returns and transform it back to p&l.
- In this section we shall consider four different estimates of the VaR which rely on different sets of hypotheses.
- Each estimate shall be presented in a very simple form, the reader is warned that actual implementation of any of these estimates requires a detailed study of the return distribution and of the properties of each security or portfolio of securities
- However, for single stocks or portfolios of stocks, the actual procedure is very similar to what you shall see here

Gaussian, or Parametric, VaR

- Point estimate of the Gaussian VaR
- Here and in what follows we shall always consider quantiles of the return distribution, once found these have to be transformed in P&L quantiles
- A word of notice, in what follows we shall use R as the symbol meaning the random variable return and r as a possible value of such random variable.
- To avoid heavy notation we shall not indicate the kind of return we are speaking about or the time interval considered.
- Both these informations shall be clear from the context.

Gaussian, or Parametric, VaR

- This is the most restrictive setting used in practice.
- Since the origins of VaR with RiskMetrics, it is well understood that the use of the Gaussian distribution in computing a VaR should be limited to diversified portfolios of stocks and to time intervals of some length
- This is due to the lack of fit of the Gaussian in the case of extreme tails we already discussed in some detail

Gaussian, or Parametric, VaR

- We suppose that R is distributed according to a Gaussian density with expected value and variance (μ, σ^2) which are either known or estimated in such a way to minimize sampling error problems.
- A typical attitude is that of setting $\mu = 0$ and estimating σ^2, for instance, using the smoothed estimate described above.
- The important point to remember with the Gaussian density is that knowledge of mean and variance is equivalent to full knowledge of the distribution, hence of any quantile.
- Clearly, the Gaussian hypothesis is fully reasonable for log returns only. For linear returns it could be considered as an approximation

Gaussian, or Parametric, VaR

- The Reader shall remember of having been exposed to a number of beginner's models for decision under uncertainty, usually based on some mean-variance comparison.
- This could have been puzzling: after all, why should the decision maker be interested in these moments and not in the full shape of the probability distribution for returns?
- A possible justification of this attitude is that, if the return distribution is believed to be Gaussian, mean and variance are everything that matters.

Gaussian, or Parametric, VaR

- Under the Gaussian Hypothesis, the CDF is continuous so we can find a quantile with exactly α probability on its left for any α.
- The procedure is simple: we must find r_α such that:

$$P(R \leq r_\alpha) = \alpha$$

- Proceeding with the usual argument, already well known from confidence intervals theory, we get:

$$P(R \leq r_\alpha) = \alpha =$$

$$= P((R-\mu)/\sigma \leq (r_\alpha - \mu)/\sigma) = \Phi((r_\alpha - \mu)/\sigma) = \Phi(z_\alpha)$$

- Here z_α $(= -z_{1-\alpha})$ is the usual α quantile for the standard Gaussian CDF $\Phi(.)$.

Gaussian, or Parametric, VaR

- We have, then

$$(r_\alpha - \mu)/\sigma = z_\alpha$$

$$r_\alpha = \mu + \sigma z_\alpha$$

- Computing an estimate of this quantile is quite an easy procedure (if we have estimates of μ and σ) and we shall consider this in a moment.
- But remember: for small values of α we are considering quantiles very far on the left tail and our previous empirical analysis has shown how the Gaussian hypothesis for returns (overall not so bad) is inadequate for extreme tails.
- Typically the problem of fat tail shall imply a undervaluation of the VaR in the sense that, for a given, small α the Gaussian α quantile shall be less negative that the "heavy tail" one.

Gaussian, or Parametric, VaR

- Estimate and approximate confidence interval (lower bound) for the VaR
- Now a problem: we do not know both μ and σ.
- We must estimate them.
- The usual RiskMetrics procedure sets $\mu = 0$ and estimates σ with the smoothed estimates introduced above.
- The corresponding estimate of the quantile of the return distribution, that is of: $r_\alpha = \sigma z_\alpha$, shall then be: $\widehat{r}_\alpha = \hat{\sigma} z_\alpha$.

Gaussian, or Parametric, VaR

- According to sound statistical practice we should implement this joint with a measure of sampling variability.
- Here we show a possible approximate and simple way to do so by computing a lower confidence bound.
- (Bocconi's BAs did study an exact conficence interval for σ^2 but this is based on the standard estimate, not the smoothed estimate)
- Under the assumptions of uncorrelated observations with constant variance and zero expected value, it is easy to compute the variance of $\widehat{r}_\alpha^2 = \hat{\sigma}^2 z_\alpha^2$.
- In fact we have

$$V(\widehat{r}_\alpha^2) = V(z_\alpha^2 \hat{\sigma}^2) = z_\alpha^4 V\left(\hat{\sigma}^2\right)$$

Gaussian, or Parametric, VaR

- During the discussion about the different precisions in estimating $E(r) = \mu$ and $V(r) = \sigma^2$ we derived, for the Gaussian with, $\mu = 0$ and the smoothed estimate of σ^2 the formula

$$V\left(\hat{\sigma}^2\right) = \frac{\left(\sum_{i=0,\ldots,n} \lambda^{2i}\right)}{\left(\sum_{i=0,\ldots,n} \lambda^{i}\right)^2} 2\sigma^4$$

$$\frac{\left(\sum_{i=0,\ldots,n} \lambda^{2i}\right)}{\left(\sum_{i=0,\ldots,n} \lambda^{i}\right)^2} \approx \frac{1-\lambda}{(1+\lambda)}$$

$$V\left(\hat{\sigma}^2\right) = \frac{1-\lambda}{(1+\lambda)} 2\sigma^4$$

Gaussian, or Parametric, VaR

- We can simply plug this in our $V(\widehat{r}_\alpha^2) = z_\alpha^4 V\left(\hat{\sigma}^2\right)$ and get

$$V(\widehat{r}_\alpha^2) = z_\alpha^4 V\left(\hat{\sigma}^2\right) = z_\alpha^4 \frac{1-\lambda}{(1+\lambda)} 2\sigma^4$$

- We can then estimate the σ^4 term by taking the square of the estimate of the variance $(\hat{\sigma}^2)^2 = \hat{\sigma}^4$: and get

$$\hat{V}(\widehat{r}_\alpha^2) = z_\alpha^4 \frac{1-\lambda}{(1+\lambda)} 2\hat{\sigma}^4$$

Gaussian, or Parametric, VaR

- Using the "two sigma" rule (Tchebicev), and recalling that $\widehat{r}_\alpha^2 = z_\alpha^2 \hat{\sigma}^2$, an approximate upper bound for $\widehat{r}_\alpha^2$ is

$$\widehat{r}_\alpha^2 + 2\sqrt{\hat{V}(\widehat{r}_\alpha^2)} = \widehat{r}_\alpha^2 + 2\hat{\sigma}^2 z_\alpha^2 \sqrt{\frac{1-\lambda}{(1+\lambda)}2} =$$

$$\widehat{r}_\alpha^2 + 2\widehat{r}_\alpha^2 \sqrt{\frac{1-\lambda}{(1+\lambda)}2} = \widehat{r}_\alpha^2 \left[1 + 2\sqrt{\frac{1-\lambda}{(1+\lambda)}2}\right]$$

Gaussian, or Parametric, VaR

- We need to convert this to a lower bound for r_α
- To do this, we take the square root of the above bound and write $\widehat{r}_\alpha$ with its original value (typically negative, for small α) and not its absolute value.

$$\widehat{r}_\alpha \sqrt{1+2\sqrt{\frac{1-\lambda}{(1+\lambda)}2}}$$

Gaussian, or Parametric, VaR

- Now let us see some numbers.
- Suppose a smoothed estimate of the (daily) σ similar to the (unsmoothed) one obtained above from I.B.M. data between 1962 and 2005: 0.0164.
- You are computing a daily Gaussian VaR with $\alpha = 0.025$.
- In this case $z_{0.025} = -1.96$
- This gets a return distribution quantile point estimate equal to

$$\widehat{r_\alpha} = \hat{\sigma} z_\alpha = 0.0164 * (-1.96) = -0.0321$$

Gaussian, or Parametric, VaR

- let us assume that our variance estimate comes from a typical implementation of the smoothed estimate formula with daily data, $n = 256$ (meaning roughly one year of data) and $\lambda = 0.95$.
- In this case we have

$$\sqrt{1 + 2\sqrt{\frac{1-\lambda}{(1+\lambda)}2}} = 1.2053$$

- and the bound shall be, roughly, 20% more negative that the point estimate of the quantile, that is

$$\widehat{r}_\alpha \sqrt{1 + 2\sqrt{\frac{1-\lambda}{(1+\lambda)}2}} = -0.0321 * 1.2053 = -.0387$$

Gaussian, or Parametric, VaR

- Notice that

$$\sqrt{1+2\sqrt{\frac{1-\lambda}{(1+\lambda)}2}} = 1.2053$$

- only depends on the choice of λ, so that it can be pre computed for any estimate sharing the same choice of λ

Gaussian, or Parametric, VaR

- What we found are a point estimate and a confidence bound for the quantile of the log return.
- We understand that we used the log return due to the fact that the data used for estimating the standard deviation were log return data
- In order to transform these into a point estimate and confidence bound for the VaR we need to know the amount invested W at time t_0.
- The point estimate of VaR shall then be $W_{t0} * (e^{-.0321} - 1) = W_{t0} * (-.0316)$, the confidence bound shall be $W_{t0} * (e^{-.0387} - 1) = W_{t0} * (-.0380)$,
- That is: a loss of 3.16% with a bound of 3.8%
- (Here the use of % is correct, why?)

- In order to further understand the consequences of using the smoothed estimate consider the case of the "classic" estimate with $\lambda = 1$ in

$$V\left(\hat{\sigma}^2\right) = \frac{\left(\sum_{i=0,\ldots,n} \lambda^{2i}\right)}{\left(\sum_{i=0,\ldots,n} \lambda^{i}\right)^2} 2\sigma^4$$

- So that

$$V\left(\hat{\sigma}^2\right) = \frac{2}{n+1}\sigma^4$$

Gaussian, or Parametric, VaR

- The bound shall be

$$\widehat{r}_\alpha \sqrt{1+2\sqrt{\frac{2}{n+1}}}$$

- Due to n, we quickly have that the extreme of the interval becomes similar to the point estimate.
- For $n = 256$, and the same value for the estimate of σ as before, we get

$$\widehat{r}_\alpha \sqrt{1+2\frac{\sqrt{2}}{\sqrt{257}}} = \widehat{r}_\alpha * 1.084 = -0.0348$$

Gaussian, or Parametric, VaR

- As expected, if we do not discount data the sampling standard deviation of the estimate is smaller
- This means the bound shall be less negative...
- ...And we shall be less conservative in our risk quantification

Lecture: VaR, non Parametric

Summary

- VaR without a distributional hypothesis: non parametric VaR
- Estimation and confidence lower bound
- Pros and cons of non parametric VaR

Non Parametric, VaR

- Point estimate
- The non parametric VaR estimate stands, in some sense, at the opposite of the Gaussian VaR.
- In the non parametric case we suppose only that returns are i.i.d. but we avoid assuming anything about the underlying distribution.
- However, in order to find the VaR we need an estimate of the unknown theoretical distribution.

Non Parametric, VaR

- In standard parametric settings, where we assume, e.g., normality, this is done by estimating parameters
- The required probabilities and quantiles are then computed using the parametric model with estimated parameters.
- Since we now are making no specific parametric assumption about the return distribution we need to find an estimate of it which is "good" whatever the unknown distribution might be.

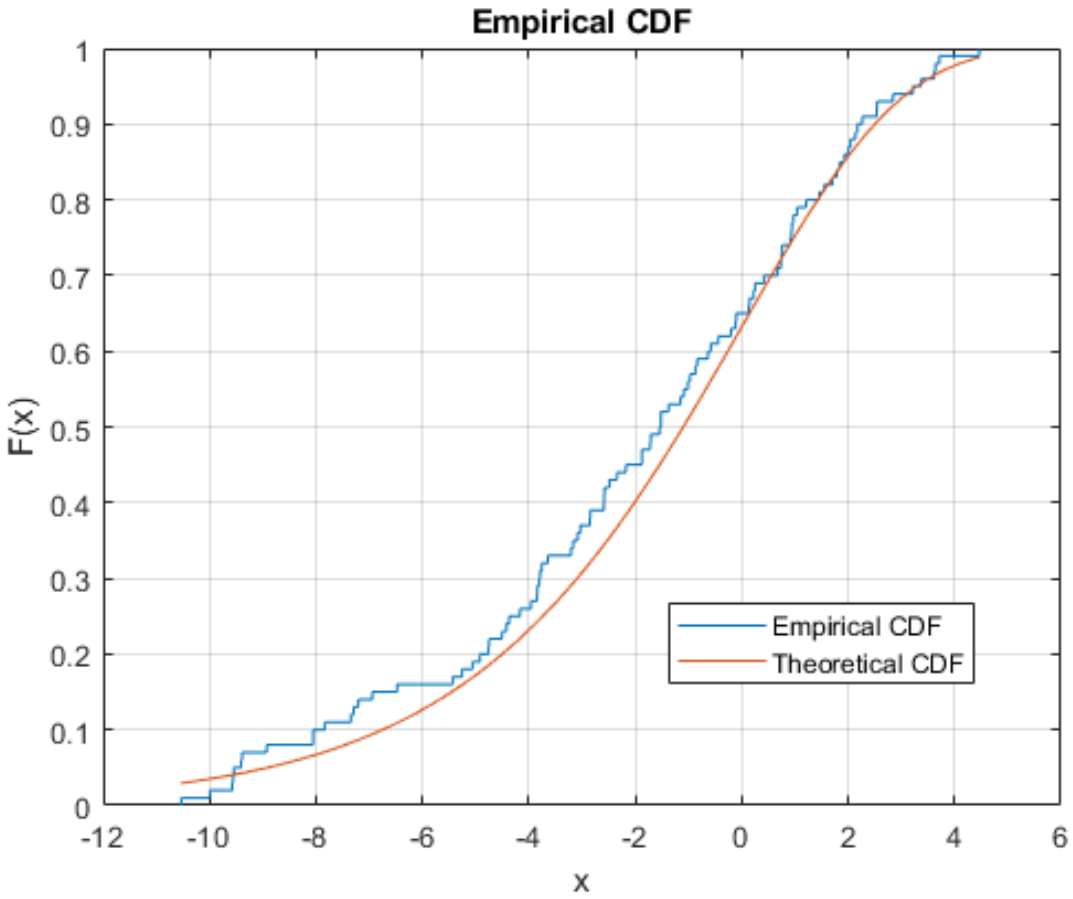
- The starting point of all non parametric procedures is to estimate the theoretical distribution using the empirical distribution function.
- Suppose we have a sample of n i.i.d. returns with common distribution $F(.)$ which yield observed values $\{r_1, r_2, ..., r_n\}$ then our estimate of $F(.)$. shall be:

$$\hat{P}(R \leq r) = \hat{F}_R(r) = \frac{\#r_i \leq r}{n} = \sum_{i=1}^{n} I(r_i \leq r)/n$$

- Where: $\#r_i \leq r$ means ' the number of observed returns less than or equal to r and $I(r_i \leq r)$ is a function which is equal to 1 if $r_i \leq r$ and 0 otherwise.

Non Parametric, VaR

- This is a plot of an empirical CDF compared witha theoretical CDF
- The vertical segnents in the empirical CDF should not be there (a problem in Excel)



- Under our hypothesis of i.i.d. returns with unknown distribution $F(.)$ the above defined estimate works quite well in the sense that

$$E(\sum_{i=1}^{n} I(r_i \leq r)/n) = \frac{n}{n} E(I(r_i \leq r) = P(r_i \leq r) = F(r)$$

- and

$$V(\sum_{i=1}^{n} I(r_i \leq r)/n) = \frac{n}{n^2} V(I(r_i \leq r)) = F(r)(1 - F(r))/n$$

- The last passage depends on the fact that, for given r, $I(r_i \leq r)$ is a Bernoulli random variable with $P = F(r)$.

Non Parametric, VaR

- Given this estimate of F, the non parametric quantile estimate is, in principle, very easy to compute.
- According to (one of) the standard definition(s) of empirical quantile we should:
- Order the observed r_i in an increasing way,
- Then find the smaller r_i such that the observed frequency of data less than or equal to it is at least α
- (Linear or other interpolations between consecutive observations are frequently used but we shall consider this in the "semi parametric VaR" section).
- This is nothing but the inversion of the empirical CDF, but this is not exactly what we shall do

Non Parametric, VaR

- Once we choose α, in the case of an empirical CDF we have two possibilities
- Either there is no value of R such that $\hat{F}_R(r) = \alpha$ (α is between two values taken by $\hat{F}_R$)
- Or there is an interval of values of R such that $\hat{F}_R(r) = \alpha$
- Since here we do risk management and should be cautious, our estimate $\hat{r}_\alpha$, in the first case, shall be the biggest observed value of R before the jump that takes $\hat{F}_R(r)$ from below α to above α
- In the second case our estimate shall be the leftmost value of R in the interval where $\hat{F}_R(r) = \alpha$
- (only in this second case we are using the above standard definition of empirical quantile)

Non Parametric, VaR

- The problem with the non parametric estimate is that, if α is small and n not VERY big, we are looking for $\hat{r}_\alpha$ in areas of the left tail where, probably, we made very few observations.
- In this case the estimate could be quite unstable and unreliable.
- The reader should compare this estimate with the estimate of a quantile in the Gaussian case.
- In the Gaussian case we estimate quantiles by inverting the CDF which, on its turn, is estimated, indirectly, by estimating μ and σ, the unknown parameters.
- This implies that any data point tells us something about any point of the distribution (maybe very far from the observed point) as it contributes to the estimate of both parameters.

Non Parametric, VaR

- In other terms, a parametric hypothesis allows us to estimate the shape of the distribution in regions where we do not make any observations.
- Instead, in the non parametric case, each data point, in some sense, has only a "local" story to tell.
- To be more precise: the non parametric estimate of the CDF at a given point r does not change if we change in any way the observed values in our dataset, provided we keep constant the number of observations smaller than and greater than r.

Non Parametric, VaR

- So, we use very little information from the data, in a non parametric estimate, while the influence of any data point on a parametric estimate is big.
- An unwritten law of Statistics is that, if you use little information, you are going to get an estimate which is robust to many possible hypotheses about the data distribution but with a high sampling variability;
- On the other hand, if you use a lot of information in your data, as you do in a parametric model, you are going to have an estimate which is not robust but with a smaller sampling variability.
- This is what happens in the case of non parametric VaR when compared to, say, Gaussian VaR.

Non Parametric, VaR

- Confidence interval for the non parametric quantile estimate
- Due to the possible instability of the estimate, it is very important to compute a one side confidence interval (bound) for the (true) α quantile $r_\alpha$on the basis of a simple random sample of size n.
- Suppose we order our data from the smallest to the biggest value (that is: we compute the "order statistics" of the sample).
- Call the $j-th$ order statistic $r_{(j)}$.

Non Parametric, VaR

- If we have n observations, and if there are no ties, the above defined estimate $\hat{r}_\alpha$ is nothing but $r_{[(n\alpha)]}$ where $[n\alpha]$ is the greates integer smaller that or equal to $n\alpha$ (the "integer part" of $n\alpha$)
- Clearly, this estimate could be on the left or on the right of the true r_α.
- If it is on the left we are overstating risk and this, if we do not exceed this is not bad
- If it is on the right we understating risk and this could be bad
- in order to reduce the probability of understating the risk we coul use a $r_{(j)}$ with $j < [n\alpha]$. Which one?

Non Parametric, VaR

- The general question is this: what is the probability that the "true" α quantile, say r_α is smaller than or equal to any given $j-th$ order statistic (ordered observation) $r_{(j)}$ in the sample?
- If this event happens and we use this quantile as estimate, the estimate shall be wrong, in the sense that we shall undervalue the possible loss (the "equal" part we put in as an added guarantee).
- This error is going to happen, if we use the *jth* ordered observation as estimate of r_α and, by chance, only $j-1$ are smaller then r_α
- (Note: as said above: the *jth* could be equal to r_α but we still count this as anerror)
- When the number of observations less than r_α is, say, $j-1$ the *jth* empirical quantile shall be bigger (less negative) than r_α (or equal to, see the previous sentence).

- Now some probability:
- We suppose iid observations so, if the underlying distribution of returns is continuous (with a continuous and increasing CDF), the probability of observing a return less than or equal to r_α is, by definition, α
- In this case, using the Binomial distribution, we find that the probability of making exactly i observations less than (or equal to) r_α (and so $n-i$ bigger than r_α) is

$$\binom{n}{i} \alpha^i (1-\alpha)^{n-i}$$

- We then have that the probability of making at most j observations less than (or equal to) r_α, that is, the probability that $r_{(j)}$ be greater than or equal to r_α (so that we make a mistake) is equal to the sum of the probabilities of observing exactly i returns smaller that or equal to r_α for $i = 0, 1, 2, ..., j$.

- For $i = 0$ all observations are greater than r_α; for $i = 1$ only the smallest observation is smaller that or equal to $r_\alpha$and so on up to $i = j$ where we have exactly j observations smaller than or equal to r_α
- (Again: we are including the case $r_{(j)} = r_\alpha$ because we want to be on the "safe side" and avoid a possible undervaluation of the risk).
- Obviously, from $i = j + 1$ onward, we have $j + 1$ or more observations smaller than or equal to r_α
- If we have $j + 1$ or more observations smaller than or equal to $r_\alpha$this implies that $r_{(j)}$ shall be (apart from ties) strictly smaller that r_α (supposing a probability of "ties": identical observations, equal to 0, as is the case with a continuous F)

- In the end, the probability of "making a mistake" in the sense of undervaluing the possible loss, that is the probability of choosing an empirical quantile $r_{(j)}$ greater than r_α, is given by:

$$P(r_{(j)} \geq r_\alpha) = \sum_{i=0}^{j} \binom{n}{i} \alpha^i (1-\alpha)^{n-i}$$

- Now the confidence limit: to be conservative, we want to estimate r_α with an empirical quantile $r_{(j)}$ such that we have a small probability β that that the true quantile r_α is smaller than its estimate.
- This, again, is because we are willing to overstate and not to understate risk, hence, we "prefer" to choose an estimate more negative than $r_\alpha$that a less negative one.
- Obviously, we would also like not to exaggerate on the safe side.

- Our strategy shall be as follows:
- Choose the maximum probability of under estimating risk (i.e. choosing a an estimate for the α quantile that is actually bigger that the true α quantile) you accept
- Call this probability β (the smaller it is, the more you are UNwilling to make mistakes)
- Then choose a $r_{(j)}$ such that $P(r_{(j)} \geq r_\alpha) \leq \beta$ for a given β
- Since you do not want to be overcautious you shall choose among all $r_{(j)}$ which satisfy the above condition, the biggest one (the one with the biggest (j))
- In short

$$\max(j) : P(r_{(j)} \geq r_\alpha) \leq \beta$$

- This is going to be the extreme of our one tail confidence interval.
- Notice this: the expected value of the random variable i with probability function $\binom{n}{i}\alpha^i(1-\alpha)^{n-i}$ is equal to $n\alpha$
- We could "expect" the empirical quantile corresponding to the index j just smaller than or equal to $n\alpha$, that is $[n\alpha]$ to be the "right choice"
- (and exactly this was our choice of point estimate made above)
- E.g. if $\alpha = .01$ and $n = 2000$, intuitively we could use as an estimate of r_α the empirical quantile $r_{(20)}$ since $n\alpha = 2000 * .01 = 20$

- However, if we make this choice, we are going (for n and α not too small) to have roughly fifty fifty probability that the true quantile is on the left or on the right of the estimate.
- This is due to the central limit theorem according to which

$$\sum_{i=0}^{j} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \approx \Phi_{n\alpha;n\alpha(1-\alpha)}(j)$$

- So that the empirical quantile r_j with $j = [n\alpha]$, for big n (almost) exactly 50% probability of being to the right of the true r_α(this probability is its β)

- If the approximation works for our n and α we see that $n\alpha$ becomes the mean of (almost) a Gaussian, hence the probability on the right and on the left of this becomes .5.
- For reasons of prudence fifty/fifty is not good for us, we go for a smaller probability that the chosen quantile be bigger than r_α, that is for a β smaller than .5.

- For this reason we choose an empirical quantile corresponding to a j whose value is smaller than $[n\alpha]$, the smaller j the smaller β.
- Since a direct use of the binomial CDF could be cumbersome, we again approximate it via the CLT
- If n is big and α not too small, we have:

$$P(r_{(j)} \geq r_\alpha) = \sum_{i=0}^{j} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \approx \Phi_{n\alpha;n\alpha(1-\alpha)}(j) =$$

$$= \Phi_{0;1}\left(\frac{j - n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right)$$

- With this approximation, we want to solve

$$\max(j) : \Phi_{0;1}\left(\frac{j-n\alpha}{\sqrt{n\alpha(1-\alpha)}}\right) \leq \beta$$

- So that our solution is given by the biggest (integer) j such that $\frac{j-n\alpha}{\sqrt{n\alpha(1-\alpha)}} \leq z_\beta$ or, that is the same, the biggest (integer) j such that $j \leq n\alpha + \sqrt{n\alpha(1-\alpha)}z_\beta$
- using the "integer part" notation we have: $j = [n\alpha + \sqrt{n\alpha(1-\alpha)}z_\beta]$
- (Note: for $\beta < .5$, as it shall be, we have $z_\beta < 0$ so j is going to be smaller that $n\alpha$, the more so the smaller is β)

Non Parametric, VaR

- Notice that the solution, in terms of the choice of j (obviously NOT in terms of $r_{(j)}$) does not depend on observed data.
- the index j of the empirical quantile to be used is known before sampling and only depends on α, β, n.
- Suppose, for instance, you have 1000 observations and look for the 2% VaR.
- The most obvious empirical estimate of the 2% quantile is the 20th ordered observation, but, according to the central limit theorem, the probability that the true 2% quantile is on the left of this (as on its right, but this is not important for us) is roughly 50%.

- To be conservative, you wish for a quantile which has only 2.5% probability of being on the right of the 2% quantile.
- Hence you choose a β of 2.5% ($z_\beta = -1,96$) and you get

$$n\alpha + z_\beta \sqrt{n\alpha(1-\alpha)} = 1000 * .02 - 1.96 * \sqrt{1000 * .02 * .98} = 11.32$$

- According to this result your choice for the lower (97.5%) confidence bound for the (2%) VaR is given by the 11th ordered observation (11=[11.32]) that is, roughly, the 1% empirical quantile (n was 1000).
- Beware: do not mistake α for β. The first defines the quantile you want to estimate (r_α) and the second the confidence level of the confidence interval.

Non Parametric, VaR

- Is this prudential estimate much different w.r.t. the simple "expected" quantile?
- It depends on the distance between ranked observations on the tail for observed cumulative frequencies of value about α.
- If the tail goes down quickly the distance is small and the difference between, in this case, the 11th and the 20th quantile shall not be big.
- On the contrary, with heavy tails the difference between the 1% and the 2% empirical quantile can be quite big.
- For instance: for a Gaussian the difference is of $.2726\sigma$ while for a T with 5 df it is $.4712\sigma$, almost double

Non Parametric, VaR

- As an example: consider the case of I.B.M. data between May 19th 1995b ($n = 2640$) and Sep 28th 2005.
- The point estimate of the 2.5% quantile is the 2.5% empirical quantile ($(j) = (66)$) which is -4.105%
- The point estimate of the 1% quantile is the 1% empirical quantile ($(j) = (26)$) which is -5.57%.
- To these point estimates correspond to $\beta = .025$ confidence bonds of -4.67% ($(j) = (50)$) and -6.53% ($(j) = (66)$).
- In the first case the bound is, roughly, .5% more negative than the point estimate, in the second case 1%.
- The reason for the difference is that around the 1% empirical quantile observations are more "rarefied", hence with large intervals in between, than around the 2.5% empirical quantile.

Non Parametric, VaR

- With the same data, a Gaussian VaR estimate and using, for comparison, an estimate of σ of 0.021421, we get, for the the 2.5% VaR, a point estimate of -4.14%, to be compared with the 2.5% empirical quantile -4.105% (confidence bound -4.67%).
- However, in the Gaussian case the (approximate 2σ) lower confidence limit, given the more than 2600 observation and the unsmoothed estimate, is -4.23%: very similar to the point estimate.
- As we did see a moment ago this is not true for the empirical VaR (.5% of difference between the estimate and the confidence limit).

Non Parametric, VaR

- Things are worse on more extreme quantiles.
- If we compute the 1% quantile in the Gaussian case, we get -4.93% with a (two σ) bound of -5.02% to be compared with the non parametric -5.57% and the corresponding bound of -6,53%.
- When we are evaluating extreme quantiles two "negative" forces sum.
- First, the empirical distribution is very "granular" in the tails (very few observations).
- Second, the empirically observed heavy tails, imply the possibility of considerable difference between contiguous quantiles, bigger that expected in the case of Gaussian data.

Non Parametric, VaR

- Non parametric VaR, sometimes dubbed "historical" VaR because it uses the observed history of returns in order to estimate the empirical CDF, is probably the most frequently used in practice.
- Again, confidence limits are often ignored and this could be due to their dismal "big" size.
- The problem of a big sampling variance for such estimates is very well known.

Non Parametric, VaR

- Applied VaR practitioners and academics have suggested, in the last years, an amazing quantity of possible strategies for improving the quality of the non parametric tail estimate.
- Most of these suggestion fall in two categories, semi parametric modeling of the distribution tails and filtered resampling algorithms.
- In the following subsection we shall consider a simple example of semi parametric model.
- The resampling approach is left for more advanced courses.
- The recent movement of Basel Agreements towards "tail conditional expectation" as a measure of risk, makes the semi parametric method of great interest for the future

Lecture: VaR, Semi Parametric

Summary

- A mixed model: semi parametric VaR
- Basic idea: a parametric model for the extreme left tail of the return distribution
- Polynomial model and its origin
- Estimate of the semi parametric VaR
- A quick hint to the Gaussian mixture model

Semi Parametric, VaR

- Semi parametric VaR mixes a non parametric estimate with a parametric model of the left tail.
- The aim is that of conjugating the robustness of the non parametric approach with the greater efficiency of a parametric approach.
- As we did see in the previous section, the non parametric approach can result in good evaluations of the VaR for values of α not very small.
- For small α its sampling variability may be non negligible even for big samples.
- However, in VaR computation we look for a quantile estimate for small α.

Semi Parametric, VaR

- The idea of a semi parametric approach is that of using a parametric model just for the tail of the distribution beyond a small but not to small α quantile.
- The plug in point of the parametric model is estimated in a non parametric way, from that point onward a parametric model is used in order to estimate the required extreme quantile.
- The reason why this may work is that, on the basis of arguments akin to the central limit, but considering not means but extreme order statistics, we can prove that, while we may have many different parametric models for probability distributions, the tails of such distributions behave in a way that can be approximated with few (typically three) different parametric models.

Semi Parametric, VaR

- Here is not the place to introduce the very interesting topic of "extreme value theory", a recent fad in the quantitative Finance milieu
- For this reason we shall not be able to fully justify our choice of tail parametric model.
- Be it sufficient to say that a rigorous justification is possible.

- We suppose that for r negative enough:

$$P(R \leq r) = L(r)|r|^{-a}$$

- where, for such negative enough r, $L(.)$ is a slowly varying function for $r \to -\infty$
- (Formally this means $\lim_{r \to -\infty} \frac{L(\lambda r)}{L(r)} = 1 \, \forall \lambda > 0$ and you can understand this as implying that the function L to be approximately a constant for big negative values of r)
- a is the speed with which the tail of the CDF goes to zero with a polynomial rate.

- This is sometimes called a "Pareto" tail because a famous density showing this tail behaviour bears the name of Vilfredo Pareto.
- This choice of tail behaviour could be justified on the basis of limit theory, or on the basis of good empirical fitting to data (e.g. the T distribution has such a behaviour).
- Notice that the Gaussian CDF has exponential tails, which go to 0 much faster than polynomial tails.
- Pareto tails are, thus, a model for "heavy" tails.

- Provided we know where to plug in the model (that is: which value of r is negative enough) our first task is that of estimating a, the only parameter in the model.
- In order to do so we take the logarithm of $P(R \leq r) = L(r)|r|^{-a}$ and we get:

$$\log(P(R \leq r)) = \log(L(r)) - a\log(|r|)$$

- We then assume that, maybe with an error, $\log(L(r))$ can be approximated by a constant C:

$$\log(P(R \leq r)) \approx C - a\log(|r|)$$

- This expression begins to be similar to a linear model.
- In fact, if, in correspondence of any observed r_i we may estimate $\log(P(R \leq r_i))$ with $\log(\hat{F}_R(r_i))$ and summarize the various approximations in an error term u_i, we have:

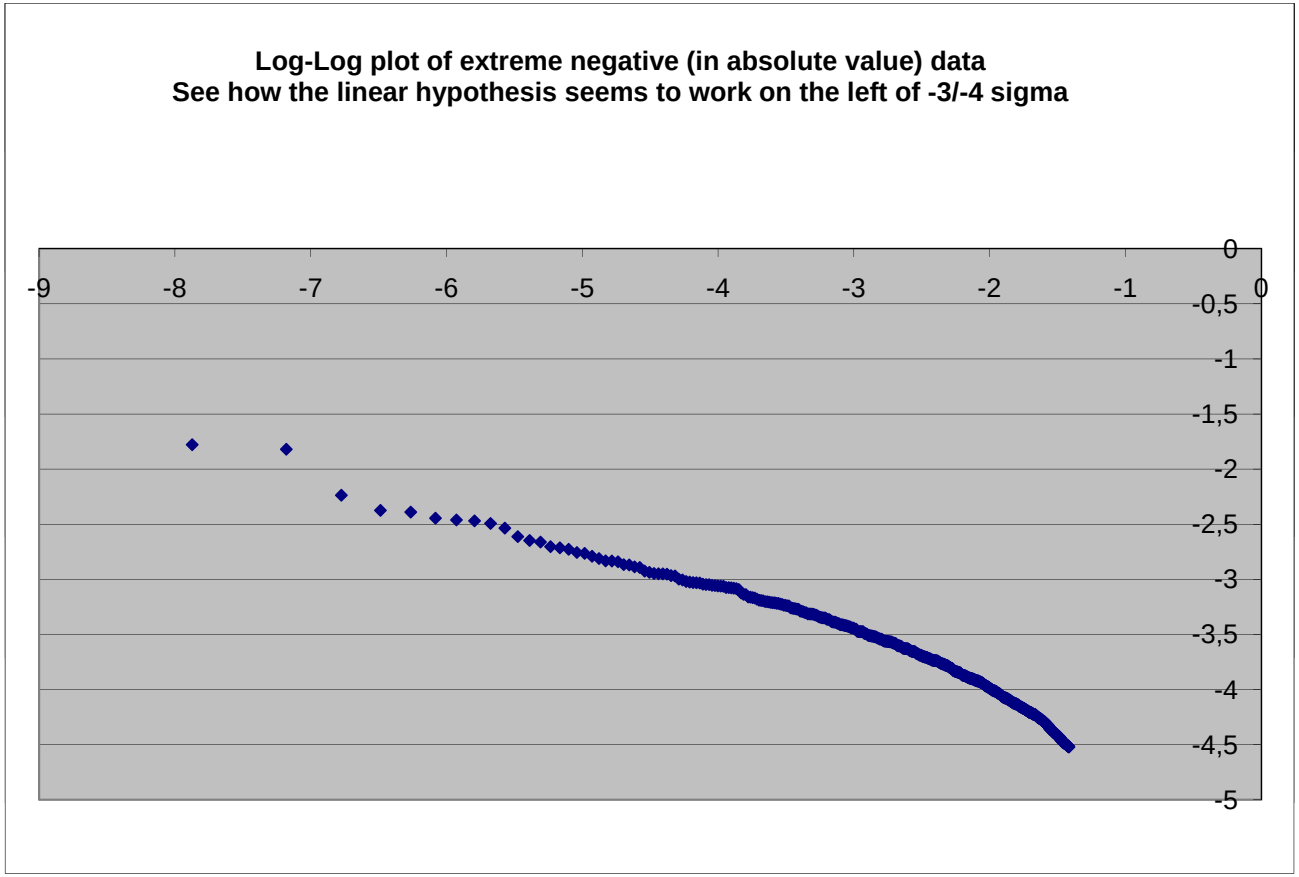
$$\log(\hat{F}_R(r_i)) = C - a\log(|r_i|) + u_i$$

- A linear regression based on this model shall not work for the full dataset of returns (why?), but it shall work for a properly chosen subset of extreme negative returns.
- A simple way to find the proper subset of tail observations to which it is reasonable to apply this simple model, is that of plotting $\log(\hat{F}_R(r_i))$ against $\log(|r_i|)$ for the left tail of the distribution.

Semi Parametric, VaR

- Typically, this plot shall show a parabolic region (consistent with the Gaussian hypothesis) or in any case a non linear region, followed by a linear region (consistent with the polynomial hypothesis).
- The regression shall be run with data from the second region.

Semi Parametric, VaR



Semi Parametric, VaR

- Suppose we now have an estimate for a out of our regression, how do we get an estimate of the quantile?
- The problem is that of plugging in the parametric tail to a different estimate of the CDF, in most cases this is the non parametric estimate.
- The solution is simple if we suppose to have a good non parametric (or of other kind) estimate for the quantile r_{α_1} where α_1 is too big for this quantile estimate to be used as VaR but is big enough so that the estimate of r_{α_1} is good.
- What we need is, now,an estimate of r_{α_2} for $\alpha_2 < \alpha_1$.

- If we suppose that the polynomial tail model is approximately true for both quantiles we have:

$$\frac{\alpha_1}{\alpha_2} = \frac{L(r_{\alpha_1})}{L(r_{\alpha_2})} \left(\frac{r_{\alpha_2}}{r_{\alpha_1}} \right)^a$$

- But the ratio $\frac{L(r_{\alpha_1})}{L(r_{\alpha_2})}$ should be very near to 1 (the same slow varying function computed at non very far away points) so that we can directly solve for r_{α_2}:

$$r_{\alpha_2} = r_{\alpha_1} \left(\frac{\alpha_1}{\alpha_2} \right)^{\frac{1}{a}}$$

Semi Parametric, VaR

- This formula solves our problem.
- Given the estimated r_{α_1}, an estimate of a (based on the above described regression), our choice of α_2 we are then able to estimate the quantile r_{α_2}.
- In the Handouts you shall find the estimate for one sided confidence bound corresponding to the semi parametric estimate (not for the exam)

Semi Parametric, VaR

- Show the Excel file on VaR computation

Lecture: Intro to Matrix Algebra

Summary

- The bare minimum of what should be known about matrix algebra in the risk management and asset management business

Basic definitions: matrix

- A matrix is a doubly indexed array of elements a_{ij} and is usually indicated with an uppercase letter A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix}$$

- i is the row index and j the column index
- The elements a_{ij} may be many different things, in our applications these shall be real numbers or real valued (maybe random) variables
- Sometimes a matrix is indicated as $A = \{a_{ij}\}$
- Sometimes, when useful, dimensions are explicitly given as: $A(n \times k)$ or as: $A_{n \times k}$

Basic definitions: vector

- A single row or single column matrix is called a "vector"
- Usually, a vector is indicated with a lower case letter z and its elements shall be indicated with a single index z_i
- In order to avoid confusion, in this course, if nothing else is said, vectors shall be introduced as columns

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

Basic definitions: transpose

- The transpose of a matrix A is indicated with A'
- A' contains all the elements of A but with flipped indexes: $A' = \left\{ a'_{ij} \right\} = \{ a_{ji} \}$
- The transpose of a transpose is the original matrix $(A')' = A$

Basic definitions: sum

- Given a matrix A and a matrix B the sum $C = A + B$ is defined iff A and B have the same sizes, that is: the same number of rows and columns
- In this case $C = \{c_{ij}\} = \{a_{ij} + b_{ij}\}$
- The sum is commutative: $A + B = B + A$
- The sum is associative, e.g.: $A_1 + A_2 + A_3 = (A_1 + A_2) + A_3$ (note the index which here distinguishes among matrices)
- The transpose of a sum is the sum of the transposes: $(A + B)' = A' + B'$

Basic definitions: sum

- NOTE: when we write a sum or other operations which require conditions on the size of the involved matrices we implicitly assume these conditions are satisfied.
- This is always the case if you follow the correct algebra.
- It could be a good strategy, when beginners in these things, to check the involved dimensions as a way to verify if you are following the right algebra

Basic definitions: product

- Given a matrix A and a matrix B the product $C = AB$ is defined iff the number of columns of A and the number of rows of B coincide
- In this case $C = \{c_{ij}\} = \left\{\sum_{l=1,\ldots,k} a_{il} b_{lj}\right\}$
- The product, in general is NOT commutative: $AB \neq BA$ and one of the two may not exist even if the other exists
- The product is associative, e.g.: $A_1 A_2 A_3 = (A_1 A_2) A_3$
- Transpose of a product: $(AB)' = B'A'$
- The product and the sum interact according to a distributive law: $(A+B)C = AC + BC$ and $D(E+F) = DE + DF$
- Notice that the C which is on the right stays on the right while the D which is on the left stays on the left

Basic definitions: interaction with scalars

- A scalar: c, is a matrix. However, in doing matrix algebra, sometimes we use scalars with a certain freedom
- Sometimes we write $c + A$ and this has no meaning as a matrix sum except in the case where A is a scalar, however $c + A$ is usually intended as summing c to each a_{ij}
- In a similar way cA has a meaning, as matrix product, only if A has a single row, however cA is usually intended as the result of multiplying each a_{ij} times c
- It is clear that, with these definitions, $c + A = A + c$ and $cA = Ac$

Basic definitions: rank of a matrix

- A row vector x is said to be linearly dependent from the row vectors of a matrix A if it is possible to find a row vector z such that $x = zA$. The same for a column vector
- $r(A)$ (or $rank(A)$): the "rank" or a matrix A is defined as the number of linearly independent rows or (the number is the same) the number of linearly independent columns of A
- A square matrix of size n is called non singular if $r(A) = n$
- If B is any $n \times k$ matrix, then $r(AB) \leq \min(r(A), r(B))$
- If B is an $n \times k$ matrix of rank n, then $r(AB) = r(A)$
- If C is an $l \times n$ matrix of rank n, then $r(CA) = r(A)$

Basic definitions: special matrices

- A matrix with the same number of rows and columns is called "square"
- If a square matrix A is such that $A' = A$ the matrix is said to be "symmetric"
- A matrix such that $A = AA$ is termed "idempotent"
- We shall find a number of important idempotent matrices. Notice that, among real scalars, only 0 and 1 satisfy $A = AA$

Basic definitions: special matrices

- A square matrix A such that $a_{ij} = 0 \; \forall i \neq j$ is called "diagonal"
- A diagonal matrix which has all ones on the diagonal ($a_{ii} = 1 \forall i$) is called "identity" and indicated with I.
- The name comes from the fact that $IA = AI = A$ but notice that, in order for both product to have a meaning, the two I shall have different sizes except in the case if a square A.
- Sometimes, to avoid errors, the symbol of the identity matrix is accompanied by its size, e.g. I_n

Basic definitions: inverse

- The matrix sum and the matrix difference are, obviously, the same operation as the sum is valid both for positive and negative elements
- We defined the matrix product. Does there exist a "matrix division"?
- You remember that, with scalars, we can take the ratio of any two numbers provided the denominator is not zero. With matrices things are a little more complicated

Basic definitions: inverse

- let us start with $AB = C$ and $DA = F$. Suppose now we have C and F. Does there exist a matrix, say A^{-1} such that $A^{-1}C = B$ whatever is B and $FA^{-1} = D$ whatever is D?
- If this matrix exists it is called the "inverse" of A
- While things may be considered in a more general setting we shall concentrate on square matrices
- Clearly, if A^{-1} exists it must be that $A^{-1}A = I = AA^{-1}$, otherwise there would exist some B or D such that $A^{-1}AB \neq B$ and/or $DAA^{-1} \neq D$

Basic definitions: inverse

- The inverse of a square matrix A exists if and only if the matrix is non singular, that is: if the size and the rank of A are the same (if you remember the notion of "determinant" of a square matrix, you may simply ask for the determinant of A not to be null)
- If the products and inversions in the following formula are well defined (that is dimensions agree and the inverses exists), then $(AB)^{-1} = B^{-1}A^{-1}$ moreover $(A')^{-1} = (A^{-1})'$
- There is no simple law for the inverse of a sum (one exists but is not simple)

Basic definitions: inverse

- As it should be clear from the previous slides, inversion has to do with the solution of linear non homogeneous systems
- Problem: find a column vector x such that $Ax = b$ with A and b given
- If A is square and invertible then the unique solution is $x = A^{-1}b$
- Important note: if A is $n \times k$ with $n > k$ but $r(A) = k$ then the system $Ax = b$ has in general no exact solution, however the system $A'Ax = A'b$ has the solution $x = (A'A)^{-1}A'b$

Quadratic forms

- A quadratic form with coefficient matrix given by the symmetric matrix A and variables vector given by the column vector x (with size of A equal to the number of rows of x) is the scalar given by:

$$x'Ax = \sum_i \sum_j a_{ij} x_i x_j$$

.

- A symmetric matrix A is called semi positive definite if and only if
- $x'Ax \geq 0$ for all x
- It is called positive definite if and only if
- $x'Ax > 0$ for all non null x

Quadratic forms

- If a matrix A can be written as $A = C'C$ for any matrix C then A is surely at least PSD
- In fact $x'Ax = x'C'Cx$ but this is the product of the row vector $x'C'$ times itself, hence a sum of squares and this cannot be negative
- It is also possible to show that any PSD matrix can be written as $C'C$ for some C

Derivatives of functions written in matrix terms

- Often, we must compute derivatives of functions of the kind $x'Ax$ (a quadratic form) or $x'q$ (a linear combination of elements in the vector q with weights x) with respect to the vector x
- In both cases we are considering a (column) vector of derivatives of a scalar function w.r.t. a (column) vector of variables (commonly called a "gradient")
- There is a useful matrix notation for such derivatives which, in these two cases, is simply given by:

$$\frac{\partial x'Ax}{\partial x} = 2Ax$$

and

$$\frac{\partial x'q}{\partial x} = q$$

Derivatives of functions written in matrix terms

- The proof of these two formulas is quite simple (and you do not need to remind it, but do it just once).
- In both cases we give a proof for a generic element k of the derivative column vector
- For the linear combination we have

$$x'q = \sum_j x_j q_j$$

$$\frac{\partial x'q}{\partial x_k} = q_k$$

Derivatives of functions written in matrix terms

- For the quadratic form

$$\frac{\partial x'Ax}{\partial x'} = 2x'A$$

$$x'Ax = \sum_i \sum_j x_i x_j a_{i,j}$$

$$\frac{\partial \sum_i \sum_j x_i x_j a_{i,j}}{\partial x_k} = \sum_{j \neq k} x_j a_{k,j} + \sum_{i \neq k} x_i a_{i,k} + 2x_k a_{k,k} =$$

$$= \sum_{j \neq k} x_j a_{k,j} + \sum_{j \neq k} x_j a_{k,j} + 2x_k a_{k,k} = 2A_{k,.}x$$

- Where $A_{k,.}$ means the $k-th$ row of A and we used the fact that A is a symmetric matrix

Derivatives of functions written in matrix terms

- An important point to stress is that the derivative of a function with respect to a vector always has the same dimension as the vector w.r.t. which the derivative is taken, in this case x, so, for instance

$$\frac{\partial x'Ax}{\partial x} = 2Ax$$

- and not

$$\frac{\partial x'Ax}{\partial x} = 2x'A$$

- (remember that A is symmetric)

Lecture: Matrix Algebra and Statistics

Summary

- We must be able to work, jointly, with: Matrix Algebra + Probability + Statistics
- This is what you should know (minimum)

Random vectors and matrices

- A random vector, resp matrix, is simply a vector (matrix) whose elements are random variables

Functions of random vectors (or matrices)

- A function of a random vector (matrix) is simply a matrix (or scalar) function of the components of the random vector (matrix)
- Simple examples are: the sum of the elements of the vector and the determinant of a random matrix (both are scalars), sums or products of matrices and vectors (these could be scalar, vector or full matrices) and so on
- We shall be interested in functions of the vector (matrix) X of the kind: $Y = A + BXC$ where A, B and C are non stochastic matrices of dimensions such that the sum and the products in the formula are well defined
- A quadratic form $x'Ax$ with a non stochastic coefficient matrix A and stochastic vector x is and example of non linear, scalar function of a random vector

Expected values of random vectors (or matrices)

- These are simply the vectors (matrices) containing the expected values of each element in the random vector (matrix)

$$E(X) = \begin{bmatrix} E(x_{11}) & E(x_{12}) & \cdots & E(x_{1k}) \\ E(x_{21}) & E(x_{22}) & \cdots & E(x_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ E(x_{n1}) & E(x_{n2}) & \cdots & E(x_{nk}) \end{bmatrix}$$

- The expected value of the transpose is the transpose of the matrix of expected values: $E(X') = E(X)'$
- An important result which generalizes the linear property of the scalar version of the operator $E(.)$ for the general linear function defined above (recall: only X stochastic), is this

$$E(A + BXC) = A + BE(X)C$$

Variance covariance matrix

- For random column vectors, and here we mean vectors only, we define the variance covariance matrix of a column vector X as: $V(X) = V(X') = E(XX') - E(X)E(X') = E((X - E(X))(X - E(X))')$
- The Varcov (or simply "variance") matrix is symmetric, on the diagonal we have the variances ($V(X_i) = \sigma^2_{X_i}$) of each element of the vector while in the upper and lower triangles we have the covariances ($Cov(X_i; X_j)$)
- The most relevant property of this operator is:

$$V(A + BX) = BV(X)B'$$

- From this property we deduce that Varcov matrices are always (semi) positive definite. In fact if $A = V(z)$ and x is a (non random) column vector of the same size as z, then $V(x'z) = x'Ax$ which cannot be negative for any possible x

Variance covariance matrix

- It is sometimes useful to define the covariance matrix of two different (column) vectors, say: z_1 and z_2
- This is defined as $Cov(z_1, z_2) = E(z_1 z_2') - E(z_1)E(z_2')$
- Notice that this is, in general, neither square nor symmetrical (it contains covariances between different vectors, and the vectors may have different sizes) and reduces to the standard definition of variance covariance matrix when $z_1 = z_2$

Correlation coefficient

- The correlation coefficient between two random variables is defined as:
$$\rho_{X_i;X_j} = \frac{Cov(X_i;X_j)}{\sigma_{X_i}\sigma_{X_j}}$$
- The correlation matrix $\rho(x)$ of the vector x of random variables is simply the matrix of correlation coefficients or, that is the same, the Varcov matrix of the vector of standardized X_i
- The presence of a zero correlation between two random variables is defined, sometimes, linear independence or orthogonality.
- The reader should be careful using these terms as they exist also in the setting of linear algebra but their meaning, even if connected, is slightly different. Stochastic independence implies zero correlation, the reverse proposition is not true

Minimization of a PD quadratic form, approximate solution of over determined linear system

- Now Let us go back to the linear system $Ax = b$ with A an $n \times k$ matrix of rank k. If $n > k$ this system has, in general, no solution. However, let's try to solve a similar problem. By solving a system we wish for $Ax - b = 0$ in our case this is not possible so let us try and change the problem to this $\min_x (Ax - b)'(Ax - b)$. In words: try to minimize the sum of squared differences between Ax and b if you cannot make it equal to 0
- We have

$$(Ax - b)'(Ax - b) = x'A'Ax + b'b - 2b'Ax$$

- Now let us take the derivative of this w.r.t. x

$$\frac{\partial}{\partial x} x'AA'x + b'b - 2b'Ax = 2A'Ax - 2A'b$$

- (remember the rule about the size of a derivatives vector)

Minimization of a PD quadratic form, approximate solution of over determined linear system

- We now create a new linear system equating these derivatives to 0

$$A'Ax = A'b$$

- And the solution is

$$x = (A'A)^{-1}A'b$$

- This is the "least squares" approximate solution of a (over determined) linear system. (see the Appendix on least squares and Gauss Markov model)

Minimization of a PD quadratic form under constraints. Simple applications to Finance

- Suppose we are given a column vector r where r_j is the random (linear) return for the stock j
- Suppose we are holding these returns in a portfolio for one time period and that the (known) relative amount of each stock in our portfolio is given by the column vector w such that $1'w = 1$ where 1 indicates a column vector of ones of the same size as w
- Then the random linear return of the portfolio over the same time period is given by $r_\pi = w'r$
- Since w is known we have $E(w'r) = w'E(r)$ and $V(w'r) = w'V(r)w$

Minimization of a PD quadratic form under constraints. Simple applications to Finance

- The fact that, over one period of time, the expected linear return and the variance of the linear return of a portfolio only depend on the expected values and the covariance matrix of the single returns and the weight vector, is what allows us to implement a simple optimization method
- For the moment let us suppose that the problem is

$$\min_{w:1'w=1} w'V(r)w$$

- Where, again, the constraint $1'w = 1$ is necessary otherwise we would not be considering portfolios

Minimization of a PD quadratic form under constraints. Simple applications to Finance

- In this problem we want to minimize a quadratic form under a linear constraint
- It is to be noticed that, without the constraint, the problem would be solved by $w = 0$ (no investment). The constraint does not allow for this
- Such problems can be solved with the Lagrange multiplier method
- The idea is to artificially express, in a single function, both the need of minimizing the original function and the need to do this with respect to the constraint $1'w = 1$

Minimization of a PD quadratic form under constraints. Simple applications to Finance

- In order to do this we define the Lagrangian of the problem given by

$$L(w, \lambda) = w' V(r) w - 2\lambda(1'w - 1)$$

  In this function the value of the unconstrained objective function is summed with the value of the constraint multiplied by a dummy parameter 2λ

- We now take the derivatives of the Lagrangian w.r.t. w and λ

$$\frac{\partial}{\partial w}(w' V(r) w - 2\lambda(1'w - 1)) = 2V(r)w - 2\lambda 1$$
$$\frac{\partial}{\partial \lambda}(w' V(r) w - 2\lambda(1'w - 1)) = 2(1'w - 1)$$

- If we set both these to zero we get, supposing $V(r)$ invertible

$$V(r)w = \lambda 1$$
$$1'w = 1$$

- Notice the difference between the 1-s. In the first equation 1 is a column vector which is required because we cannot equate a vector to a scalar. The same for 1' in the second equation which while the r.h.s. is a scalar one (for dimension compatibility with the l.h.s.) We do not stress this using, e.g., boldface for the vector 1 because the meaning follows unambiguously from the context
- It is clear that the second equation is satisfied if and only if w satisfies the constraint

Minimization of a PD quadratic form under constraints. Simple applications to Finance

- What is the meaning of the first equation (or, better, set of equations)? The unconstrained equation would have been

$$V(r)w = 0$$

whose only solution (due to the fact that $V(r)$is invertible) would be $w = 0$. But this solution does not satisfy the constraint

- What we shall be able to get is $V(r)w = \lambda 1$. For some λ chosen in such a way that the constraint is satisfied
- To find this λ, simply put together the result of the first set of equations: $w = \lambda V(r)^{-1}1$ and the equation expressing the constraint: $1'w = 1$. Both equations are satisfied if and only if

$$\lambda = 1/1'V(r)^{-1}1$$

Minimization of a PD quadratic form under constraints. Simple applications to Finance

- We now know λ, that is we know of exactly how much we must violate the unconstrained optimization condition ($V(r)w = 0$) in order to satisfy the constraint
- In the end, putting this value of λin the solution for the first set of equations, we get

$$w = \frac{V(r)^{-1}1}{1'V(r)^{-1}1}$$

- It is to be noticed that these are only necessary conditions but, for our purposes, this is enough
- What we got is the one period "minimum variance portfolio" made of securities whose returns covariance is $V(r)$. Notic ethat weights do not epend on expected values: we only minimize variance
- Suppose $V(r)$ diagonal, then w is the column vector containing the ratio between the reciprocal of each variance

- What is the variance of this portfolio?

$$V(w'r) = w'V(r)w = \frac{1'V(r)^{-1}}{1'V(r)^{-1}1} V(r) \frac{V(r)^{-1}1}{1'V(r)^{-1}1} =$$

$$= \frac{1'V(r)^{-1}1}{(1'V(r)^{-1}1)^2} = \frac{1}{1'V(r)^{-1}1}$$

  In the diagonal case: the precision of the optimal portfolio is the sum of the precisions of the components

- The expected value shall be

$$E(w'r) = w'E(r) = \frac{1'V(r)^{-1}E(r)}{1'V(r)^{-1}1}$$

Two last results/exercises useful for the rest of the courses

- Note: In what follows we shall consider, very briefly, a basic asset allocation result and a fundamental theorem which simplifies the use of varcov matrices
- We shall also give two basic results concerning the regression function
- These are "fragments" of results we do not have the time to consider in depth. And we should!
- From this point to the end of this lecture you shall find many slides with an astrerisk; *, in the title
- This means that, while the content of these slides is relevant for understanding the financial meaning of what we are going to do, those slides are not required for the exam
- For your general knowledge, however, I strongly suggest you to give some of your study time to these slides

deFinetti, Markowitz and Roy Asset Allocation Model: Summary

- The basic mean variance asset allocation model as an exercise in matrix algebra and probability
- The spectral theorem: the basic way for understanding a variance covariance matrix
- A recall to the definition and properties of the regression function (we are about to deal with the linear model)

The deFinetti, Markowitz and Roy model for asset allocation

- This section on the deFinetti-Markowitz-Roy model is here both as an exercise in matrix calculus and because it shall be useful to us in what follows.
- Suppose one investor is considering a buy and hold portfolio strategy for a single period of any fixed length (the length is not a decision variable in the asset allocation) from time t to time T.
- Let us indicate with R the random vector of linear total returns from a given set of (k) stocks for the given time period.
- We suppose the investor knows (not "estimates") both $\mu_R = E(R)$ and $\Sigma = V(R)$. We suppose Σ to be invertible.
- Moreover, there exists a security which can be bought at t and whose price at T is known at t (typically a non defaultable bond) called "risk free security".
- Let r_f be the (linear) non random return from this investment over the period.

- The investor's strategy is to invest in the risk free security and in the stocks at time t and then liquidate the investment at time T.
- The relative amounts of each investment in stocks are in the column vector w, while $1 - w'1$ is the relative amount invested in the risk free security.

- The return of the portfolio between t and T is:

$$R_\Pi = (1 - w'1)r_f + w'R$$

- So that the expected value and the variance of the portfolio return are

$$E(R_\Pi) = (1 - w'1)r_f + w'\mu_R$$

- and

$$V(R_\Pi) = w'\Sigma w$$

- The problem faced by the investor is to choose w so that, for a given expected value c of the portfolio return, the variance of the portfolio return is minimized.
- In formulas

$$\min_{w:(1-w'1)r_f+w'\mu_R=c} w'\Sigma w$$

- Equivalently the investor could choose the variance and find w such that the expected return is maximized.

- In both problems it would be sensible to use an inequality constraint.
- For instance, in the first problem, we could look for

$$\min_{w:(1-w'1)r_f+w'\mu_R \geqq c} w'\Sigma w$$

- We choose the $=$ version just for allowing direct use of Lagrange multipliers as we'll see in what follows.

- Notice that we do not assume the sum of the elements of w (the riski part of the investment) to be 1.
- This shall be true only if no risk free investment is made.
- However, obviously, if we complement the vector w with the fraction of portfolio invested in the risk free security, the sum of all the portfolio fractions is 1.
- Moreover we do not require each element of w to be positive.
- This can be done but not in the the straightforward way we are going to follow

- In order to solve the problem we consider its Lagrangian function (notice the rearranged constraint):

$$L = w'\Sigma w - 2\lambda(r_f - c + w'(\mu_R - 1r_f))$$

- And differentiate this with respect to the vector w and the scalar λ (remember the differentiation rules).

$$\frac{\partial L}{\partial w} = 2\Sigma w - 2\lambda(\mu_R - 1r_f)$$

$$\frac{\partial L}{\partial \lambda} = -2(r_f - c + w'(\mu_R - 1r_f))$$

- We then define the system of "normal equations" as

$$\begin{cases} \Sigma w - \lambda(\mu_R - 1 r_f) = 0 \\ r_f - c + w'(\mu_R - 1 r_f) = 0 \end{cases}$$

- It is easy to solve the first sub system for w as:

$$w = \lambda \Sigma^{-1}(\mu_R - 1 r_f)$$

- At this point we already notice that the required solution is a scalar multiple (λ) of a vector which does not depend on c.
- In other words the relative weights of the stocks in the risky portfolio are already known and do not depend on c.
- More on this in a moment.
- What is still not known is the relative weight of the portfolio of stocks with respect to the risk free security.

- We get this by substituting the result for w in the last equation of the system (notice that in the last equation we have w')

$$r_f - c + \lambda(\mu_R - 1r_f)'\Sigma^{-1}(\mu_R - 1r_f) = 0$$

- So that

$$\lambda = \frac{c - r_f}{(\mu_R - 1r_f)'\Sigma^{-1}(\mu_R - 1r_f)}$$

- This formula tells us a very important fact: the relative weights in the optimal risky portfolio are unique and depend only on Σ, r_f and μ_R, not on c
- The amount of expected return we want to achieve (c) only influences the allocation between the risk free security and the stock portfolio
- (Notice that the denominator of λ is a constant incependent on c and, then, the bigger is c the bigger (in absolute value) shall be the risky part of the portfolio of an amount which is bigger if the denominator of λ is smaller)
- This is a first instance of a "separation theorem": the decision of investment is split in two
- First step: find the optimal risky portfolio. Second step: then decide how much to invest in it

The deFinetti, Markowitz and Roy model for asset allocation

- This may seem "just" mathematics. It is not so: there are important and very direct practical implications
- A relevant example: you may have heard that over a given horizon, you should invest more in bonds and less in stocks if you are old and vicevesa if you are young.
- Our result tells us this is WRONG! The risky portfolio is always the same, for given Σ, r_f and μ_R, (and time horizon) for this reason a possible suggestion could be to invest more or less in it but NOT to change its composition
- In order for such suggestion to have a meaning (possibly), you should suppose horizons to be different (not unreasonable) AND asset statistics not to behave as in the log random walk case (otherwise the optimal weights would be the same over any horizon and the only change would be in the optimal Sharpe ratio, multiplied by the square root of the number of periods in the horizon)

- In the handouts, this section goes on with details as: more on the value of λ, its connection with the Sharpe ratio, the statistical properties of the optimal portfolio and so on.
- These are interesting but not required for the exam
- We limit ourselves to two comments (both not for the exam but important)

The deFinetti, Markowitz and Roy model for asset allocation*

- First: the expected excess return of R_π is, obviously, $E(R_\Pi) - \mu_R = c - \mu_R$
- The variance of $R_\Pi - r_f$ shall be

$$V(R_\Pi - r_f) = \frac{(E(R_\Pi) - r_f)^2}{(\mu_R - 1r_f)'\Sigma^{-1}(\mu_R - 1r_f)}$$

- So that its standard deviation shall be (suppose $R_\Pi > r_f$)

$$\sqrt{V(R_\Pi - r_f)} = \frac{E(R_\Pi) - r_f}{\sqrt{(\mu_R - 1r_f)'\Sigma^{-1}(\mu_R - 1r_f)}}$$

- Rearranging, we recognize in this the "Sharpe ratio" for the portfolio return

$$\sqrt{(\mu_R - 1r_f)'\Sigma^{-1}(\mu_R - 1r_f)} = \frac{E(R_\Pi) - r_f}{\sqrt{V(R_\Pi - r_f)}}$$

The deFinetti, Markowitz and Roy model for asset allocation*

- Notice that this ratio does not depend on $c = E(R_\Pi)$ so that in the mean standard deviation plane all (optimal) portfolios can be found on the same straight line whose slope is the Sharpe ratio (in other words: it is meaningless to compare Sharpe ratios either a single one exists or some portfolio is suboptimal)
- Moreover, since we now see that the denominator of λ is the (optimal) Sharpe ratio we also understand that the bigger c the bigger is going to be the (absolute) amount invested in risky securities at a ratio which is greater if the Sharpe ratio is smaller
- In markets where the optimal portfolio(s) have a small Sharpe ratio you must increase a lot the (absolute) exposure to risky securities in order to increase the expected return
- This should be almost obvious, less obvious the fact that the proportionality factor is NOT the (optimal) Sharpe ratio but its square

The deFinetti, Markowitz and Roy model for asset allocation*

- Second: we see that, had our objective been that of solving

$$\max_{w} \left\{ (1 - w'1) r_f + w' \mu_R - \frac{1}{2\lambda} w' \Sigma w \right\}$$

- That is: had we wished to maximize some "mean variance" utility criterion, our result would have been exactly the same.
- Since the (negative) weight of the variance in this criterion is given by $-\frac{1}{2\lambda}$ as a rule $1/\lambda$ is termed "risk aversion parameter".
- Notice that, according to our formula for λ, the bigger is c, the bigger is λ (the smaller the risk aversion) and the bigger is the fraction of wealth invested in the optimal risky portfolio

Spectral Theorem

- A multi purpose fundamental result in matrix algebra is the so called "spectral theorem":

Theorem

If A is a $(n \times n)$ symmetric, pd matrix then there exist a $(n \times n)$ orthonormal matrix X and a diagonal matrix Λ such that:

$$A = X\Lambda X' = \sum_j \lambda_j x_j x_j'$$

where x_j is the $j-th$ column of X and is called the $j-th$ eigenvector of A, the elements λ_j on the diagonal of Λ are called the eigenvalues of A. These are positive, if A is pd, and can always be arranged (rearranging also the corresponding columns of X) in non increasing order.

Spectral Theorem

- The formula $A = X\Lambda X'$ is called the "spectral decomposition" of A.
- If the matrix A is only psd with rank $m < n$ a similar theorem holds but the matrix X has only M columns and the matrix Λ is a square $m \times m$ matrix.
- Notice that the spectral theorem implies that the rank of a p(s)d matrix A is equal to the number of non null eigenvalues of A.

Spectral Theorem

- A property of the eigenvectors of a pd matrix A is that $XX' = I$ (and, since in the pd case X is square, we also have $X'X = I$).
- That is: the eigenvectors are orthonormal.
- A nice result directly implied by this theorem (when A is pd) is this: $A^{-1} = X\Lambda^{-1}X'$.
- In fact $A^{-1} = (X\Lambda X')^{-1} = X'^{-1}\Lambda^{-1}X^{-1} = X\Lambda^{-1}X'$.

Spectral Theorem

- The Reader must notice that computing the spectral decomposition of a matrix, while straightforward from the numerical point of view, is by no mean easy to do by hand.
- In order to understand this we can post multiply A times the generic x_i:

$$Ax_i = X\Lambda X' x_i = \sum_j \lambda_j x_j x_j^{'} x_i$$

Spectral Theorem

- Each term in the sum is going to be equal to 0 (orthonormal $x_j$vectors) except the i_{th} which is going to be equal to $x_i\lambda_i$ so that x_i solves the equation $Ax_i = \lambda_i x_i$ which we could better write as $(A - \lambda_i I)x_i = 0$.
- For this homogeneous linear sistem to have a non trivial x_i solution we need to choose, if possible, a λ_i such that: $|A - \lambda_i I| = 0$
- Hence, eachλ_i-s is a root of the equation: $|A - \lambda I| = 0$ (the so called "characteristic equation" for A).
- If this determinant equation is written down in full it shall be seen that it is a polynomial equation in the variable λ of degree equal to the rank k of A (its size, if A is of full rank).

Spectral Theorem

- As it is well known since high school days, polynomial equations have k real and/or complex roots (the so called "fundamental theorem of algebra").
- However, an explicit formula for finding such roots "in radicals" (that is: using the four operations and root extraxtion) only exists (in the general case) for $k \leq 4$.
- On the other hand, finding the roots of this equation is so relevant a problem in applied Mathematics that numerical algorithms for computing them exist at least from the times of Newton.

Spectral Theorem, some implications*

- Notice that, if A is only psd, say of size k and rank $q < k$, then it has q eigenvectors x_l corresponding to positive eigenvalues
- However, there exist $k - q$ vectors z_j^*such that $(z_j^*)'z_j^* = 1$, $(z_j^*)'z_{j'}^* = 0$ for $j \neq j'$ and $(z_j^*)'x_l = 0$ for any vector x_l which is eigenvector of A.
- Using any of these z_j^* ,or any non zero scalar multiple of them, is then possible to build linear combinations of the random variables whose varcov matrix is A such that these linear combinations have zero variance.

Spectral Theorem, some implications*

- If A is a variance covariance of linear security returns, this implies that it is possible to build a portfolio of such securities and, possibly, the risk free rate, such that the single securities returns are random but the overall position return is non random
- Compare this with the example about the constrained minimization of a pd quadratic form
- By no arbitrage, any such position must yield the same risk free rate (otherwise you can borrow at the lower rate and invest at the higher rate for a sure, possibly unbounded, profit).
- This very important property is central in asset pricing theory and asset management

Spectral Theorem, some implications*

- As we shall see in the section dedicated to factor models and principal components, it is often the case that covariance matrices of returns for large sets of securities are approximately not of full rank
- That is: it may be that all eigenvalues are non zero but many of these are almost zero
- In this case it is possible to build portfolios of risky securities whose return is "almost" riskless
- This has important applications in (hedge) fund management and, more in general, trading and asset pricing

Conditional expectation (regression function)

- NOTE: not for the classroom
- You should read this if you do not remember what is a conditional expectation
- Remember these two formulas

$$E(Y) = E(E(Y|X))$$

$$V(Y) = E(V(Y|X)) + V(E(Y|X))$$

Conditional expectation (regression function)*

- The concept of conditional expectation is of paramount importance for the study of asset pricing and for correct understanding and interpretation of the linear model
- The conditional expectation of a (vector) Y given a (matrix) X is, simply, the expected value of the conditional distribution of Y given X: $E(Y|X)$
- This is a (vector) function of X and, as such, inherits its distribution from the distribution of X
- This function of X is sometimes also called "regression function" of Y on X (the name "regression" has historical origins and derives from an, at first non well understood, property of $E(Y|X)$)

Conditional expectation (regression function)*

- Being an expected value, the regression function shall share all the properties of expected values we discussed above, however it has some new properties which derive from the fact of being a function of X. Let us summarize these
- Suppose $Y = g(X)$. That is: Y is a function of X. Then $E(Y|X) = E(g(X)|X) = g(X)$. In other words: if, once we know X, we know also the value of Y as this is a function of X, then the conditional expectation is this function. Consider the implications of this joint with the linear property of the expected value operator
- $E(E(Y|X)) = E(Y)$. In this you must understand the meaning of the two E signs. Each expected value is made wrt a different distribution. The "outside" E is wrt the distribution of Z, that is $P(Z)$ the "inner" E is wrt the conditional distribution of Y GIVEN X

Conditional expectation (regression function)*

- $E(Y - E(Y|X)) = 0$ this is a corollary of the second property
- $E((Y - E(Y|X))E(Y|X)') = 0$ That is: suppose we use the regression function to "forecast" Y given X, all the covariances between the forecasts ($E(Y|X)$)and the forecast errors ($Y - E(Y|X)$) are zero

Conditional expectation (regression function)*

- Now we are ready for a very simple result of paramount relevance. This result justifies the use of the regression function in order to "forecast" Y on the basis of X
- Consider any (vector) function $h(X)$ (the vector has the same dimension as Y). Suppose you want to "forecast" Y using $h(X)$ and call $E(Y|X) = \phi(X)$
- Suppose you want to minimize the "expected forecast error" $E((Y - h(X)(Y - h(X))')$ In the sense that any other choice of $h^*(X)$ gives you an expected forecast error $E((Y - h^*(X)(Y - h^*(X))')$ such that $E((Y - h^*(X)(Y - h^*(X))') - E((Y - h(X)(Y - h(X))')$ is PSD
- The "best" way of choosing $h(X)$ is to set $h(X) = \phi(X) = E(Y|X)$

Conditional variance (covariance)*

- The conditional variance (covariance) of Y given X is the variance covariance of the conditional distribution of Y given X

$$V(Y|X) = E(YY'|X) - E(Y|X)E(Y'|X) =$$

$$= E((Y - E(Y|X))(Y - E(Y|X))'|X)$$

- This is just the variance around the regression function computed for any GIVEN X

Conditional variance (covariance)*

- Notice that

$$E(E((Y - E(Y|X))(Y - E(Y|X))'|X)) = E(V(Y|X)) =$$

$$= E((Y - E(Y|X))(Y - E(Y|X))')$$

- This is the measure of error we did minimize above: $E((Y - h(X)(Y - h(X))')$, and this is NOT the unconditional (sometimes: "marginal") variance (covariance) of Y which is:

$$V(Y) = E((Y - E(Y))(Y - E(Y))')$$

Conditional variance (covariance)*

- In fact, while for $E(Y)$ we have $E(Y) = E(E(Y|X))$, for the variance (covariance) we have a slightly more complicated formula $V(Y) = E(V(Y|X)) + V(E(Y|X))$ which involves the expected value of the conditional variance AND the variance of the regression function
- This is easy to prove (not required but useful exercise)

$$V(Y) = E(YY) - E(Y)E(Y') =$$

$$= E(E(YY'|X)) - E(E(Y|X))E(E(Y|X)') =$$

$$= E(V(Y|X)) + E(E(Y|X)E(Y|X)') - E(E(Y|X))E(E(Y|X)') =$$

$$= E(V(Y|X)) + V(E(Y|X))$$

Conditional variance (covariance)*

- The result is more interesting that it seems and is fundamental, for instance, in order to measure the quality of a forecast made with the regression function
- It tells us that the unconditional variance (covariance) of Y can be split in two parts: the expected value of variance around the regression function and the variance of the regression function
- For an ideal "perfect" forecast, i.e. when Y is a function $g(X)$ of X and, as we did see above, in this case $g(X) = E(Y|X)$, the first term would be equal to 0 and the unconditional variance of Y would be the same as that of its regression function on X simply because Y would be equal to this regression function
- This is obviously an extreme case we rarely observe

Lecture: Linear Regression 1

Summary

- We begin with a simple interpolation method, not a model
- It becomes a model if we add probability hypotheses
- The weak OLS hypotheses
- Some implication of the OLS hypotheses

The linear model in matrix notation

- Suppose you have a matrix X of dimensions $n \times k$ containing n observations on each of k variables

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

- You also have a $n \times 1$ vector Y containing n observations on another variable

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The linear model in matrix notation

- You would like to approximate Y with a linear function of X that is: $X\beta$ for some $k \times 1$ vector β

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

The linear model in matrix notation

- In general, if $n > k$ it shall not be possible to exactly fit $X\beta$ to Y so that the approximation shall imply a vector of errors

$$\varepsilon = Y - X\beta = \begin{bmatrix} y_1 - X_{1.}\beta \\ y_2 - X_{2.}\beta \\ \vdots \\ y_n - X_{n.}\beta \end{bmatrix} = \begin{bmatrix} y_1 - \sum_{j=1,\dots k} x_{1j}\beta_j \\ y_2 - \sum_{j=1,\dots k} x_{2j}\beta_j \\ \vdots \\ y_n - \sum_{j=1,\dots k} x_{nj}\beta_j \end{bmatrix}$$

- You would like to minimize ε but this is a vector not a scalar.
- So you have to define some scalar function of ε you want to minimize

The linear model in matrix notation

- A possible solution is $\varepsilon'\varepsilon$ that is: the sum of squares of the errors
- We then wish to minimize

$$\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) = Y'Y + \beta'X'X\beta - 2Y'X\beta$$

The linear model in matrix notation

- If we take the derivative of this w.r.t β we get

$$\frac{\partial}{\partial \beta}(Y'Y + \beta'X'X\beta - 2Y'X\beta) = 2X'X\beta - 2X'Y$$

- (again remember the size rule and remember that $Y'X\beta = \beta'X'Y$ each is the transpose of the other but both are scalars)
- Now, set this vector of derivatives to zero and get the system of equations $2X'X\beta - 2X'Y = 0$. The solution is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- This simple application of the rule for the approximate solution of an over determined system yields the most famous formula in applied (multivariate) statistics. But...

Linear regression. Weak OLS hp

- Now: all this is NOT a "model" except in the sense that it is a choice of an approximation method
- The choice is reasonable but arbitrary
- It is NOT a statistical model for Y and, even less, it is NOT a model for $E(Y|X)$
- It is NOT a "regression model", we cannot use it to justify forecasts based on it
- In order for it to become a model, we need to add hypotheses which connect actual observations to potential observations (sample to population)
- Since we shall use this model for making forecasts (not for interpolating data) such hypotheses should contain the reason for a forecasting use
- It turns out that there exist several choices of such hypotheses, all worded in the language of Statistics and Probability

Linear regression. Weak OLS hp

- We begin with the so called "weak" OLS hypotheses.
- Let Y be an $n \times 1$ random vector containing a sample of n (potential) observations on a "dependent" variable y.
- Let X be a non random $n \times k$ matrix of rank k, β a $k \times 1$ non random vector and ε a $n \times 1$ random vector.
- Let:

$$Y = X\beta + \varepsilon$$

- Suppose $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma_\varepsilon^2 I_n$.

Linear regression. Weak OLS hp

- These hypotheses are best understood in a statistical setting (that is: estimation in repeated samples).
- Each sample we are going to draw shall yield a "realization" of Y (i.e. a vector of numbers).
- In each sample X (observable) and β (unobservable) shall always be the same (this is the meaning of "non random").
- What makes Y "random" that is: changing in a (partially) unpredictable (for us) way from sample to sample, is the random "innovation" or "error" vector ε which we cannot observe.

Linear regression. Weak OLS hp

- As for the "partially" clause in "partially unpredictable": under the assumed hypotheses is clear that

$$E(Y) = E(Y|X) = X\beta$$

- So that the expected value of the random Y is not random as it is "made of" non random things (while it is unknown as β is unknown).
- In this sense we may say that we are modeling the regression function of Y on X that is: the conditional expectation of Y given X.
- However, since the matrix X is non random, the conditional expectation shall be the same as the unconditional expectation.
- A more interesting model, from the point of view of financial applications, shall be considered below when we shall allow X to be itself random.

Linear regression. The OLS estimator

- A note on the use of the term "estimator"
- In english language statistics literature, the terms "estimator" and "estimate" are often used in a promiscuous way
- A precise use should require the term "estimator" when we indicate that the sample is still to be observed and so we consider a random variable
- The term "estimate" should be used to indicate the specific numerical value on the estimator computed on a given sample
- In the handouts and in these slides we widely agree to the common promiscuous use, when the context is sufficient to avoid misunderstandings
- In this section, since we shall mostly deal with the statistical properties of the OLS estimator as a random vector, we shall use the term estimator and the term estimate in a rigorous way

Linear regression. The OLS estimator

- The problem is to estimate β.
- Under the above hypotheses, different estimation procedures lead all to the same estimate: the Ordinary Least Squares Estimator $\widehat{\beta}_{OLS}$.
- The simplest way for deriving $\widehat{\beta}_{OLS}$ is trough its namesake, that is: find the value $\widehat{\beta}_{OLS}$ that minimizes $\varepsilon'\varepsilon$: the sum of squared errors.
- This is what we just did, with NO statistical hypotheses
- Let's do it once more

- The objective function shall be:

$$\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

- The first derivatives with respect to β are:

$$\frac{\partial (Y - X\beta)'(Y - X\beta)}{\partial \beta} = 2X'X\beta - 2X'Y$$

- The system of "normal equations" (where we ask for the β that sets to zero the first derivatives) is:

$$X'X\beta = X'Y$$

- Since the rank of X is k, $X'X$ can be inverted and the (unique) solution of the system is:

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

- We do not check the second order conditions, we should!
- Informally: we see that we are minimizing a sum of squares.
- This may go to plus infinity, but not to minus infinity.
- Hence, our (single) stationary point should be a minimum non a maximum
- (this is by no means rigorous but it could be made so).

- Why should this "sum of squares minimizing $\hat{\beta}_{OLS}$" be a "good" estimator for β in our model?
- Use the statistical hypoteses you made to check that this estimate satisfies "goodness" statistical criteria for estimators
- It is easy to show that $\widehat{\beta}_{OLS}$ is unbiased for β.
- In fact:

$$E(\widehat{\beta}_{OLS}) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$$

- Here, in the first passage, we use the fact that X is non random and in the second one we use the hypothesis that β is non random and that $E(\varepsilon) = 0$.

- It is also easy to compute $V(\widehat{\beta}_{OLS})$:

$$V(\widehat{\beta}_{OLS}) = (X'X)^{-1}X'V(Y)X(X'X)^{-1} =$$

$$= (X'X)^{-1}X'\sigma_\varepsilon^2 I_n X(X'X)^{-1} =$$

$$= \sigma_\varepsilon^2 (X'X)^{-1}X'X(X'X)^{-1} =$$

$$= \sigma_\varepsilon^2 (X'X)^{-1}$$

Lecture: Linear Regression 2

Summary

- The fundamental theorem of OLS analysis: The Gauss-Markoff theorem
- The OLS as an "optimal" estimator
- Definition of "optimal" and its implications
- Proof of the theorem under weak OLS hypotheses with diagonal varcov matrics of the errors
- The same with a general varcov matrics of the errors

Linear regression. The Gauss Markoff theorem

- The choice of $\widehat{\beta}_{OLS}$ as an estimator of β based only on the minimization of $\varepsilon'\varepsilon$ could be disputed: in which sense this should be a "good" estimate from a statistical point of view?
- In fact, this estimator just minimizes a specific measure of "error of fit"
- OK it is unbiased but...
- Is it the case that some different, and in some sense "better", estimators exist?
- A strong result in favor of $\widehat{\beta}_{OLS}$ and against the use of different estimators, at least under weak OLS hypotheses, is the Gauss-Markoff theorem.

- First: we need a reasonable definition of "better" or at lest "not worse"

Definition

If β^* and β^{**} are both unbiased estimators of β we say that β^* is "not worse" than β^{**} iff $D = V(\beta^{**}) - V(\beta^*)$ is at least a psd matrix.

- Notice that in the univariate case this definition boils down to the standard definition of "more efficient unbiased estimator".
- Let us now justify the multivariate version.
- Suppose what you really want is to estimate a set of linear functions of β say $Q\beta$ where Q is any nonrandom matrix (Q could also be the identity I).
- Suppose you know that β^* is better than β^{**} according to the previous definition.

- In this case it is easy to show that $Q\beta^*$ is better than $Q\beta^{**}$, according to the same definition, as an estimator of $Q\beta$
- In fact:

$$V(Q\beta^{**}) - V(Q\beta^*) = Q(V(\beta^{**}) - V(\beta^*))Q' = QDQ'$$

- And if D is at least psd then QDQ' is at least psd (why?)

- This "invariance to linear transform" property is the strongest argument in favor of this definition of "not worse" estimator.
- Obviously this definition of not worse would amount to little, if the estimators were allowed to be biased.
- In this case any vector of constants would be not worse than any other estimator, in terms of variance.
- For this reason in the definition we require unbiasedness.

Linear regression. The Gauss Markoff theorem

- We are now just a step short of being able to state an important result in OLS theory.
- We would like to prove a theorem of the kind: the best unbiased estimator of β is $\hat{\beta}_{OLS}$.
- Alas, this is actually not true in this generality.
- The theorem results to be true, if we further ask for the class of competing estimators to be linear in Y that is: each competing estimator must be of the form HY with H a known nonrandom matrix.

Definition

We say that $\hat{\beta}$ is a linear estimator of β iff $\hat{\beta} = HY$ where H is a non random matrix.

- We thus arrive to the celebrated Gauss Markoff theorem.

Theorem

Under the weak OLS hypotheses $\widehat{\beta}_{OLS}$ is the Best Linear Unbiased Estimator (BLUE) of β.

Proof.

Any linear estimator of β can be written as $\hat{\beta} = ((X'X)^{-1}X' + C)Y$ with an arbitrary C. The estimator must be unbiased, so:

$$E(\hat{\beta}) = ((X'X)^{-1}X' + C)X\beta = \beta + CX\beta = \beta$$

This is possible only if $CX = 0$. Let us now compute $V(\hat{\beta})$:

$$V(\hat{\beta}) = \sigma_\varepsilon^2((X'X)^{-1}X' + C)((X'X)^{-1}X' + C)' =$$

$$= \sigma_\varepsilon^2((X'X)^{-1} + CC' + (X'X)^{-1}X'C' + CX(X'X)^{-1})$$

Recalling that $CX = 0$, due to unbiasedness, we have:

$$V(\hat{\beta}) = \sigma_\varepsilon^2((X'X)^{-1} + CC')$$

This is $V(\widehat{\beta}_{OLS})$ plus $\sigma_\varepsilon^2 CC'$ which is an at least psd matrix. □

Linear regression. The Gauss Markoff theorem

- We have shown that the covariance matrix of any linear unbiased estimator of β can be written as the covariance matrix of $\widehat{\beta}_{OLS}$ plus $\sigma_\varepsilon^2 CC'$ which is at least psd
- If $C \neq 0$ this is "worse", according to the Gauss Markoff definition, than the OLS estimator
- The best we can do, then, is to set $C = 0$
- This takes us back to the OLS estimator
- To summarize: we have shown that $\widehat{\beta}_{OLS}$ is BLUE

- As we shall see in the "stochastic X" section, the Gauss-Markoff theorem still holds under suitable modifications of the weak OLS hypotheses in the case of stochastic X.
- There is an equivalent theorem which is valid when $V(\varepsilon) = \Sigma$ with Σ any non random (pd) matrix known up to a multiplicative constant.
- In this case the BLUE estimator is

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

- where GLS stands for Generalized Least Squares, but we are not going to use this in this course.
- Notice that if Σ is not known up to a multiplicative constant the above is not an estimator.

- The proof begins by recalling that any pd matrix has a pd inverse.
- Moreover any pd matrix A can be written as $A = PP'$ with P invertible.
- We then have $\Sigma = PP'$ and $\Sigma^{-1} = (PP')^{-1} = P'^{-1}P^{-1}$.
- Now, multiply the model $Y = X\beta + \varepsilon$ times P^{-1} as $P^{-1}Y = P^{-1}X + P^{-1}\varepsilon$.

- Call now $Y^* = P^{-1}Y$;$X^* = P^{-1}X$ and $\varepsilon^* = P^{-1}\varepsilon$.
- We have that $Y^* = X^*\beta + \varepsilon^*$ satisfies the weak OLS hypotheses
- In fact:

$$V(\varepsilon^*) = V(P^{-1}\varepsilon) = P^{-1}\Sigma P'^{-1} = P^{-1}PP'P'^{-1} = I$$

- We can then follow the standard proof up to the result:

$$\hat{\beta}^*_{OLS} = (X'^* X^*)^{-1} X'^* Y^*$$

- This $\hat{\beta}^*_{OLS}$ is BLUE on the transformed data
- The transform of the data: $Y^* = P^{-1}Y$;$X^* = P^{-1}X$ and $\varepsilon^* = P^{-1}\varepsilon$ was linear and invertible
- We can transform back to the original data and we get

$$\hat{\beta}^*_{OLS} = (X' P'^{-1} P^{-1} X)^{-1} X' P'^{-1} P^{-1} Y =$$

$$= (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y = \hat{\beta}_{GLS}$$

- Where GLS stands for Generalized Least Squares.

- The result seems very general and, in a sense, it is so.
- However, we should take into account that the above proof implies Σ non diagonal but KNOWN.
- Otherwise we could not compute P and the estimator.

- In most cases of linear model with correlated residuals, Σ is not "known" and the GLS estimate cannot directly be used.
- In these cases some estimator of Σ is used in the place of the "true" Σ.
- A wide number of different regression models used in practice can be seen as versions of the GLS estimator where Σ is itself "estimated" in some way.
- We do not consider this topic (interesting and very relevant) in this introductory course.

Linear regression. The Gauss Markoff theorem

- A last comment about these results
- Under the standard weak OLS hypotheses, both with non stochastic and with stochastic X, the OLS estimator works fantastically well: it minimizes the sum of squares of the errors, is unbiased, is BLUE.
- This is perhaps too much and we should surmise that some of this bonanza strictly depends on a clever choice of the hypotheses.
- This is exactly the case.
- As an instance: we did just prove that, even in the case of non stochastic X, when the covariance matrix of residuals is NOT $\sigma_\varepsilon^2 I$, the BLUE estimator is NOT the OLS estimate.
- In this case "minimizing the sum of squared errors" is not equivalent to finding the "best" estimate in the Gauss Markoff sense.

Lecture: Linear Regression 3

Summary

- Definitions and statistical properties of "fit" and "errors of fit"
- How much did we minimize the sum of squared errors?
- R^2 and variance decomposition (with stress on the role of the intercept)

Linear regression. Fit and errors of fit

- We call $\hat{Y} = X\hat{\beta}_{OLS}$ the "fitted" values of Y. In fact $\hat{Y}$ is to be understood as an estimator of $E(Y) = X\beta$.
- On the other hand $\hat{\varepsilon} = Y - \hat{Y}$ bear the name of "errors of fit".
- Notice that, by the first order conditions of least squares, we have:

$$X'X\hat{\beta}_{OLS} = X'Y$$

$$X'(X\hat{\beta}_{OLS} - Y) = 0$$

$$X'\hat{\varepsilon} = 0$$

Linear regression. Fit and errors of fit

- This in particular implies

$$\widehat{\beta}'_{OLS} X' \hat{\varepsilon} = \hat{Y}' \hat{\varepsilon} = 0$$

- This result is independent on the OLS hypotheses and depends only on the fact that $\widehat{\beta}_{OLS}$ minimizes the sum of squared errors.

Linear regression. R^2

- A useful consequence of this result, joint with the assumption that the first column of X is a column of ones, allows us to define an index of "goodness of fit" (read: how much did I minimize the squares?) called R^2.
- In fact:
$$Y'Y = (\hat{Y} + \hat{\varepsilon})'(\hat{Y} + \hat{\varepsilon}) = \hat{Y}'\hat{Y} + \hat{\varepsilon}'\hat{\varepsilon}$$
- where the last equality comes from the fact that $\hat{Y}'\hat{\varepsilon} = 0$.

Linear regression. R^2

- Moreover if X contains as first column a column of ones $X'\hat{\varepsilon} = 0$ implies that the sum, hence the arithmetic average, of the vector $\hat{\varepsilon}$: $\bar{\hat{\varepsilon}}$ is equal to zero. So:

$$\bar{Y} = \bar{\hat{Y}} + \bar{\hat{\varepsilon}} = \bar{\hat{Y}}$$

- and:

$$Y'Y - n\bar{Y}^2 = \hat{Y}'\hat{Y} - n\bar{\hat{Y}}^2 + \hat{\varepsilon}'\hat{\varepsilon}$$

- where n is the length of the vectors (number of observations).

Linear regression. R^2

- In other words, indicating with $Var(Y)$ the *numerical* variance of the vector Y (that is: the mean of the squares minus the squared mean), we have:

$$Var(Y) = Var(\hat{Y}) + Var(\hat{\varepsilon})$$

- We see that the variance of Y neatly decomposes in two non negative parts. There is no covariance!
- This is totally peculiar to the use of least squares and implies the definition of a very natural measure of goodness of fit:

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = 1 - \frac{Var(\hat{\varepsilon})}{Var(Y)}$$

Linear regression. R^2

- Notice that, in order to be meaningful, R^2 requires for the presence of a column of ones (or any constant, in fact) in X.
- Otherwise the mean of the errors of fit may be different than 0 and the passage from sum of squares to variances is no more possible (the mean of $\hat{Y}$ shall not, in general, be the same as the mean of Y).
- Due to sampling error we may expect to observe a positive R^2 even when there is non "regression" between Y and X.
- We can also say something about the expected size of an R^2 when it should actually be 0.

Linear regression. R^2

- In fact, suppose, conditionally to a $n \times k$ matrix X of regressors, Y is a vector of n iid random variables of variance σ^2 so that there should be an R^2 of 0 out of the regression of Y on X.
- However, under weak OLS hp it is immediate to prove that the expected value of $Var(Y)$ (that is: the denominator of the R^2) is σ_ε^2
- Under the same hypotheses, the expected value of sampling variance of the elements of the error of fit vector $\hat{\varepsilon}$ is $\sigma_\varepsilon^2 \frac{n-k}{n}$
- Hence, the expected value of the sampling variance of $\hat{Y}$ is going to be $\sigma_\varepsilon^2 - \sigma_\varepsilon^2 \frac{n-k}{n} = \sigma^2 \frac{k}{n}$ and this is the numerator of the R^2.
- (Recall: $Y = \hat{Y} + \hat{\varepsilon}$ and $Var(Y) = Var(\hat{Y}) + Var(\hat{\varepsilon})$ due to the zero correlation between $\hat{Y}$ and $\hat{\varepsilon}$)

Linear regression. R^2

- While we know that the expected value of the ratio is not the ratio of the expected values this could be a good first order approximation
- (If you wish for a "sketch of proof", not required for the exam, read the rest of this slide)
- Expand z_1/z_2 around the point $E(z_1)/E(z_2)$ up to the first order
- This gives $z_1/z_2 \approx E(z_1)/E(z_2) + (z_1 - E(z_1))/E(z_2) - (z_2 - E(z_2))E(z_1)/E(z_2)^2$
- Take the expected value on both sides and recal that $E(z_i - E(z_i)) = 0$. You get $E(z_1/z_2) \approx E(z_1)/E(z_2)$

Linear regression. R^2

- Using this approximation, we can say that, in the case of no regression at all between Y and X, the expected value of the R^2 is approximated by

$$\frac{\sigma^2 \frac{k}{n}}{\sigma^2} = \frac{k}{n}$$

- This number could be quite big if you use many variables and do not have many observations.

Linear regression. R^2

- In fields the like of social sciences, where the complexity of observed phenomena, arguably depending in variable way on a multitude of underlying, mostly unobservable, variables, should imply very small "true" $R^2 - s$ for any simple model.
- So, often, if n is not so big, $\frac{k}{n}$ could be quite similar to the observed value of R^2.
- This should not stop us from using linear regression as an inferential tool, but should make us careful in avoiding what is usually called "overfitting".
- So, once more, be careful!

Lecture: Linear Regression 4

Summary

- Some further implications of the weak OLS hypotheses
- Strong OLS hypoteses and testing linear hypotheses in the linear model
- What happens if X is stochastic
- "Forecasts"

Linear regression. More properties of $\hat{Y}$ and $\hat{\varepsilon}$

- Let us now study some property of $\hat{Y}$ and $\hat{\varepsilon}$ depending on the OLS hypotheses. First we compute the expected values and covariance matrices of both vectors.

$$E(\hat{\varepsilon}) = E(Y - X\widehat{\beta}_{OLS}) = X\beta - X\beta = 0$$

$$V(\hat{\varepsilon}) = V(Y - X\widehat{\beta}_{OLS}) = V(Y - X(X'X)^{-1}X'Y) =$$

$$= V((I - X(X'X)^{-1}X')Y) =$$

$$= (I - X(X'X)^{-1}X')\sigma_{\varepsilon}^2 I(I - X(X'X)^{-1}X') =$$

$$= \sigma_{\varepsilon}^2(I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') = \sigma_{\varepsilon}^2(I - X(X'X)^{-1}X')$$

Linear regression. More properties of $\hat{Y}$ and $\hat{\varepsilon}$

- This because, by direct computation, we see that

$$(I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') = (I - X(X'X)^{-1}X')$$

- That is: $(I - X(X'X)^{-1}X')$ is an idempotent matrix.

Linear regression. More properties of $\hat{Y}$ and $\hat{\varepsilon}$

- With similar passages we find that:

$$E(\hat{Y}) = X\beta$$

$$V(\hat{Y}) = \sigma_{\varepsilon}^2 X(X'X)^{-1}X'$$

Linear regression. More properties of $\hat{Y}$ and $\hat{\varepsilon}$

- In summary: we see that $\hat{Y}$ is indeed an unbiased estimator of $E(Y)$.
- On the other hand, we see that $\hat{\varepsilon}$ shows a non diagonal correlation matrix even if (or, better, just because) the vector ε is made of uncorrelated errors.
- This property of the estimated residuals is, in some sense, unsatisfactory and led some researchers into defining a different estimator of the residuals (non OLS based) with the property of being uncorrelated under OLS hypotheses.

Linear regression. Testing linear hypotheses in the linear model

- This is a very short section for a very relevant topic.
- We are not going to deal with general testing of linear hypotheses but only with those tests which are routinely found in the output of standard OLS regression computer programs.
- Let us begin with introducing Strong OLS hypotheses.
- In short: these are the same as the weak OLS hypotheses (with non stochastic X) plus the hypothesis that the ε vector is not only made of uncorrelated, zero expected value and constant variance random variable but also is a vector distributed as an n dimensional Gaussian density with expectation vector made of zeros and diagonal variance covariance matrix $\sigma_\varepsilon^2 I_n$.

Linear regression. Testing linear hypotheses in the linear model

- When we wish to test hypotheses we need to find distributions of sample functions, for instance we are going to need the distribution of $\hat{\beta}_{OLS}$.
- Up to now we know that under weak OLS hypotheses $\hat{\beta}_{OLS}$ has expected value vector β and variance covariance matrix $\sigma_\varepsilon^2(X'X)^{-1}$.
- Moreover we know that $\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'\varepsilon$ that is: it is a linear function of ε (and β and X are non stochastic).

Linear regression. Testing linear hypotheses in the linear model

- An important result is that, if z is a $q \times 1$ Gaussian random vector with $E(z) = \mu_z$ and $V(z) = Q$, and if A and B are non stochastic and such that $s = A + Bz$ is defined, then s is Gaussian with $E(s) = A + B\mu_z$and $V(s) = BQB'$
- We can then conclude that, IF ε is Gaussian, THEN $\hat{\beta}_{OLS}$ has Gaussian distribution with expected value vector β and variance covariance matrix $\sigma_\varepsilon^2 (X'X)^{-1}$.

Linear regression. Testing linear hypotheses in the linear model

- Adding the Gaussian hypothesis to the weak OLS hypotheses makes them into the strong OLS hypotheses
- Well...This may seem too expedient: OK, computations are now simple but why Gaussian errors?
- In fact it often IS too expedient, and the pros and cons of the hypothesis could (and are) discussed at length.
- For the sake of this course, we shall take it as a beginner's choice in the econometric world we live in, a choice to be taken with much care.

Linear regression. Testing linear hypotheses in the linear model

- We suppose everybody knows what a Statistical Hypothesis is (see Appendix in case).
- We now define a "linear" statistical hypothesis.
- A linear hypothesis on β can be written as $R\beta = c$ where R is a matrix of known constants and c a vector of known constants.
- For the purpose of this summary we shall concentrate on two particular R:

1. a $1 \times k$ vector R where only the j^{th} element is 1 and the others are zeros
2. a $(k-1) \times k$ matrix where the first column is of zeros an the remaining $(k-1) \times (k-1)$ square matrix is an identity.

- In both cases c is made of zeros (in the first case a single 0 and in the second a $k-1$ vector of zeros).

Linear regression. Testing linear hypotheses in the linear model

- The first kind of hypothesis states that the j^{th} element of β is zero (while all other parameters are free to have any value)
- The second kind of hypothesis states that all parameters (but the intercept which is not constrained) are *jointly* zero
- For (non trivial) historical reasons these hypothesis are considered so frequently relevant that any program for OLS regression tests them.
- The fact that these hypotheses be of *your* interest is for you to evaluate.
- We shall not detail the procedure for the test of the second hypothesis that all parameters, except, possibly, the intercept are jointly equal to zero.
- Just remember the fact that the result of this test is displayed, usually, in any OLS program output. The name of this test is F test.

Linear regression. Testing linear hypotheses in the linear model

- A little more detail on the test of the first hypothesis.
- The standard test for the Hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ (and the other parameters unconstrained) requires the distribution of $\hat{\beta}_{OLS}$
- This, as we wrote above, a strengthening of the OLS hypotheses which takes the form of the assumption that ε is distributed according to an $n-$variate Gaussian: $\varepsilon \approx> N_n(0, \sigma_\varepsilon^2 I_n)$
- Alternatively, a CLT argument could be used for the same purpose

Linear regression. Testing linear hypotheses in the linear model

- Under this hypothesis: $\widehat{\beta}_{OLS} \approx> N_k(\beta, \sigma_\varepsilon^2 (X'X)^{-1})$.
- Hence the ratio:

$$\frac{\hat{\beta}_{OLSj} - \beta_j}{\sqrt{\sigma_\varepsilon^2 \{(X'X)^{-1}\}_{jj}}}$$

- Is distributed according to a standard Gaussian (i.e. $N_1(0,1)$).

Linear regression. Testing linear hypotheses in the linear model

- Suppose now we set $\beta_j = 0$ in the above ratio. In this case the distribution of the ratio shall be a standard Gaussian only if $H_0 : \beta_j = 0$ is true.
- This allows us to define a reasonable rejection region for our test.
- Reject $H_0 : \beta_j = 0$ with a size of the error of the first kind equal to α iff:

$$\frac{\hat{\beta}_{OLSj}}{\sqrt{\sigma_\varepsilon^2 \{(X'X)^{-1}\}_{jj}}} \notin [-z_{1-\frac{\alpha}{2}}; +z_{1-\frac{\alpha}{2}}]$$

- (Because, if $H_0 : \beta_j = 0$ is true, there is a probability of "only" α of NOT observing $\hat{\beta}_{OLSj}$ inside the interval $[-z_{1-\frac{\alpha}{2}}; +z_{1-\frac{\alpha}{2}}]$)

Linear regression. Testing linear hypotheses in the linear model

- In the above formula $\{(X'X)^{-1}\}_{jj}$ is the $j-$th element in the diagonal of $(X'X)^{-1}$.
- $z_{1-\frac{\alpha}{2}}$ is the quantile of the standard Gaussian which leaves on its left $1-\frac{\alpha}{2}$ of probability.

- This solves the problem if σ_ε^2 is known.
- In the (likely) case σ_ε^2 is unknown, estimate it with:

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

- and use as critical region:

$$\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_\varepsilon^2 \left\{(X'X)^{-1}\right\}_{jj}}} \notin \left[-_{n-k}t_{1-\frac{\alpha}{2}}; +_{n-k}t_{1-\frac{\alpha}{2}}\right]$$

- where ${}_{n-k}t_{1-\frac{\alpha}{2}}$ is the quantile in a T distribution with $n-k$ degrees of freedom which leaves on its left a probability equal to $1-\frac{\alpha}{2}$.

Linear regression. Testing linear hypotheses in the linear model

- The use of the T distribution is the reason for the name given to this test: the T test.
- Another test whose results are as a rule reported in all standard outputs of a regression package is the F test.
- As in the case of The T test there exist many different hypotheses which can be tested using a F test but the standard hypothesis tested by the universally reported F test is:
- H_0 : all the betas corresponding to non constant regressors are jointly equal to 0
- H_1 : at least one of the above mentioned betas is not 0
- The Idea is that, if the null is accepted (i.e. big P-value), no "regression" exists (provided we did not make an error of the second kind, obviously). Hence the popularity of the F test.

Linear regression. A note on P-values

The standard procedure for a test is:

- Choose H_0 and H_1 (they should be an exclusive and exhaustive partition of the parameter space).
- Choose a size of the error of the first kind: α. Be careful: too little α implies usually a big error of the second kind. Your choice should be based on a careful analysis of the costs implied in the two kind of errors.
- Find a critical region for which the maximum size of error of the first kind is α and, possibly, with a sensible size of error of the second kind.
- Reject H_0 if your sample falls in the critical region and accept it otherwise.

Linear regression. A note on P-values

- This procedure typically requires the availability of statistical tables.
- When computer programs for performing tests came to the fore two alternative paths were possible.
- Let the user input the α for each test and, as output, point out if the null hypothesis is accepted or rejected on the observed dataset with that α.
- Let the user input nothing, but give the user as output the value of α for which the observed data would have been exactly on the brink of the rejection region.
- This value is called the P value.

- With this information the user, knowledgeable about his or her preferred α, would be able to state whether the null hypothesis was accepted or rejected by simply comparing the preferred α with the value given by the program.
- If the researcher's α is smaller than the value given by the program, this means that the observed data is inside the acceptance region as would have been computed by the researcher, so that H_0 is accepted.
- If the researcher's α is bigger than the value given by the program, this means that the observed data is outside the acceptance region as would have been computed by the researcher, so that H_0 is rejected.

Linear regression. A note on P-values

- Conceived for the simple reason of avoiding the use of tables, the P value has become, in the hands of Statistics illiterates, the source or numberless misunderstandings and sometimes amusing bad behaviors.
- The typical example is the use of terms like "highly significant" for the case of null hypotheses rejected with small P values or the use of "stars" for indicating in a pictorial way "significance level" of an hypothesis.
- Another quite wrong and actually strange, but widespread, attitude is that of considering optimal, a posteriori, a small P value "strongly significant" which, a priori, would never be considered a proper α value.
- Sometimes, worst of all, in an explicit and, more frequently, implicit way, the P value is considered as an estimate of the probability for the null hypothesis to be true given the data.

Linear regression. A note on P-values

- An amazing error, since in testing theory we only compute the probability of falling in the rejection region given the truth of the null hypothesis (our α) and not the probability of the null hypothesis to be true, given the data.
- Please, avoid this and other trivial errors: the fact that such errors are widespread, even in the scientific community, does not make them less wrong.
- If you find these errors in some published paper, and you will, don't use this as an excuse for making the same mistake
- Exercise: under the hypotheses used for analyzing the t-test, find confidence intervals for (univariate) linear functions of β.

Linear regression. Confidence intervals

- Just a word on confidence intervals
- As we just did see, under strong OLS hypotheses, for any given non stochastic A and B we have:

$$A + B\hat{\beta}_{OLS} \approx> N(A + B\beta; \sigma_\varepsilon^2 B(X'X)^{-1}B')$$

- This allows us to build both rejection/acceptance regions for tests and confidence intervals (obviously related) on linear functions of $\hat{\beta}_{OLS}$

Linear regression. Confidence intervals

- For tests we dis see that a common choice of A, B is a null A and $B = [0, 0, ...1..., 0]$ where the 1 is in the $j-th$ position
- In this case

$$A + B\hat{\beta}_{OLS} = \hat{\beta}_{OLS,j} \approx> N(\beta_j; \sigma_\varepsilon^2 [(X'X)^{-1}]_{j,j})$$

- On this basis we can build a confidence interval for β_j

$$[\hat{\beta}_{OLS,j} \pm z_{1-\alpha/2} \sigma_\varepsilon \sqrt{[(X'X)^{-1}]_{j,j}}]$$

Linear regression. Confidence intervals

- Another example is a confidence interval for the difference of two elements in β, say $\beta_{j_1} - \beta_{j_2}$
- In this case B shall be a row vector of zeroes except in the positions j_1(where the value is 1) and j_2 (where the value is -1)
- the confidence interval shall then be

$$[(\hat{\beta}_{OLS,j_1} - \hat{\beta}_{OLS,j_2}) \pm$$

$$\pm z_{1-\alpha/2}\sigma_\varepsilon \sqrt{[(X'X)^{-1}]_{j_1,j_1} + [(X'X)^{-1}]_{j_2,j_2} - 2[(X'X)^{-1}]_{j_1,j_2}}]$$

- Proceed in similar ways for any combination of elements in β

- When σ_ε^2 is unknown, use the same formulas just changing σ_ε^2 into $\hat{\sigma}_\varepsilon^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$ and $z_{1-\alpha/2}$into $_{n-k}t_{1-\alpha/2}$
- This is the $(1-\alpha/2)$ quantile of a T distribution having as degrees of freedom the difference between n and k: the number of rows and columns in X, the regressors matrix in the estimation sample.

Linear regression. Stochastic X

- In applications of the linear model to Economics and Finance, only infrequently is it realistic to suppose X to be a non random matrix.
- Typically the X shall be as much random (that is unknown before observation and variable between samples) than Y.
- Just think to the CAPM "beta" regression where the "dependent" variable is the excess return of a stock and the "independent" variable is the contemporaneous excess return of a market index which contains the same stock.
- If X is random, the results we gave above about $\widehat{\beta}_{OLS}$ are no more valid, in general.

Linear regression. Stochastic X

- Under the hypothesis of a stochastic matrix X we can extend the OLS results in several ways.
- Each of these ways means the addition of an hypothesis to the standard weak OLS setting.
- Here I choose a very simple way which shall be enough for our purposes.
- We shall extend the weak OLS hypotheses in this way:

$$E(\varepsilon|X) = 0$$

$$V(\varepsilon|X) = E(\varepsilon\varepsilon'|X) = \sigma_\varepsilon^2 I_n$$

Linear regression. Stochastic X

- In other words we shall assume that, conditionally on X, it is true what we assumed unconditionally in the weak OLS hypotheses setting.
- It is clear that our new hypotheses imply the old one, not vice verse.
- Notice that the equality between conditional covariance matrix and conditional second moments matrix is true only because we assumed that the conditional expectation of ε is zero. (See by yourself what happens otherwise).
- An immediate result is that:

$$E(Y|X) = E(X\beta + \varepsilon|X) = X\beta$$

- and this property justifies the name "linear regression" for our model.

Linear regression. Stochastic X

- With a stochastic X, and our added hypothesis, our model becomes a linear model for a conditional expectation: a regression function.
- This was true in the non stochastic X case, but not very interesting as X, being non stochastic, could only assume a given set of values, so that $E(Y|X) = E(Y)$

Linear regression. Stochastic X

- Let us see now if the OLS estimator is still unbiased in the new setting.

$$E(\widehat{\beta}_{OLS}) = E(\beta + (X'X)^{-1}X'\varepsilon) = \beta + E_X(E_{\varepsilon|X}((X'X)^{-1}X'\varepsilon|X)) =$$

$$= \beta + E_X((X'X)^{-1}X'E_{\varepsilon|X}(\varepsilon|X)) = \beta$$

- Notice the use of the iterated expectation rule, in general: $E(f(a,b)) = E_a(E_{b|a}(f(a,b)|a))$.

- Now compute the variance covariance matrix.

$$V(\widehat{\beta}_{OLS}) = V(\beta + (X'X)^{-1}X'\varepsilon) = V((X'X)^{-1}X'\varepsilon) =$$

$$= E((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}) - E((X'X)^{-1}X'\varepsilon)E(\varepsilon'X(X'X)^{-1}) =$$

- The second term was just shown to be equal to 0 so, again using iteraded expectations:

$$= E((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}) =$$

$$= E_X((X'X)^{-1}X'E_{\varepsilon|X}(\varepsilon\varepsilon'|X)X(X'X)^{-1}) =$$

- Now the term $E_{\varepsilon|X}(\varepsilon\varepsilon'|X)$ is, by hypothesis, equal to $\sigma_\varepsilon^2 I_n$ so that:

$$= E_X((X'X)^{-1}X'\sigma_\varepsilon^2 I_n X(X'X)^{-1}) =$$

$$= \sigma_\varepsilon^2 E_X((X'X)^{-1}X'X(X'X)^{-1}) = \sigma_\varepsilon^2 E_X((X'X)^{-1})$$

Linear regression. Stochastic X

- In short: with a stochastic X and the new OLS hypotheses, $\widehat{\beta}_{OLS}$ is still unbiased but now its covariance matrix is "fully unknown" as it depends on the expected value of $X'X$ and not on the observed value of $X'X$.
- We conclude this section with just a hint at two results of standard non stochastic X OLS theory, one of which can and the other cannot (in general) be proved to hold in the new setting.

Linear regression. Stochastic X

- The first result is that, with a suitable modification of the definition of "best", a Gauss-Markoff theorem still holds.
- Let us see how this works.
- In case of stochastic X, the theorem proof (using OLS hypotheses for the case of stochastic X) goes, now conditional on X, exactly as before.
- The last statement then becomes

$$V(\hat{\beta}|X) = \sigma_\varepsilon^2((X'X)^{-1} + CC')$$

Linear regression. Stochastic X

- We must then find $V(\hat{\beta})$. In the general case, this is NOT $E(V(\hat{\beta}|X))$ it is $E(V(\hat{\beta}|X)) + V(E(\hat{\beta}|X))$.
- However, we have $E(\hat{\beta}|X) = \beta$ (due to unbiasedness), so that $E(\hat{\beta}|X)$ is a constant (vector) i.e. it has no variance.
- Hence, the term $V(E(\hat{\beta}|X))$ is always equal to 0 if you use OLS under (conditional to X) weak hypotheses.
- We then have

$$V(\hat{\beta}) = E(V(\hat{\beta}|X)) = \sigma_\varepsilon^2(E((X'X)^{-1}) + E(CC'))$$

- This is the varcov matrix of the OLS estimate when X is stochastic plus the expected value of an at least psd matrix.

Linear regression. Stochastic X

- If you are puzzled by the term $E(CC')$ consider that, while CC' is non stochastic GIVEN X, it can be stochastic unconditional to X so the expected value is not, in general, the (trivial) expected value of a matrix of constants
- We must show that $E(CC')$ is at least psd
- This is easy, we must show that for any (non stochastic) vector z we have $z'E(CC')z \geq 0$ but, by the basic properties of the expected value operator and since z is non stochastic, we have $z'E(CC')z = E(z'CC'z)$.
- What is in the expected value is a (in general random) sum of squares. So it is a non negative random variable: you do not know its value but you know the value shall be non negative.
- The expected value of a non negative random variable cannot be negative (internality property of the expected value, mean value theorem of integral calculus) and we have our proof: as in the non stochastic X case the best we can do is to set C to 0

Linear regression. Stochastic X

- The second result is about the (very basic) statistical hypothesis testing setting discussed in the non stochastic X case
- The basic results underlying hypothesis testing do not directly hold in the new setting, except if we work conditionally on X (which somewhat vanifies the scope of the new setting) and under the (strong) hypothesis that, conditional on X, the distribution of ε is Gaussian.
- However, for all those tests whose distribution is independent on X when X is non stochastic, as, for example, the $T-$test, the assumption of conditional Gaussianity of ε given X implies that the standard result in the non stochastic X case still holds for stochastic X.

Linear regression. Stochastic X

- Consider, for instance, the standard confidence interval for a generic β_j: $[\hat{\beta}_{j,OLS} \pm_{n-k} t_{1-\alpha/2} s\hat{\sigma}_\varepsilon \sqrt{(X'X)^{-1}_{jj}}]$
- This contains in its specification $(X'X)^{-1}_{jj}$ which is stochastic if X is stochastic while, in the non stochastic X or conditional to X case, only $\hat{\beta}_{j,OLS}$ is stochastic
- However, both in the case of stochastic and non stochastic X, the interval with random extremes $[\hat{\beta}_{j,OLS} \pm_{n-k} t_{1-\alpha/2} s\hat{\sigma}_\varepsilon \sqrt{(X'X)^{-1}_{jj}}]$ contains β_j with probability $1-\alpha$ (that is: a probability which does not depend on X)
- Hence, if, as usually is the case, a confidence interval is used only after the observation of the data, $(X'X)^{-1}jj$ included, its interpretation and use does not change be X is stochastic or not
- The same for the critical region of a test
- Things are different if we consider, as in what follows, "forecast" intervals

Linear regression. "Forecasts"

- Here we intend the term "forecast" in a very narrow meaning.
- Suppose you estimated β using a sample of Y and X, let us sat n observations (rows).
- Now suppose a new set of q rows of X is given to you and you are asked to assess what could be the corresponding Y.
- Set as this the question does not allow for an answer. We need to assume some connection between the old rows and the new rows of data.
- A possibility is as follows.
- Let the model for the n rows of data used to estimate β with $\hat{\beta}_{OLS}$ be

$$Y = X\beta + \varepsilon$$

Linear regression. "Forecasts"

- Suppose we now have data for m more lines for the variables in X and call these X_f. Let the model for the corresponding new "potential" observations be

$$Y_f = X_f\beta + \varepsilon_f$$

- And suppose (we consider here the general case where X can be stochastic) that $E(\varepsilon|X, X_f) = 0 = E(\varepsilon_f|X, X_f)$, $V(\varepsilon|X, X_f) = \sigma_\varepsilon^2 I_n, V(\varepsilon_f|X, X_f) = \sigma_\varepsilon^2 I_m$ and $E(\varepsilon\varepsilon_f'|X, X_f) = 0$.
- (Notice the double conditioning to both X and X_f)

Linear regression. "Forecasts"

- In this case the obvious (BLUE) estimate for $E(Y_f|X, X_f)$ is $\hat{Y}_f = X_f\hat{\beta}$ with expected value $X_f\beta$ and variance covariance matrix
$$\sigma_\varepsilon^2 X_f(X'X)^{-1}X_f'$$
- Equivalently, the error in the estimate of $E(Y_f|X, X_f)$ is $X_f\beta - X_f\hat{\beta}$ whose expected value is 0 and variance covariance matrix exactly as above

Linear regression. "Forecasts"

- However, we have an other error to consider which, arguably, is much more interesting
- Define the "point forecast error" as $Y_f - \hat{Y}_f$, the expected value of this shall be 0 and the variance covariance matrix

$$\sigma_\varepsilon^2(I_m + X_f(X'X)^{-1}X_f')$$

- Be very careful not mistaking these formulas with the corresponding (and different) ones for $\hat{Y}$.
- On the basis of these formulas and working either under strong OLS hp (or using Tchebicev), it is possible to derive (exact or approximate) confidence intervals for the estimate of the expected value of each element in the new set of observations and for the corresponding point forecast errors.

Linear regression. "Forecasts"

- On the basis of these formulas and working either under strong OLS hp (or using Tchebicev), it is possible to derive (exact or approximate) confidence intervals for the estimate of the expected value of each element in the new set of observations AND the (different!) interval for the corresponding point forecast (this last takes into account both the estimation error for β and the presence of ε_f).
- For instance, under the Gaussian hypothesis, the two tails confidence ($\alpha\%$) interval for the expected value of a single observation in the forecasting sample corresponding to a row of values of X_f equal to x_f, that is the confidence interval for : $E(x_f\hat{\beta}_{OLS}) = x_f\beta$ under the hypothesis of a known error variance, is given by:

$$\left[x_f\hat{\beta}_{OLS} \pm z_{1-\alpha/2}\sigma_\varepsilon\sqrt{x_f(X'X)^{-1}x_f'}\right]$$

- The corresponding confidence interval for the point forecast, that is, the forecast interval which keeps into account both the estimation error for β and the presence of ε_f, is:

$$\left[x_f \hat{\beta}_{OLS} \pm z_{1-\alpha/2} \sigma_\varepsilon \sqrt{(1 + x_f (X'X)^{-1} x_f')} \right]$$

Linear regression. "Forecasts"

- Notice that

$$\left[x_f \hat{\beta}_{OLS} \pm z_{1-\alpha/2} \sigma_\varepsilon \sqrt{x_f (X'X)^{-1} x_f'} \right]$$

- is a true confidence interval, as $x_f \beta$ is a "parameter" and the interval is there to account for the fact that we do not know β and estimate it with sampling error thru $\hat{\beta}_{OLS}$

Linear regression. "Forecasts"

- On the other hand:

$$\left[x_f \hat{\beta}_{OLS} \pm z_{1-\alpha/2} \sigma_\varepsilon \sqrt{(1 + x_f (X'X)^{-1} x_f')} \right]$$

- Is a mix of a confidence and probability interval, as it keeps into account both the uncertainty due to sampling error (which acts thru the estimate of β), AND the further uncertainty involved in the fact that y_f depends on ε_f

Linear regression. "Forecasts"

- In the case the error variance σ_ε^2 is not known, estimate it with

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

- As described above, the only changes to be made in the formulas are: σ_ε substituted with its estimate $\hat{\sigma}_\varepsilon = \sqrt{\hat{\sigma}_\varepsilon^2}$ and $z_{1-\alpha/2}$ substituted with ${}_{n-k}t_{1-\alpha/2}$
- It is easy to see that the point forecast interval shall always be bigger than the expected value interval (for the same α), as it takes into account not only of the sampling uncertainty in estimating β but also of the uncertainty added by ε_f.

Linear regression. "Forecasts"

- Now, what happens in the case of stochastic X
- Clearly, while the probabilities of these intervals are independent both on X and x_f, these forecast intervals are useful only if we condition to X and x_f
- In fact, while the probability statements would be exact also in the unconditional case, lack of knowledge of X and x_f would give us an interval with random extremes, hence not really useful for forecasts
- This is not a true problem if, when making forecasts in the linear regression setting, we know both X and x_f
- The more general, case where we need to forecast both y_f and x_f, is very interesting but goes beyond the topics of this course

Lecture: Linear Regression 5

How to read the results of a regression

- See the specific handouts

Lecture: Style Analysis

Summary

- A classic use of the linear model in finance: style analysis
- The need for a (relative) benchmark in comparing asset managers
- Why a relative benchmark may be useful even when we do not have a precise asset pricing model
- Sharpe's style analysis procedure: hypotheses and implementation

Style analysis

- Style analysis is interesting both from the point of view of practitioner's Finance and as an application of the linear regression model.
- The current version of the model was elaborated by William F. Sharpe in a series of papers beginning in 1989.
- In this summary we shall refer to the 1992 paper (as of November 2018 you may download it at http://www.stanford.edu/∼wfsharpe/art/sa/sa.htm).

Style analysis

- In order to understand the origin of the model we must recall the intense debate developing during the eighties about the validity of the CAPM model, its possible substitution with a multifactor model and the evaluation of the performance of fund managers.
- In a nutshell (some more detail in the future): a factor model is a tool for connecting expected returns of securities or securities portfolios to the exposure of these securities to non diversifiable risk factors.
- The CAPM model asserts that a single risk factor, the "market", or, better, the random change in the "wealth" of all agents which is, by hypothesis, fully invested in the market, is priced in terms of a (possible) excess expected return.

Style analysis

- This factor is empirically represented by the market portfolio, that is: the sum of all traded securities.
- The expected return of a security in excess of the risk free rate (remember that we are considering single period models) is proportional to the amount of the correlation between the security and the market factor.
- The proportionality factor is the same for all securities and is called price of risk.

Style analysis

- Multifactor models, such as the APT, suggest the existence of multiple risk factors (not necessarily traded) with different prices of risk, so that the cross section of expected security (or security portfolios) excess returns is "explained" by the set of the security expositions to each factor.
- Classical implementations of the APT were based on economic factors, some were tradable, like the slope of the term structure of interest rates, some, at least at the time, non tradable, as GNP growth and inflation.
- At the turn of the nineties Fama and French, followed by others, produced a number of papers where factors were represented by spread portfolios.

Style analysis

- The most frequently used factors were based on the price to book value ratio, on the size of the firm and on some measure of market "momentum" (relative recent gain or loss of the stock w.r.t. the market).
- These factors were represented, in empirical analysis, by spread portfolios.
- As an instance: the price to book value ratio was represented by the p&l of a portfolio invested, at time zero, in a zero net value position long in a set of high price to book value stocks and short in a set of low price to book value stocks.
- Fama and French asserted that the betas w.r.t. this kind of factor mimicking portfolios were "priced by the market", that is, the correlation of a stock return with such portfolios implied a non null risk premium.
- Securities excess expected returns, should be equal to the sum of each security beta wrt each factor multiplied by the price of risk of the factor

Style analysis

- Consider now the problem of evaluating the performance of a fund manager.
- A preliminary problem is to understand for which reason you, the fund subscriber, should pay the fund manager.
- Obviously, you should not pay the fund manager beyond implementation costs (administrative, market transactions etc) for any strategy which is known to you at the moment you subscribe to (or do not withdraw from) the fund, if this strategy gives "normal" returns and if you can implement it by yourself.

Style analysis

- Suppose, for instance, that the asset allocation of the fund manager is known to you before subscribing the fund.
- Since the subscription of the fund is your choice the fund manager should not be paid for the fund results due to asset allocation, or, better, should not be paid for this beyond implementation costs.
- A bigger fee could be justified only if, by the implementation of management decisions you cannot forecast on the basis of what you know when you subscribe the fund, the fund manager earns some "exceptional" (and supposedly positive) return.
- This is the reason why index funds should (and, in markets populated by knowledgeable investors, usually do) ask for small management fees.

Style analysis

- This should happen for any fund managed with some, say, algorithm, replicable on the basis of a "style" model
- For instance, funds which follow asset selection procedures based on variants of the Fama and French approach (that is: stock picking based on observable characteristics of the firms issuing the equity as, for instance, accounting ratios, momentum etc.)
- While implementing such models requires some care and a lot of good data management, the reader should be aware of the fact that nothing magic or secret is required for the implementation of these algorithms.

Style analysis

- The fund manager contribution, with a possible value for you, if any, should be something you cannot replicate
- That is: either something arising from (unavailable to you) abilities or information of the manager or, maybe, from some monopolistic or oligopolistic situation involving the manager.
- Let us suppose (a very naive idea!) that the second hypothesis is not relevant (in most jurisdictions this is, in principle, against the law).

Style analysis

- A formal way to say that the manager ability is not available to you is to say that you cannot replicate its contribution to the fund return with a strategy conceived on the basis of your knowledge.
- It is important to understand that, according to this view, the evaluation of the performance of a fund manager is, first of all, subjective.
- It depends on your knowledge.
- It is the addition of hypotheses on the set of information used by subscribers and on their willingness to optimize using these information that can convert the subjective evaluation into an economic model.
- The problem here is, obviously, to define what we mean by "normal return" and "known strategy".

Style analysis

- Here a market model, representing efficient financial use of public information, could be the sensible solution.
- Were the market model and the effective asset manager's asset allocation available, the first could be used to define the efficiency of the second and, by difference, possible unexpected (by the model) over or under performances on the part of the fund manager.
- Alas, for reasons that shall be discussed in following sections, satisfactory empirical versions of market models still have to appear or, at least, versions of market models, and statistical estimates of the relative parameters, strong enough to be agreed upon by everybody an so useful in an inter-subjective performance analysis.

Style analysis

- A less ambitious and more empirically oriented alternative is return based style analysis.
- This alternative yields a (model dependent) subjective statement about the quality of the fund.
- We shall return on this point but we stress the fact that, if the purpose of the method is for a potential subscriber or for someone already invested in the fund to judge the fund manager performance and not for some agency to award prizes, the subjective component of the method is by no means a drawback.

Style analysis

- Return based style analysis can be seen as a specific choice of "normal return" and "known strategy" definitions.
- The "known strategy" is the investment in a set of tradable assets (typically total return indexes) according to a constant relative proportion strategy
- The "normal return" is the out of sample return of this strategy previously tuned in order to replicate the historical returns of the fund.
- This point has to be hammered in, so we repeat: the strategy is not chosen in order to yield "optimal" returns (in any case the lack of a market model would impede this) but only in order to replicate as well as possible, in the least squares sense, the returns of the fund strategy.

Style analysis

- In order to estimate the replica weights, that is: the weights that best fit a constant relative proportion strategy to the fund performance, we run a regression of the returns (R_t^{Π}) of the fund under investigation on those of the "indexes" involved in the to constant relative proportion strategy
- The weights shall be β_j one for each of k predetermined indexes with returns R_{jt}:

$$R_t^{\Pi} = \sum_{j=1,\ldots,k} \beta_j R_{jt} + \varepsilon_t$$

- The term "constant relative weights strategy" indicates, as usual, a strategy where the proportion of wealth invested in any given index is kept constant over time.
- This implies that, when some index over performs other indexes, a part of the investment in the over performing index must be liquidated and invested in the under performing indexes.

Style analysis

- For the sake of comparison other possible strategies could be the buy and hold strategy, where a constant number of shares is kept for each index and, alternatively, the trend following strategy, where shares of "loser" indexes are sold to buy shares of "winner" indexes.
- Both these strategies have variable weights on returns and could reasonably be used as reference strategies.
- There exist variants of the constant relative proportions strategy itself.
- In a constrained version the weights could be required to be non negative (short positions are not allowed).
- In another version weights could be allowed to change over time (in this case we should assume that the sum of all weights is constant over time).

Style analysis

- In typical implementations no intercept is in the model and the sum of betas is constrained to be one.
- The constant is dropped because it is usually interpreted as a constant return and, over more than one period, a constant return cannot be achieved even from a risk free investment.
- The assumption that the sum of all weights is one is required for the interpretation of the weights as relative exposures and, in the case of a multi period strategy, in order for the portfolio to be self financing.

Style analysis

- While both interpretations and both constraints could be challenged, in our applications we shall stick to the common use.
- We only relate the fact that, sometimes, instead of imposing the sum to one constraint explicitly at the estimation time, this is implemented on a *a posteriori* basis by renormalizing estimated coefficients.
- The two methods do not yield the same results.

Style analysis

- A relevant point in the choice of the reference strategy is that it should not cost too much.
- In this sense the constant relative proportions strategy could be amenable to criticism as it can imply non negligible transaction costs.
- The reason for its use in style analysis seems more leaning on tradition than on suitability.

Style analysis

- Notice that we are not supposing that the fund under analysis actually follows a constant relative proportion strategy invested in the provided set of indexes.
- We are NOT trying to discover the true investment of the fund but only to replicate its returns as best as we can with some simple model.
- This point has to be underlined because, at least in the first paper on the topic, Sharpe himself seems to state that the purpose of the analysis is to find the actual composition of the fund.
- This is obviously impossible if it is not the case that the fund is invested, with a constant relative proportions strategy, in the indexes used in the analysis.

Style analysis

- In fact, the actual discovery of the composition of the fund and its evolution over time would hardly add anything to the purpose of identifying the part of the fund's strategy not forecastable by the fund subscriber.
- A model would still be needed in order to divide what is forecastable from what is unforecastable in the fund evolution.
- Let us go back to the identity:

$$R_t^{\Pi} = \sum_{j=1,\ldots,k} \beta_j R_{jt} + \varepsilon_t$$

- Up to now this is not an estimable model but, as said above, an identity.

Style analysis

- In order to convert this into a model, we must assume something on ε_t.
- A way of doing this is to recall the chapter on linear regression.
- The style model is clearly similar to a linear model.
- In particular it is similar to a linear model where both the dependent and independent variables are stochastic.
- In this case we know that a minimal hypothesis for the OLS estimate to work is that $E(\varepsilon|R^I) = 0$ where ε is the vector containing the observations on the n ε_t-s and R^I is the matrix containing the n observations on the returns of the k indexes.
- The second, less relevant, hypothesis is the usual $E(\varepsilon\varepsilon'|R^I) = \sigma_\varepsilon^2 I_n$.

Style analysis

- The hypothesis $E(\varepsilon|R^I) = 0$ has a sensible financial meaning: we are supposing that any error in our replication of the fund's returns is uncorrelated with the returns of the indexes used in our replication.
- Sharpe's suggestion for the use of the model in fund performance evaluation is as follows:

1. given a set of observations (typically with a weekly or lower frequency, Sharpe uses monthly data) from time $t = 1$ to time $t = n$
2. fit the style model from $t = 1$ to $t = m < n$ and
3. use the estimated coefficient for forecasting R_{m+1}
4. then add to the estimating set the observation $m+1$ (and, in most implementations) drop observation 1.
5. Forecast R_{m+2} and so on.

- These forecast represent the fund's performances due to its style as interpreted by our replicating model.

Style analysis

- The important point is that this "style" result is forecastable and, in principle, replicable by us.
- The possible contribution of the fund manager, at least with respect to our replication strategy, must be found in the forecast error.
- The quality of the fund manager has to be evaluated only on the basis of this error.

Style analysis

- There are three possibilities:

1. The fund manager return is similar (in some sense to be defined) to the replicating portfolio return. In this case, since you are able to replicate the result of the fund manager strategy using a dumb strategy, you shall be willing to pay the fund manager only as much as the strategy costs.
2. The fund manager returns are less than your replica returns.
3. The fund manager returns are better than your replica strategy. In this case it seems that the manager adds to the fund strategy something which you cannot replicate.

Style analysis

- The important point to remember, here, is that the result is relative to the strategy and the asset classes used.
- No attempt is made to build optimal portfolios with the given asset classes, only replica portfolios are built. The reader should think about the possible extensions of procedures like style analysis were a market model available
- A simple example of style analysis using my version of Sharpe's data and three well known US funds is in the worksheet style analysis.xls.

Style analysis

- Traditional approaches with some connection to style analysis
- The idea that you should find some "normal" return with which to compare a fund return and that this definition of "normal" return is to be connected with the return of some simple strategy related with the fund's strategy is so basic that many empirical attitudes are informally justified by it.
- On a first level, we observe very rough fund classifications in families of funds, defined by broad asset classes.
- This suggests comparisons of funds to be made only inside the same family.
- In a sense the comparison strategy is implicitly considered as a mean of the strategy in the same asset class.

Style analysis

- Another shadow of this can be found in the frequently stressed idea that the result of any fund management must be divided between asset allocation and stock picking.
- In common language this partitioning is not well defined and asset allocation may mean many different things as, for instance, the choice of the market, the choice of some sector, the choice of some index.
- Moreover there is no precise definition of how to distinguish between asset allocation and stock picking.
- But it is clear that this distinction, again, hints at some normal return, derived by asset allocation, and some residual: stock picking.

Style analysis

- The benchmarking idea is another crude version of the same:
- You try to separate the fund manager's ability from the overall market performance by devising a benchmark which should summarize the "market" part of the fund manager strategy.

Style analysis

- Market models can be seen as a step up the ladder.
- Here the benchmark idea is expressed in a less naive way.
- Under the hypothesis that the market model holds and is known and the beta (CAPM) or betas (APT) of the fund are known, the part of the result due to the market factor(s) is to be ascribed to the overall fund strategic positioning and, as such, its consequences are in principle a choice of the investor.
- Any other over or under performances can be ascribed to the fund manager abilities and private information.

Style analysis

- Critiques to style analysis
- Under the hypotheses and the interpretation described in the previous section style analysis can be considered an useful performance evaluation tool.
- However, at least in the version suggested by Sharpe, it lends itself to some strong critique.
- A first very simple critique concerns the choice of the replicating strategy.
- While the use of indexes does not create big problems, at least when these indexes can be reproduced with some actual trading strategy, a big puzzle lies in the choice of a constant relative proportion strategy.

Style analysis

- This is both an unlikely and a costly strategy, due to portfolio rebalancing.
- The typical simple strategy is the buy and hold strategy, most indexes are, in principle, buy and hold strategies and the market portfolio of CAPM is a buy and hold strategy.
- As seen in chapter 1 the buy and hold strategy is NOT a constant relative proportion strategy.
- Moreover, a buy and hold strategy, typically, implies very small costs (the reinvestment of dividends is the main source of costs if there is no inflow or outflow of capital from the fund) while a constant relative proportion strategies implies a frequent rebalancing of the portfolio.

Style analysis

- Now, the replicating strategy is a free choice of the analyzer
- However, if we simply suppose that the fund follows a buy and hold strategy in the same indexes used by the style analyzer we end with a strange, if perfectly natural, result.
- Obviously the R^2 of the model shall not be 1, except in the case of identical returns for all the indexes involved in the strategy, moreover the analysis shall point out as non forecastable and so due to the fund manager action, any return of the fund due to the lack of rebalancing implied in a buy and hold strategy.

Style analysis

- A second critique, of theoretical interest but hardly relevant in practice, is connected with Roll's critiques to CAPM tests and, more in general, to CAPM based performance evaluation.
- If the Constant proportion strategy does not contain all the indexes required for composing an efficient portfolio, any investment by the fund manager into the relevant excluded indexes shall result in an over performance.
- This would be relevant if the evaluated fund manager should know, ex ante, the style model with which his/her strategy shall be evaluated AND if the fund manager has a more thorough information on the structure of the efficient portfolio.

Style analysis

- The point is that, while it is rather easy (and not so useful) to compute an efficient portfolio ex post, this is not so easy ex ante.
- The reason is always the same, we discussed it in a previous lecture: it is very difficult to estimate the expected excess returns vector.
- However, this is exactly one of the reasons why style analysis only ask for a relative benchmark.
- In fact, if we accept the idea that the style decomposition depends on the information of the analyzer, this critique loses much of its stance.

Style analysis

- A third, and more subtle, critique can be raised to style analysis as well as to any OLS based factor model used for performance evaluation.
- If the model is fitted to the fund returns, the variance (or sum of squares, if no intercept is used) of the replicating strategy shall always be less than or equal to that of the fund returns.
- In a CAPM or APT logic this is not a problem, since only non diversifiable risk should be priced in the market.
- However, as stressed above, we are NOT in a CAPM or APT world.
- With this lack of variance we are giving a possible advantage to the fund.

Style analysis

- Ways for correcting this problem can be suggested and, in fact, performance indexes which take into account this problem do exist.
- However, since, as we saw above, the positive (for the fund) result is already a weak result in style analysis, this undervaluation of the variance is only another step in the same direction: negative valuations are strong, neutral or positive valuations could be challenged.

Style analysis

- A last word of warning. Many data providers and financial consulting firms sell style analysis.
- As far as I know, the advertising of commercial style models invariably asserts the ability of such models to discover the true composition of the fund portfolio and most reports produced by style analysis programs concentrate on the time evolution (estimated by some rolling window OLS regression) of portfolio compositions.
- This is quite misleading (Sharpe is somewhat responsible as in the original papers he seems to share this opinion) and can be accepted only if interpreted as a misleading way to asses the true purpose of the strategy, that is: return replication.

Style analysis

- Discuss the example (Exercise 10)

Lecture: Linear Factor Models

Summary

- Understanding portfolio risk: linear factor models for asset pricing
- Some empirical evidence
- The basic questions of asset pricing (why so many securities with so high return correlation ans so different statistical properties?)
- The answer given by linear asset pricing models and their basic mathematics

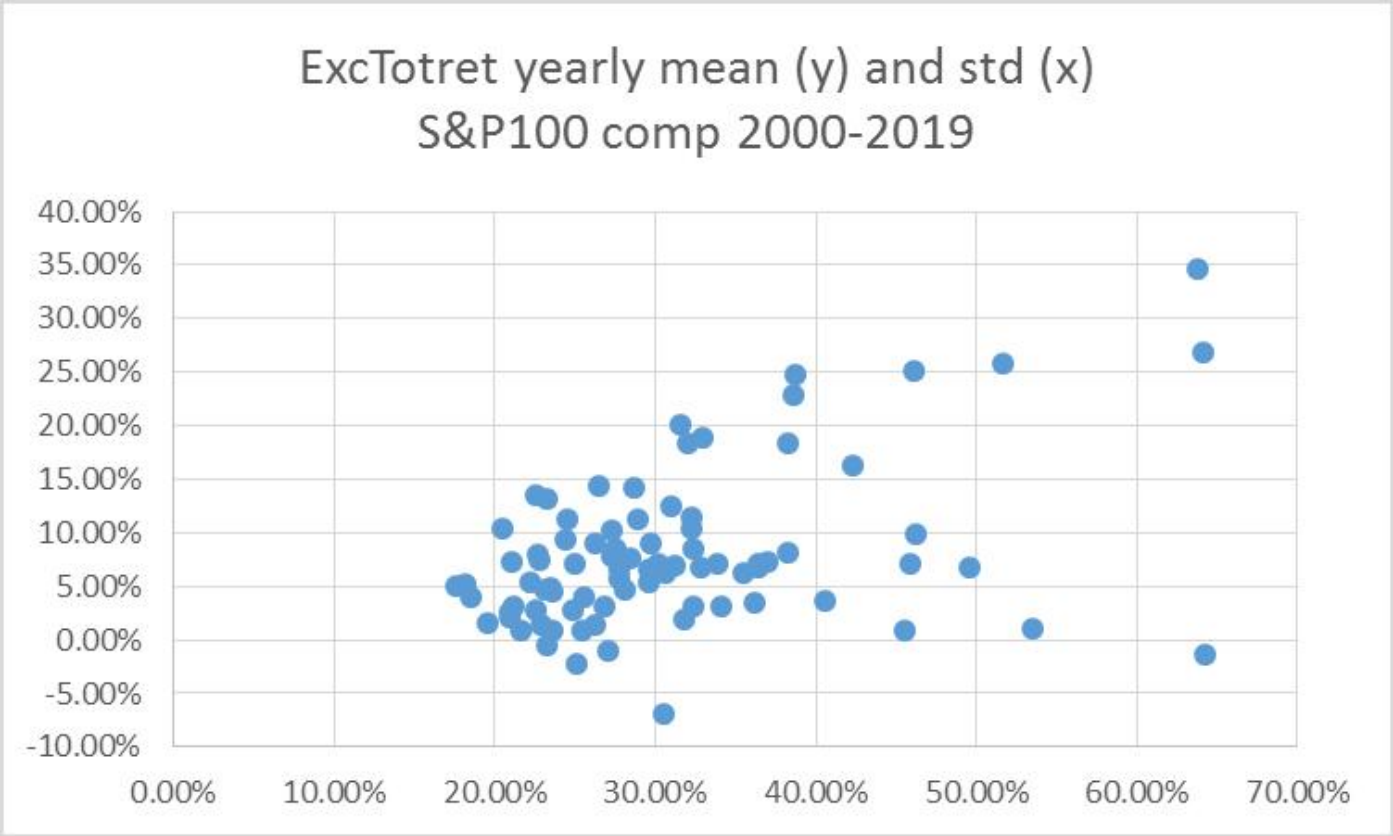
Linear factor models

- This is a very interesting and relevant topic in asset pricing and management
- This is not the course where you study this topic in detail
- Howeve, since most of the applications of regression in finance are connected with factor models, we need a quick summary
- By the way: implicitly, style analysis, while not strictly requiring factor models, can be interpreted in that direction

Linear factor models

- Let us begin by considering the following plot.
- Here you find yearly total return averages and standard deviations for those stocks which were in the S&P 100 index from 2000 to 2019 (weekly data, 84 stocks).

Linear factor models



Linear factor models

- As you can see, stocks with similar average total return show very different standard deviations and vice versa.
- We know that the statistical error in the estimate of expected returns using average returns may be big,
- A back of the envelope computations gives us $\pm 2\sigma$ confidence intervals of about ± 11 (20 years and average yearly std of about .25)
- However, if we believe that average return has anything to do with expected return and standard deviation with risk the plot is puzzling
- (And these are 84 BIG companies).

Linear factor models

- We can see asset pricing models as tools devised to answer the kind of puzzles which plots like this one may raise.
- Among these two of the oldest and most relevant questions of Finance:

1. in the market we see securities whose prices evolve in completely different ways. There may even be securities that have both mean returns lower and standard deviations of returns higher than other securities. Why are all these securities, with such apparently clashing statistical behaviours, still traded in equilibrium?
2. which are the "right" equilibrium relative prices of traded securities?

Linear factor models

- We anticipate here the answers to these two questions given by asset pricing models:

Linear factor models

1. securities prices can be understood only when securities are considered within a portfolio. Completely different securities (in terms, say, of means and variances of returns) are traded because they contribute in improving the overall quality of the portfolio (in the classic mean variance setting this boils down to the usual diversification argument), what is relevant is not the total standard deviation of each security but how much of this cannot be diversified out in a big portfolio, for this reason the expected return of a security return should not be compared with its standard deviation but only with the part of this standard deviation which cannot be diversified ;
2. the right (excess) expected returns of different securities should be proportional with the non diversifiable risks "contained" in the returns, to equal amounts of the same risk should correspond equal amounts of expected return.

Linear factor models

- There are other observed properties of asset prices/returns that asset pricing models try to account for.
- Another striking property is as follows:
- While thousands of securities are quoted, there seems to be a very high correlation, on average, among their returns.
- In a sense it is as if those many securities were "noisy" versions of much less numerous "underlying" securities/"risk sources".

Linear factor models

- For instance, the 83 stocks of the S&P 100 plotted above show an average (simple) return correlation (over 20 years!) of .31.
- If we recall the discussion connected with the spectral theorem and compute the eigenvalues of the covariance matrix of these returns, while we see no eigenvalue equal to 0, the sum of the first 5 eigenvalues is greater the 50% of the sum of all eigenvalues, where the last, say, 50 eigenvalues count for about 15% of the total.
- The sole first eigenvalue is about 1/3 of the total.
- This suggest the idea that, while not singular, the overall covariance matrix can be well approximated by a singular and low rank covariance matrix.
- In simpler words, these returns seem the result of linear transforms of few, uncorrelated "risk factors" plus a bit of uncorrelated noise

Linear factor models

- It should be clear how finding an answer to these questions and modelling the the high average correlation of returns is important for any asset manager
- In fact, asset pricing models are central to any asset management style not purely based on gut feelings
- We can deal with these problems within a simple class of asset pricing models known as: "linear (risk) factor models"
- Here we show some hint of how this is done in practice

Linear factor models

- An asset pricing model begins with a "market model" that is a model which describes asset returns (usually linear returns) as a function of "common factors" and "idiosyncratic noise".
- A typical market model for the $1 \times m$ vector of excess returns r_t observed at time t, the $1 \times k$ vector f of "common risk factors" observed at time t and and the $1 \times m$ vector of errors is:

$$r_t = \alpha + f_t B + \varepsilon_t$$

- Where B is a $k \times m$ matrix of "factor weights" and α is a $1 \times m$ vector of constants.

Linear factor models

- We suppose to observe the vectors r_t and f_t for n time periods t.
- Stack the n vectors of observations for r_t and f_t in the $n \times m$ matrix R and the $n \times k$ matrix F, and stack the corresponding error vectors in the $n \times m$ matrix ε
- The hypotheses are: $E(\varepsilon|F) = 0$, $V(\varepsilon_t|F) = \Omega$ and $E(\varepsilon_t \varepsilon'_{t'}|F) = 0 \ \forall t \neq t'$.
- Under these hypotheses a good, if not optimal, estimator of B shall be $\hat{B} = (\underline{F}'\underline{F})^{-1}\underline{F}'\underline{R}$ this is a matrix version of the classic OLS estimator
- (Here $\underline{F}$ and $\underline{R}$ are demeaned versions of F and R (we take each return series in differences from its arithmetic mean)

Linear factor models

- In order to give meaning to the term "idiosyncratic noise" the contemporaneous correlation matrix Ω is, as a rule, supposed to be diagonal, typically with non equal variances.
- Such a time series model can, and usually shall, yield a good fit to the data on R (for instance it may show high R^2 for most return series) but, by itself, this has nothing to do with tha validity of an asset pricing model

- The OLS estimate of α can be written in a compact way as

$$\hat{\alpha} = \bar{r} - \bar{f}\hat{B}$$

- Where $\bar{r}$ is the $1 \times m$ vector of average excess return (one for each security excess return averaged over time), $\bar{f}$ is the $1 \times k$ vector of average common factors values (again: averaged over time) and $\hat{B}$ is the matrix of OLS estimated factor weights (one for each factor for each security: $k \times m$.

- If we forget the fact that we only have an estimate of B, the expected value of this, under the above hypotheses, is:

$$E(\hat{\alpha}) = E(r) - E(f)B$$

- As we shall see in a moment, an asset pricing model is valid, if, supposing Ω diagonal, we have that $\alpha = 0$.
- This is usually written as:

$$E(r) = \lambda B$$

- Where $\lambda = E(f)$ is a $1 \times k$ vector of "prices of risk" and in a moment we shall see why this name is used.

- It is now important to stress that this restriction may hold, so that the asset pricing model is valid, and this not withstanding the time series model could offer a very poor fit of r
- On the contrary, the fit could be very good and $\alpha \neq 0$.
- For asset management purposes, however, a possible good fit for the time series model with a $k << m$ could be very useful even when the asset pricing model does not hold.

Linear factor models

- Suppose, for instance, you want to use a Markowitz model for your asset allocation.
- In order to do this you need to estimate the variance covariance matrix of returns.
- This requires the estimate of $m(m+1)/2$ unknown parameters using n observations on returns. With a moderately big m this could be an hopeless task.
- Suppose now the market model works, at least in the time series sense, in the sense that the R^2 of each of the m linear models is big. In this case the variances of the errors are small and:

$$V(r_t) = B' V(f_t) B + \Omega \cong B' V(f_t) B$$

Linear factor models

- Let us now count the parameters we need to estimate the varcov matrix of the excess returns with and without the market model.
- Without the market model, the estimation of $V(r_t)$ would require the estimation of $m(m+1)/2$ parameters
- With the factor model it requires the estimation of $k \times m + k \times (k+1)/2$ parameters, that is: just B and $V(f_t)$.

Linear factor models

- Suppose for instance $m = 500$ and $k = 10$
- The direct estimation of $V(r_t)$ implies the estimation of 125250 parameters
- The the (approximate) estimate based on the factor model "only" 5000+55 parameters.

Linear factor models

- Let us move on a step.
- We must remember that our aim is the construction of portfolios of securities with weights w and excess returns $r_t w$.
- In this case we are not necessarily interested in the full $V(r_t)$ but just in the variance of the portfolio

$$V(r_t w) = w' B' V(f_t) B w + w' \Omega w$$

- It is well possible that $w'\Omega w$ be small, so that $w' B' V(f_t) B w$ be a good approximation of $V(r_t w)$, even if it is not true that all R^2 are big and, by consequence, the diagonal elements of Ω small.

- Suppose that the weights w of different securities in this portfolios are all of the order of $1/m$.
- This simply means that no single security dominates the portfolio.
- We have, then (recall: Ω is a diagonal matrix: no covariances)

$$w'\Omega w = \sum_{i=1}^{m} w_i^2 \omega_i \approx \frac{1}{m} \sum_{i=1}^{m} \frac{\omega_i}{m}$$

- This, with bounded, but not necessarily small, diagonal elements of Ω: ω_i, goes to 0 for m going to infinity.

- This means that, for large, well diversified, portfolios "forgetting" Ω is irrelevant even if its diagonal elements are not small.
- Beware: the hypothesis of a diagonal Ω, that is: idiosyncratic ε_t is fundamental for this result.

- This implies that, even if the $\alpha = 0$ condition does not work, a market model shall be useful to simplify the computation of a portfolio vatiance
- However, from this result, we can shed some light on the reason why we should have $\alpha = 0$ so that

$$E(r) = E(f)B = \lambda B$$

- That is: why an asset pricing model should hold.

- Compute the expected value of our well diversified portfolio

$$E(r_t w) = E(r_t)w = \alpha w + E(f_t)Bw$$

- Suppose $\alpha \neq 0$, recall that B is a $k \times m$ matrix with (supposedly) $k << m$ and we can always suppose that the rank of B is k.

- This implies that the matrix $B'V(f_t)B$ is $m \times m$ matrix of rank $k < m$.
- The matrix $B'V(f_t)B$ is then SEMI positive definite, this implies that there exist $m-k$ orthogonal vectors z such that $z'z = 1$ and $z'B'V(f_t)Bz = 0$.
- According to what discussed in the matrix algebra section and in the presentation of the spectral theorem, under conditions we do not specify here, we can always build from these a set of weights $w^{\$}$ such that $(w^{\$})'\mathbf{1} = 1$, $w^{\$'}B'V(f_t)Bw^{\$} = 0$ and $\alpha w^{\$} > 0$.

- You should understand the reason of the dollar sign.
- The vector $w^{\$}$ is such that it defines a zero (systematic) risk portfolio (zero non diversifiable variance) with positive excess return $\alpha w^{\$}$ (this is the expected excess return but also the effective excess return if the portfolio is well diversified).
- In other words, we created a (systematic) risk free security (the portfolio) which yields a return (in principle arbitrarily) greater than the risk free rate.
- This is an "arbitrage" as one could borrow any amount of money at the risk free rate and invest it in the portfolio with a positive profit and no risk (hence the \$).

Linear factor models

- Provided all the financial operations involved (building the portfolio, borrowing money etc.) are possible and do not cost too much, this should not happen if traders are "reasonable" (and if they know of the existence of the factor model).
- The only way to unconditionally (that is: whatever the choice of $w^{\$}$) avoid this is that $\alpha = 0$ be true, so that

$$E(r) = E(f)B = \lambda B$$

- In the following, optional, slides, you can see in some detail how to perform the arbitrage

*How to make money for free (except some algebra)

- Let us go back to
$$r = fB + \varepsilon + \alpha$$
- (same dimension as before) and
$$V(r) = B'V(f)B + V(\varepsilon)$$
- with $V(\varepsilon)$ diagonal

*How to make money for free (except some algebra)

- Suppose BB' (which is $k \times k$) can be inverted (no problem usually for this) and define

$$Q = I - B'(BB')^{-1}B$$

- notice that $HBQ = 0 \ \forall H$ (recall the properties of $I - X(X'X)^{-1}X'$ in the OLS setting)
- Hence:

$$rQ = fBQ + \varepsilon Q + \alpha Q = 0 + \varepsilon Q + \alpha Q$$

*How to make money for free (except some algebra)

- In words: rQ only depends on the idiosincratic components
- Now, choose any $m \times 1$ vector $g \neq 0$ and define

$$w_g^{\$} = Qg / \mathbf{1}' Qg$$

- ($\mathbf{1}$ is the $m \times 1$ column vector of ones)
- Clearly $\mathbf{1}' w_g^{\$} = 1$ so $w_g^{\$}$ are the weights in a portfolio

*How to make money for free (except some algebra)

- Now the variance of the portfolio with weights $w_g^\$$ shall be

$$V(rw_g^\$) = w_g^{\$'} V(\varepsilon) w_g^\$$$

- which, for a big m and a choice of $w_g^\$$ with weights reasonably "spread out" (proper choice of g) shall be near 0 due to diversification
- But the expected value of the portfolio is

$$E(rw_g^\$) = \alpha w_g^\$$$

- and if this is not the risk free rate (or, if every returs is expressed in excess returns terms, if this is not 0) we have an "asymptotic arbitrage"

*How to make money for free (except some algebra)

- Except for some details...
- (e.g. why $\mathbf{1}'Qg \neq 0$? And why, even if this is not the case, we can do the trick all the same?)
- ...this is a constructive proof AND it gives the gist of the APT no arbitrage reasoning

*How to make money for free (except some algebra)

- **See how important is a factor model?**
- Any relative value arbitrage/absolute return strategy is based on this or similar reasoning
- We should duscuss how to choose g but this is not an asset management course
- Moreover: notice that in this discussion we supposed B to be known

Linear factor models

- Let us now give a "financial interpretation" of the no arbitrage result
- Each element β_{ji} of B represents the "amount" of non diversifiable factor f_j in the excess return of security i
- $E(f_j)$ represents the excess expected return of a security which has a "beta 1" with respect to the j factor and zero with respect to the others
- (if the factor f_j is the excess return of a security, this could simply be the excess return of that security, but this is not required)

- We may then understand the name
- "price of risk for factor j" used for $E(f_j) = \lambda_j$ and
- "risk premium for factor j" given to the "price times quantity" product $f_j \beta_{ji}$.

- Just a moment to notice that, if all λ_j share the same value, say: λ^* then, for each security return, $E(r_i) = \lambda^* \sum_j B_{j,i} = \lambda^* \beta_i$ where $\beta_i = \sum_j B_{j,i}$
- This means that, if "the market" makes no distinction in pricing non diversifiable risks of different origin, we can aggregate the "amounts" of all such risks, for each *ith* return, into a single sum β_i and call this: the amount of "non diversifiable market risk" in the *ith* security return

Linear factor models

- Consequently, the difference between CAPM and multi factor models is not due to the existence of multiple non diversifiable risk factors, correlated or not, but to the fact that the prices of risk paid for each of these sources of non diversifiable variance are equal (CAPM) or different (multi factor models)
- B.T.W.: by now you should have understood that "risk factor" simply means "source of nondiversifiable variance"

- It could also be useful to understand why it may be possible for different sources of non diversifiable variance in returns to be "paid" different prices
- After all, each of these just add some (nondiversifiable) variance to the same return: the same aleatory change in the same money, why should these be considered different?

Linear factor models

- The basic idea, which we shall not develop further (you do this in Theory of Finance) is that such risk generating factors may be correlated with other (economic or not) variables which are of interest for the agents ("enter in their utility function")
- In this case, security expected returns, for a given overall nondiversifiable variance, can be different for different securities if such variance comes from a (to fix ideas) positive correlation with less feared economic risks or positive correlation with more feared economic risks
- In the first case we would argue for a smaller risk premium then in the second (notice that, underlying this, we have some implicit sort of "hedging" argument)

Linear factor models

- Notice that, in order to use factor models for (non strategic) asset management, risk management, performance attribution, performance evaluation etc. we do not need to discover the economic risks behind the underlying risk factors
- We need this if we want to use factor models for strategic asset allocation and for more general purposes of macroeconomic financial policy evaluation
- In this sense, Ross' APT, with its emphasis on interpreting risks in the light of economic variables, has a more strategic flavour than, say, CAPM, Fama and French models or principal component models (to be seen shortly) where "factors" are just represented by portfolios of assets which, supposedly, are most influenced by such underlying risk factors, but with no attempt to identify the specific origin of such risks

Linear factor models

- Now, we have a rough idea of how an asset pricing model works
- Go back to the questions with which this section began
- Think a little bit about how the answers come from the asset pricing model

- We should first notice that the approximation

$$V(r_t w) = w' B' V(f_t) B w + w' \Omega w \cong w' B' V(f_t) B w$$

- is a formal interpretation of the empirical fact that correlations among quoted securities returns are on average high.
- The interpretation is based on the idea that all returns depend (in different ways) on the same underlying factors and what is "not factor" is uncorrelated across returns.
- For this reason, well diversified portfolios of securities tend to show returns whose variance only depends on that of factors.

- As a consequence, it shall be difficult to build many such well diversified portfolios which are not correlated among them.
- In fact, if we assume the above approximation to be exact, and $V(f_t)$ to be non singular, only exactly k of such portfolios can be built, the choice is unique modulus an orthogonal transform.
- In this case it is quite tempting to interpret any choice of such k non correlated portfolios as a "factor estimate".
- Some aspects of this idea shall be reconsidered when presenting the "principal component" way to risk factor estimation.

Linear factor models

- Asset pricing models give a very precise answer to the puzzle about the fact that securities are traded in the market even if they may show, at the same time, lower average returns and higher standard deviations of other traded securities.
- This is a possible equilibrium because what is relevant is not the "absolute risk" (marginal standard deviation of returns) of a security, but its contribution to the risk/return mix in a well diversified portfolio.
- For this reason, we can see a relatively low average return and a high standard deviation simply because the security showing these statistics as little correlation with systematic risk factors.

- The model tells us many other interesting things regarding this point.
- For instance: it tells us that us that if we see two securities with, say, the same average returns and very different return standard deviations, the correlation between the returns of these securities should be small.

Linear factor models

- Last: asset pricing models give us formulas for measuring the "right mix" of expected returns and correlation with systematic risk factors (betas) and this answers to the question about right equilibrium relative prices.
- On this basis, asset pricing models give us an unified framework to precisely quantify and test the equilibrium price system and to transform the statistical results into asset management tools.

Linear factor models

- In what follows we shall discuss some way for estimating linear factor models, in particular the principal component method
- There is no time to go deeper in the empirical aspects of asset pricing models
- To those of you that are curious of how such models could work with the S&P 100 data used above, just a quick peek:
- Regress each excess return on the excess return of the market index and take the residuals for each of the 83 regressions.
- The question is if the residuals of this regression are "idiosyncratic".

Linear factor models

- While the original excess returns are almost invariably positively correlated, residuals show positive and negative correlations.
- A good measure of "overall correlation", then, is the sum of the squared element of the correlation matrix.
- If we take the ratio of the sum of squares of the correlation matrix for the residuals and for the original data we get .23.
- The simple beta model, then, reduces the measure of overall correlation to less than one quarter of the original value.

Linear factor models

- Notice that this is by no means a necessary consequence of the use of OLS
- While, in general, correlations among residuals of OLS regressions on the same variable are lower than the original correlations, they could be negligibly so.
- We deduce that, while other factors may be necessary, the single beta model takes us a long way toward the separation between systematic and idiosyncratic "risk".
- In the handouts you shall find a section dedicated to some hint about how to go on and really test at least the CAPM

Linear factor models

- It is important to remember that asset pricing models are central both for asset management and for Corporate Finance, for this reason they constitute a mainstay of Finance education.
- For the interested Reader a good starting point would be:
- Kennet J. Singleton "Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment", 2006, Princeton University Press.

Lecture: Explicit and Implicit Factors, Factor Identification, Principal Components

Summary

- Empirical implementation of linear asset pricing (and market) models
- Explicit factors and the linear model
- Implicit factors and their connection with market indexes

Linear factor models. Explicit and Implicit Factors

- How do we estimate a market model?
- A first possibility is as follows.
- We suppose to observe the vectors r_t and f_t for n time periods (Explicit factors).
- Let us stack the n vectors of observations for r_t and f_t in the $n \times m$ matrix R and the $n \times k$ matrix F, and the corresponding error vectors in the $n \times m$ matrix ε

- We get
$$R = FB + \varepsilon$$
- Recall the hypotheses: $E(\varepsilon|F) = 0$, $V(\varepsilon_t|F) = \Omega$ and add the hypothesis $E(\varepsilon_t \varepsilon'_{t'}|F) = 0 \ \forall t \neq t'$.

- Under these hypotheses we could estimate B as

$$\hat{B} = (F'F)^{-1}F'R$$

- Notice that this is a MATRIX of dimension $k \times m$.
- This should not worry us: each column of $\hat{B}$ is identical to the OLS estimate we could get regressing the $n \times 1$ Column of R, that is r_j, corresponding to the $j-th$ stock on the $n \times m$ matrix R

Linear factor models. Explicit and Implicit Factors

- This, however, is possible if we can observe F
- Sharpe's "beta model", Ross multifactor APT and Fama and French models are examples of this procedure
- Sharpe's style analysis is in this same strain
- However, there is a second possibility: Implicit factors

Linear factor models. Explicit and Implicit Factors

- If the above hypotheses are correct and, in particular, Ω is diagonal, then, for any "well diversified portfolio" with weights w

$$V(r_t w) \approx w' B' V(f_t) B w$$

- Any such well diversified portfolio has a return variace which depends (almost) only on $V(f_t)$ which is of rank k
- Hence, it shall at most possible to build k of these portfolios the $k+1$ shall be (almost) a linear function of the others
- In other words, if I build $k+q$ well diversified portfolios, their return varcov matrix shall be singular for any $q > 0$

Linear factor models. Explicit and Implicit Factors

- There are many ways of choosing such k portfolios
- Choose any one of these sets mof k portfolios, call their weights W (a column per portfolio) and the returns $g_t = r_t W$

Linear factor models. Explicit and Implicit Factors

- Any other well diversified portfolios shall be a "portfolio of these portfolios"
- Infact, take other well diversified portfolio with weights, say, w^*
- We want this to be, with good approximation, a linear combinations of the k portfolios, that is

$$r_t w^* = g_t h^*$$

- for some h^*, but which h^*?

Linear factor models. Explicit and Implicit Factors

- The solution is simple: since we need $g_t h^* = r_t w^*$ and $g_t = r_t W$
- We have $r_t W h^* = r_t w^*$
- This is possible, independently on r_t, iff $W h^* = w^*$
- This is a linear system where the number of equations (m: rows of W) is greater than the number of unknowns (k: rows of $h^*$and columns of W)

Linear factor models. Explicit and Implicit Factors

- We know a possible way to deal with this: least squares
- Multiply this by W' and get $W'Wh^* = W'w^*$
- Now: $W'W$is a $k \times k$ invertible (assumption) matrix so that the solution is $h^* = (W'W)^{-1}W'w^*$
- (Notice, this is ONE solution to the problem, other exist but we just need just)

Linear factor models. Explicit and Implicit Factors

- But, now, we have

$$V(r_t w^*) = V(g_t h^*) = h'^* V(g_t) h^*$$

- And this means that, for the purpose of building well diversified portfolios, we can use the k portfolios with returns g_t as "factors" (and as "basic" investments).
- It also means that this is everything we need for computing the variance of such portfolios (and if a asset pricing model holds, also its expected value)
- In other words: if we only consider well diversified portfos, we can directly approximate "factors" with (any well diversified) set of k portfolios and do not need to define factors in any other way or even to remember the original set of securities.

Linear factor models. Explicit and Implicit Factors

- How can we build, in practice, these "factor portfolios"?
- In a sense we intuitively do this when we consider "factors" in a style analysis model
- Notice, moreover, that in that setting we try to choose such portfolios as little correlated as possible
- There are many methods, some more "formal" some more "intuitive"
- We shall only consider one of these: principal components.

Identification of factors

- Notice: since, in general, the choice of "factor mimicking portfolios" (this the name) is not unique, in order to get a unique solution we must add nypotheses
- From the beginning, let us in general consider "factors estimates" which are linear combinations of returns (that is: not necessarily portfolio returns)
- Recall: our porpose is, first of all, that of computing variances and minimize these
- Given the sum of the weights in the linear combination we can always, over each single period, add to any linear combination the risk free return with weight equal to one minus the sum of the other weights.
- This would give us the return of a portfolio with a different expected value but the same variance and covariances as the original linear combination.

Identification of factors

- We must then put some further constrain in order to "identify" the factors
- We did see that we have many possible choices
- The first idea is this: any non singular linear transform of the factors is equivalent to the original factors (portfolios of portfolios)
- In order to avoid this indeterminacy, the first idea, which is generally reasonable, is to choose factor portfolios such that their covariance matrix is diagonal.

*Identification of factors

- We can always do so.
- More in general: In a linear factor model we can always assume common factors to be orthogonal.
- In fact, from

$$r_t = f_t B + \varepsilon_t$$

- and $E(r_t|f_t) = f_t B$ *and* $V(r_t|f_t) = V(\varepsilon_t|f_t) = \Omega$.
- We get

$$V(r_t) = B' V(f_t) B + \Omega$$

*Identification of factors

- Suppose $V(f_t) = \Phi$ and use the spectral theorem: $\Phi = H\Theta H'$
- Define $f_t^* = f_t H$ so that $V(f_t^*) = H'H\Theta H'H = \Theta$
- We also have $f_t^* H' = f_t$so that we can rewrite the factor model as

$$r_t = f_t^* H' B + \varepsilon_t = f_t^* B^* + \varepsilon_t$$

- Where $B^* = H'B$

Identification of factors

- The simple choice of non correlated factors does not imply a unique choice of factors
- For this reason we shall add the requirement that each single factor has the maximum possible variance, given the requirement of being orthogonal to the other factors

Identification of factors

- This is still not enough: given a linear combination of returns satisfying orthogonality if I double its weights I get a linear combination which still satisfies orthogonality but with 4 times the variance.
- I can increase this as much as I want and this makes indeterminate the "maximize the variance" problem.
- To avoid this I add a last hypothesis:
- I choose to bound the coefficients of the linear combinations by requiring their sum of squares to be equal to 1.
- Orthogonality, maximum variance and sum of squared weights equal to 1 lead to factors=principal components

Lecture: Principal Components

Summary

- Principal components as a way for estimating implicit factor models
- Two equivalent definitions of principal components
- Definition based on the spectral theorem
- Definition based on "maximum risk portfolios"

Principal components

- As a starting point suppose that the variance covariance matrix of a $1 \times m$ vector of returns r: $V(r)$ is known.
- We introduce the principal components, at first, in an arbitrary way.
- Then we shall justify the choice by showing that principal component satisfy the above requirements
- (orthogonal linear combinations, with maximal variances and sum of squared weights equal to 1).

Principal components

- From the spectral theorem we know that $V(r) = X\Lambda X'$.
- By the rules of matrix product and recalling that Λ is diagonal, we have:

$$X\Lambda X' = \sum_i x_i x_i' \lambda_i$$

- Where x_i is the $i-$th column of X and the sum is from 1 to m.

Principal components

- Notice that, in general, only k eigenvalues of λ_j are greater than 0 while the others are equal to 0.
- Here k is the rank of $V(r)$.
- For simplicity in the following formulas we suppose $k = m$ but with proper changes of indexes the formulas are correct in general.
- Define the "principal components" as the linear combinations of returns given by: $f_j = rx_j$ and let $F = rX$

Principal components

- Clearly we have $V(F) = X'V(r)X = X'X\Lambda X'X = \Lambda$: the f_j are not correlated and the variance of each is the corresponding eigenvalue
- We also know that $X'X = XX' = I$ so that the sum of the squares of each column $x_j$of X is equal to 1
- Moreover $r = FX' = \sum_{l=1}^{k} f_l x_l'$ where x_l' is the $l-$th row of X' (the transpose of the $l-$th column of X)
- So that, for instance, $r = f_j x_j' + \sum_{l \neq j}^{k} f_l x_l'$ and $V(r) = \sum_{l=1}^{k} x_l V(f_l) x_l' = \sum_{l=1}^{k} \lambda_l x_l x_l^{'}$
- (recall: f_l are orthogonal)

Principal components

- With explicit factors, we get the factor weights by regressing returns on factors
- Now what happens if we regress each element of r on a single principal component f_j?
- Do we get, as OLS coefficients, the vector x_j'?
- Which are the residuals of these regressions?
- What is the vector of the R^2 corresponding to these regressions?

Principal components

- Notice that these are m univariate regressions on the same f_j and the "betas" (one for each return in r) of this regressions (supposing the intercept in the model) are, as usual:

$$\beta_j = \frac{Cov(f_j; r)}{V(f_j)} = \frac{E(x_j' r' r) - E(x_j' r')E(r)}{V(f_j)} = \frac{x_j' V(r)}{V(f_j)}$$

Principal components

- However:

$$x_j' V(r) = x_j' X \Lambda X' = x_j' \sum_i x_i x_i' \lambda_i = x_j' \lambda_j$$

- and

$$V(f_j) = V(r x_j) = x_j' X \Lambda X' x_j = \lambda_j$$

- so that:

$$\beta_j = x_j'$$

- Which is what we did expect (and remember: this is the $j-$th row of X' so this is the $j-$th column of X that is: the weights of each return in the linear combination which defines f_j)

Principal components

- The vector of the $R^2$of the regression on the $j-$th principal component is given by the ratio of each element on the diagonal of $V(f_j x_j') = x_j \lambda_j x_j'$ and each element on the diagonal of $V(r)$
- A generic element I on the diagonal of $x_j \lambda_j x_j'$ is given by $\lambda_j (x_{Ij})^2$ where x_{Ij} is the (I,j) element in the eigenvector matrix X
- A generic element I on the diagonal of $V(r) = X \Lambda X' = \sum_{i=1}^{k} \lambda_i x_i x_i'$ is given by $V(r)_{II} = \sum_{i=1}^{k} \lambda_i (x_{Ii})^2$
- So that $R^2_{jI} = \lambda_j (x_{Ij})^2 / (\sum_{i=1}^{k} \lambda_i (x_{Ii})^2)$
- We see that the size of the R^2 for the regression of the $I-$th return on the $j-$th component is proportional to the variance of the component (λ_j) times the square of x_{Ij} that is: the weight of the $I-$th return in the linear combination defining the $j-th$ component

Principal components

- We can prove that, for $V(r - f_j\beta_j)$ that is: the variance of the "residuals", the following is true:

$$V(r - f_j\beta_j) = V(r - rx_jx_j') = \sum_{i \neq j} x_i x_i' \lambda_i = X_{-j}\Lambda_{-j}X_{-j}',$$

- where X_{-j} and Λ_{-j} are, respectively, the X matrix, after dropping column j and the Λ matrix, after dropping row and column j.

Principal components

- In other words, the covariance matrix of the residuals $r - f_j \beta_j'$ has the same eigenvectors and eigenvalues of the original covariance matrix, with the exception of the eigenvector and eigenvalue involved in the computation of f_j.
- This result is due to the orthogonality of factors and has several interesting implications.
- We mention just three of these.

Principal components

- First: one by one "factor extraction", that is: the computation of the f_j and of the corresponding residuals, yields the same results if performed in batch or one by one.
- Second: the result is invariant to the order of computation (provided I use the same factors).
- Third: once all factors are considered the residual variance is 0.

Principal components

- This last obvious result can be written in this way.
- As we did above, If we set $F = rX$ we have $r = FX'$.
- Grouping in F_q and X_q the first q factors and columns of X and in F_{m-q} and X_{m-q} the rest of the factors and columns of X, we have:

$$r = \sum_{i=1}^{m} f_i x_i' = \sum_{i=1}^{q} f_i x_i' + \sum_{i=q+1}^{m} f_i x_i' = F_q X_q' + F_{m-q} X_{m-q}'$$

Principal components

- Which we are tempted to write as:

$$r = F_q X_q' + e$$

- Now recall the initial factor model (we drop the t suffix for the moment):

$$r = fB + \varepsilon$$

Principal components

- It is tempting to equate F_q to f and X_q' to B for some q.
- At the same time it is tempting to equate e with ε.
- Now, given the above construction, it is always possible to build such a representation of r.
- The question is whether, given a pre specified model $r = fB + \varepsilon$, the above described method shall identify f, B and ε.
- The answer is: in general no.

Principal components

- In fact the two formulas are only apparently similar and become identical only under some hypothesis. These are:

1. The dimension of f is q.
2. $V(f)$ is diagonal.
3. $BB' = I$
4. The rank of $V(\varepsilon)$ is $m - q$ and the maximum eigenvalue of $V(\varepsilon)$ is smaller than the minimum element on the diagonal of $V(f)$.

- To these hypotheses we must add the already mentioned requirement that f and ε are orthogonal.

Principal components

- As we did see, for any given fB the second and third hypothesis can always be satisfied if $V(fB)$ is of full rank.
- Hypothesis one is more problematic: all we observe is r and we do not know, a priori, the value of q.

Principal components

- But the most relevant (and interesting) hypothesis is that the rank of $V(\varepsilon)$ should be $m-q$ and its eigenvalues are all smaller than the eigenvalues of $V(fB)$.
- This may well not be the case and in fact we could consider examples where ε is a vector orthogonal to the elements of f but $V(\varepsilon)$ is of full rank and/or its eigenvalues are not all smaller than those of $V(fB)$.

Principal components

- For instance: in classical asset pricing models (CAPM, APT and the like) the main difference between residuals and factors is not that the variance contributed by the factors to the returns is bigger than the variance contributed by factors
- The main point is that factors are common to different securities, so that they generate correlation of returns, while residuals are "idiosyncratic" that is: they should be uncorrelated across securities.
- While principal component analysis guarantees zero correlation across different factors, residuals in the principal component method are by no means constrained to be uncorrelated across different securities
- In fact, since the varcov matrix of residuals is not of full rank, some correlation between residual must exist and shall in general be higher if many factors are used in the analysis.

Principal components

- While this is not the place for a detailed analysis of this important point, it is useful to introduce it as a way for remembering that $r = F_q X_q' + e$ is, before all, a "representation" of r and only under (typically non testable) hypothesis some estimate of a factor model.
- In our setting we need the representation in order to simplify the estimation of $V(r)$, while the interpretation of the result as the estimation of a factor model is very useful, when possible, the simple representation shall be enough for our purposes.

Principal components

- It should always be remembered that our purpose is not the precise estimation of each element of $V(r)$.
- What we really hope for is a sensible estimate of the variance of reasonably diversified portfolios made with the returns in r.
- In this case, even if the estimate of $V(r)$ is rough it may well be that the estimate of a well diversified portfolio variance is fairly precise as, by itself, diversification shall erase most of the idiosyncratic components in the variance covariance matrix.
- This intuitive reasoning can be made precise but it is above the purpose of our introductory course.

Principal components

- **Maximum variance orthogonal factors**
- In the preceding section we derived a principal component representation of a return vector by comparing the spectral theorem with the general assumptions of a linear factor model.
- Here we follow a different path and go back to the beginning of the discussion of factor models: we characterize each principal component (suitably renormalized) as a particular maximum risk portfolio with the constraint that each component must be orthogonal to each other component and that the sum of squared weights should be equal to one.
- Let us comment some more on this last constraint

Principal components

- Given a set of returns, it is interesting to answer the question: which are the weights of the maximum variance linear combination of returns?
- As we did see, this problem is not well defined as the variance of any portfolio (provided it is not 0) can be set to any value by multiplying its weight by a constant.
- It could be suggested to constrain the sum of weights to one, however this does not solve the problem.
- Again, By considering multiples of the different positions the requirement can be satisfied and the variance set to any number, at least, if weights are allowed to be both positive and negative.

Principal components

- A possible solution is to set the sum of the absolute values of the weights to one.
- This would both solve the problem and have a financial meaning.
- Alas, this can be done but only numerically.

Principal components

- Suppose, instead, we set the sum of squared weights to 1.
- This solves the bounding problem with the inconvenient that the resulting linear combination shall in general not be a portfolio.
- But this choice yields an analytic solution.

Principal components

- Let us set up the mathematical problem:

$$\max_{\theta:\theta'\theta=1} V(r\theta) = \max_{\theta:\theta'\theta=1} \theta' V(r)\theta$$

- The Lagrangian for this problem is:

$$L = \theta' V(r)\theta - \lambda[\theta'\theta - 1]$$

Principal components

- So that the first order conditions are:

$$V(r)\theta - \lambda\theta = 0$$

- and

$$\theta'\theta = 1$$

Principal components

- Rearranging and using the spectral theorem we have:

$$[X\Lambda X' - \lambda I]\theta = 0$$

- We see that, if we set $\theta = x_j$ and λ to the corresponding λ_j, for any j we have a solution of the problem.
- Since $V(rx_j) = \lambda_j$, the solution to the maximum variance problem is given by the pair x_1 and λ_1 where, as usual, we suppose the eigenvalues sorted by size.

Principal components

- in fact we have

$$[X\Lambda X' - \lambda_j I]x_j = X\Lambda X' x_j - x_j\lambda_j = \sum_{i=1}^{m} x_i x_i' x_j \lambda_j - x_j\lambda_j = x_j\lambda_j - x_j\lambda_j = 0$$

- Where the last passage comes from the fact that, due to orthonormality of X, m$x_i x_i' x_j = x_i 0 = 0$ if $i \neq j$ and $= x_j$ if $i = j$
- These are, in general m solutions for a rank m matrix $V(r)$.

Principal components

- Are other solutions possible?
- Since the λ_i are the solution of the m degree polynomial equation $|V(r) - \lambda I| = 0$ (where $|.|$ means the determinant) the fundamental theorem of algebra guarantees us that these are the sole solutions
- However, if (λ_j, x_j) is a solution, also (λ_j, cx_j) would be a solution for any scalar c but, in this case, $c^2 x_j' x_j = c^2$ and for $c \neq \pm 1$ the condition on the sum of squares of θ would not be satisfied
- In sum: these are the sole solutions modulus a possible change of all signs in the eigenvectors elements
- Notice that the orthogonality between different solutions directly comes from the properties of the eigenvectors

Principal components

- We know that each λ_j is the variance of the corresponding principal component
- We can, then, qualify the solutions as orthogonal linear combinations with sum of squared coefficients equal to one which, given these constraints, give us all (relative) maximal variances
- It is custumary in principal component analysis to order the eigenvalues from the biggest to the smallest

Principal components

- The question, now, is what we do observe if we compute principal component for a given variance covariance matrix
- Anything may happen betwee two limit cases

Principal components

- On one extreme side we have the case of a variance covariance matrix of a set of m identical (hence, perfectly correlated) random variables
- In this case only one eigenvalue shall be different than 0 and to this shall correspond a single component with eigenvector weights all equal to $1/\sqrt{m}$ (or minus this)
- On the opposite extreme case, all random variables could be uncorrelated with the same variance
- In this case all eigenvalues shall be identical and identical to this common variance while each principal component shall correspond to one of the variables
- Each eigenvector shall contain one 1 (or -1) in one position, different for each eigenvector due to orthogonality, and 0 in all other positionsand We know that each λ_j is the variance of the corresponding principal component

Principal components

- You must notice that these two extremes are characterized not only by "correlation" but by the possibility of summarizing the variance covariance matrix using few components
- To understand this consider the case of m uncorrelated random variables, one with a "big" variance and $m-1$ with "small" variances
- In this case, due to lack of correlations, the eigenvectors shall be exactly as in the second case on the previous slide
- However the eigenvalue pattern shall be similar to that of the first case
- In short: the eigenvalue pattern tells you if you can summarise the phenomenon in few components
- The eigenvector pattern tells you how correlated are the random variables you are studying

Principal components

- In Finance, you most frequently apply principal components to varcov matrixes of returns
- Since you can invest in the corresponding securities, suitably renormalized principal components are returns from constant relative proportion investment portfolios
- This adds a dimension to our interpretation of principal components
- These portfolios define orthogonals dimensions in the space of possible (constant relative proportion) investments in the market defined by the returns in the varcov matrix

Principal components

- Each dimension/portfolio has maximal variance (say: "risk") conditional to being orthogonal to the others
- If few eigenvalues are "big" this means that, either due to lots of correlation or to the fact that few securities have a big variance, most of the risk in the market can be summarized in few portfolios
- If eigenvectors coefficients are not concentrated in few entries this implies that correlation among securities is high

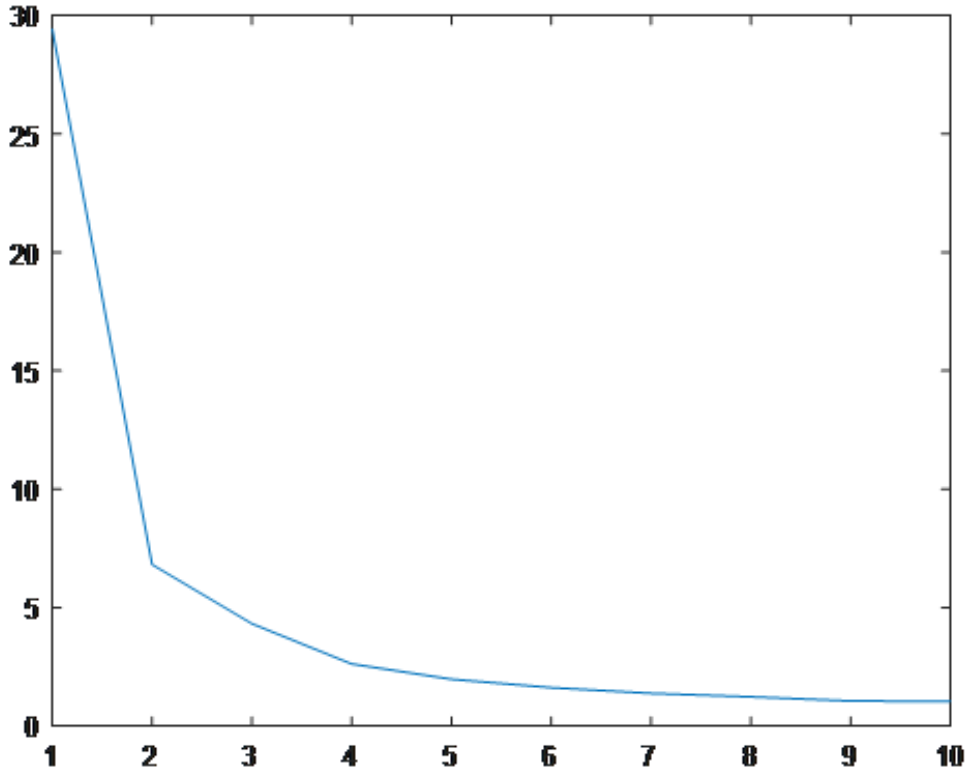
Principal components

- Almost invariably, financial data, and not only data on stock prices/returns, show patterns nearer to the high correlation case that to the low correlation, equivariance or non equivariance, case
- We observe few large eigenvalues and the eigenvectors corresponding to these eigenvelues show fairly homogeneously distributed weights

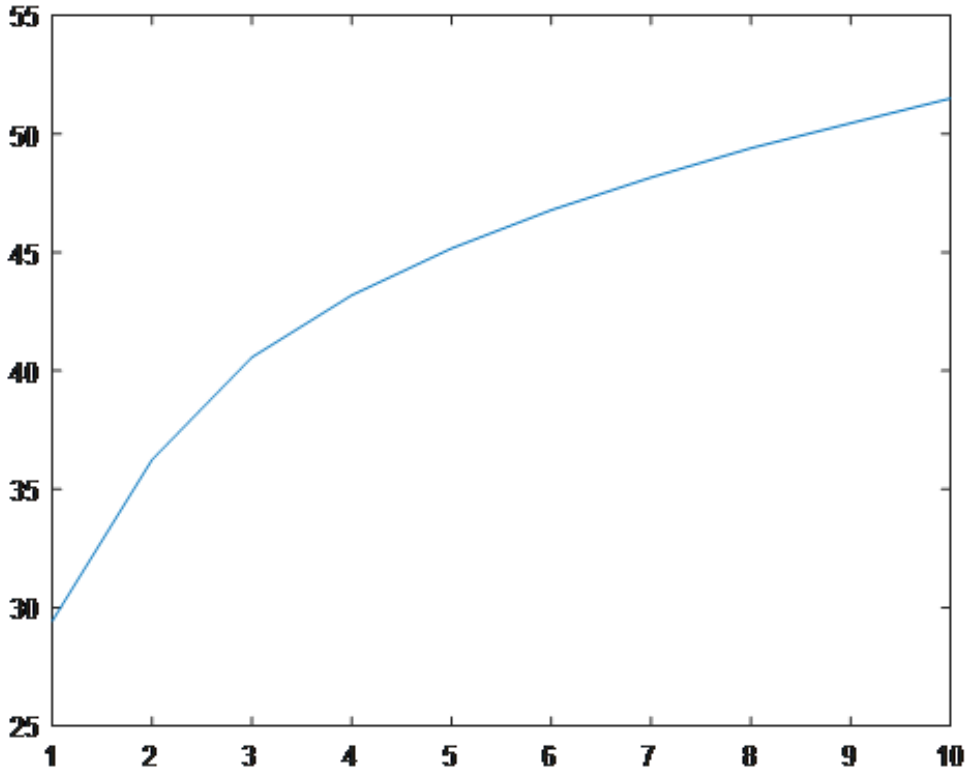
Principal components

- The following example is based on 10 years of weekly data on the S&P500 for the time interval end 1999-end 2009
- In this period 443 stocks were constant members in the index
- In the following pictures we see the eigenvalues and the cumulative eigenvalues for the first 10 PCs

Principal components



Principal components



Principal components

- As you can see, 30% of the total sum of the eigenvalues is due to the first PC
- The first 10 PCs cumulative percentage of eigenvalues is above 50%
- All this for 10 years of data in a very messy time span with two big crises (end of the internet bubble and 2008)

That's all, folks

